AN EXAMINATION OF THE OBJECTIVE EVALUATION OF STUDENT ACHIEVEMENT IN ANATOMY, WITH AN ENQUIRY INTO THE RESULTS OF CYCLING MARKING PROGRAMS AND CONFIDENCE WEIGHTING OF RESPONSES.

The survey of an investigation into the results obtained over 7 years of experiment in the Department of Anatomy, University of Cape Town.

MARCUS FREDMAN.
B.Sc. (Med).
M.B., Ch.B. (UCT).
F.R.C.S. (Eng.).

Department of Anatomy
UNIVERSITY OF CAPE TOWN.

THESIS
Submitted for the Degree of
DOCTOR OF PHILOSOPHY -
UNIVERSITY OF CAPE TOWN.

1977

To Margo

DECLARATION

I declare that the ideas presented and work done in execution of this thesis are original except where acknowledged in the text, and have not been previously published.

All the computer programs were devised by me; the writing, debugging and testing, however, of certain programs was done by Mr. I. MacArthur, (MAC Marking Program), Mr. P. Derriman (Fortran Marking Program, Cycling Program and Correlation of Components of an Examination) and Mr. B. Rassiner (Analysis of Confidence Marking) - all to my specifications.

Assistance was obtained from Mr. L.D.A. Wright and Dr. K.A. Woodhouse, both of the University of Cape Town Computer Centre, when I ran into problems that I could not solve in writing the programs to draw the graphs that illustrate this thesis.

Other than the above, all programs and work on the computer was my responsibility.

Signed

M. Fredman.

PREFACE

Every teacher must face the solution of three major problems in his field.

1. What should I teach,

2. What are the best methods that I can employ in teaching,

3. What methods can I use to assess the effectiveness of my teaching.

The present enquiry has grown from my interest in these problems with constant encouragement from Professor L.H. Wells, formerly Head of the Department of Anatomy at the University of Cape Town, and more recently from his successor, Professor E.N. Keen.

If teachers can use reliable methods for assessing the effectiveness of their teaching and apply these results to the selection of the best methods of teaching they shall earn the gratitude of those countless students of the future who will be facing the enormous learning requirements demanded by the speed of advances in knowledge and technology in this modern age.

This work is a study of the role of just one method of assessing the effectiveness of teaching.

ACKNOWLEDGEMENTS

I should like to express my gratitude to:-

Professor L.H. Wells who supported the initial project of setting multiple-choice examinations in the Department of Anatomy, encouraged the idea of undertaking this study, who acted as my supervisor in the early stages of the investigation, and who kept the enquiry alive when it showed signs of flagging.

Professor E.N. Keen who inherited the system and asked some of the penetrating questions that I have attempted to answer in this investigation, and who acted as my supervisor in the later stages of the investigation.

Professor A. Davison who initially gave me the idea of machine-recording student answers on the computer by his work in the Department of Physiology and Medical Biochemistry at the University of Cape Town.

Mr. I. MacArthur, Mr. P. Derriman and Mr. B. Rassiner, who in turn have been responsible for writing some of the marking and analysis programs that were required.

The members of the University of Cape Town M.C.Q. committee who devised the sophisticated U.C.T. Model Marking Program on the basis of the program then in use by the Department of Anatomy, namely Dr. A.V. Hall, Dr. L. Nassimbeni, Dr. S. Schach, Mr. C. Melzer, Mr. L. Gilbert and Mr. M. Fielding.

Dr. J. Juritz for advice on statistical methods.

Mr. D.E. Stegman and Mr. B. Lichtman of the Computer Centre who were responsible for the implementation of the U.C.T. Model Multiple-Choice Marking Program used in this investigation since 1973.

Mr. L. Frost - Operations Manager of the University of Cape Town's Computer Centre, for assistance in the running of some of my programs.

Mrs. M. Verchin and Mrs. D. Downie for their assistance in typing the drafts of this manuscript, and Mrs. Wakeham for typing the final manuscript.

TABLE OF CONTENTS

BOOK 1.

3./.....

1./.....

BOOK II.

BOOK 1.

ADDENDUM

AN EXAMINATION OF THE OBJECTIVE EVALUATION

OF STUDENT ACHIEVEMENT IN ANATOMY

ADDITIONAL COMMENTS

Since completion and compilation of this thesis certain aspects and conclusions have presented themselves for further comment.

These are

****

1. KUDER-RICHARDSON 20 RELIABILITY COEFFICIENT VALUES.

In the results published in the thesis when cycling was used values for the Kuder-Richardson 20 reliability coefficient in excess of 1,00 were obtained. This is manifestly false, as no coefficient can have values above unity.

At the outset of the work this possibility had not been considered and the choice of a coefficient of reliability for use in my investigations was made as discussed in pages 52 to 57 on the basis of reports available in the literature. The major drawback, as discussed on pages 56 or 57 , was that the KR20 coefficient had been designed for use with a correct answer awarded one mark and any wrong or omitted answer a zero mark, and I was aware that my practice of penalising wrong marks might lead to incorrect KR20 values.

My experience has shown that higher KR20 values were obtained when correcting for guessing than for the same examination where both answers and omitted answers were scored as zero.

The problem I faced, however, was that I could find no reference to a coefficient of reliability designed to cope with this procedure of penalising for guessing. (The reasons for not adopting the test-retest and Spearman-Brown split half methods are discussed on pages 52 and 53).

Since it was my intention to examine various tests given within the department, it was obviously necessary that some measure of reliability be adopted to compare the results of these various tests. Accordingly within the limits of credibility, as set out on page 57, the KR20 was chosen as the only coefficient then available to compare these results. It is important to note that the KR20 has been used only to compare tests within the series, all of which had been marked by correcting for guessing, and all of which could therefore be expected to show a slightly higher KR20 than if correction had not been applied. Since this effect would have operated on all tests in this series under review, and since this series is not being compared to any other series, I decided that the slight distortion seen with correcting does not alter the relation of the KR20

scores/.. .. ..

scores in this series both in the examination of cycling and confidence tests since the observation under these circumstances is merely that the KR20 rises – but the extent of the rise is not measured and does not enter into the discussion.

The most important factor influencing this decision is the fact, as stated on page 52, that any coefficient of reliability is an estimate based on the concept of a true score emerging from a test, which in itself is impossible to obtain, and the knowledge therefore that any figure for reliability, however obtained, must of this necessity be suspect.

Initially, when values of above 1,00 were obtained as a consequence of using cycling in marking tests these values were held to be due to the fact that the KR20 was in fact an estimate and that these values merely reflected this disability of the KR20 reliability coefficient (and for that matter any other reliability coefficient). However, further examination of the KR20 formula (as set out on page 55) revealed that this effect is an inherent one in the KR20 formula which arises in two ways. Firstly, from the number of questions and secondly, from the standard deviation of test scores. These defects may well not have been apparent in the early investigations but emerge when test procedures are pushed to the extent that I have done in cycling and confidence marking.

One of the objectives of multiple-choice testing in our hands was to give frequent small tests to the students for the self-evaluation purposes, and the formula

$$R_{tt} = \frac{n}{n-1} \times \frac{\sigma t^2 - \Sigma pq}{\sigma t^2}$$

tends to bias the results under these conditions.

For example

where n = 2

p = .5 (50 percent of correct answers)

and s.d = 10 percent

then the KR20 = 1,98 !!!

Note that the figure of .5 for p is the highest figure that can be obtained and this

pq is/.. .. ..

pq is maximised – which would lead to the lowest KR20 value obtainable if the other constants remain static.

For self-evaluation tests we often set tests of 30 questions, and even if we allow the maximal value of p of 0,5 in the formula we find that for a standard deviation of 15 percent in students scores (which is below the mean value of 17,22 percent (Book II , Page 57) that we have obtained in cycling), then i.e.

n = 30
p = 0,5
s.d = 15 percent

then KR20 = 1,00

This will mean that any 30-item test with

a) a p higher or lower than 0,5 (50 percent correct answers to the questions set) and

b) an sd greater than 15 percent will result in KR20 values above 1,00.

It would appear therefore that a new coefficient of reliability is required to handle the improved technology of multiple-choice testing now available, a coefficient which will handle the problem outlined above and which will allow for penalising for guessing and a non-unitary marking system. (Work is proceeding along these lines and when new formulae are devised by my colleagues I hope to have the opportunity of retesting some of the data in my thesis.)

Such a coefficient has not yet been devised and until such time as one appears any worker who pushes his experimental work to the degree to which I have done may well face the problem under these circumstances of obtaining KR20 values above 1,00. It must be remembered that the KR20 values obtained were used for comparative purposes within the series and distortion will be common to all tests, thus not detracting from the effects caused by cycling and confidence marking.

2. Analyses/.. .. ..

2. ANALYSES OF DATA ON INTERNAL ANALYSIS.

A criticism of the methods of analysing the data may conceivably be made on the grounds that these methods of analysing the data that emerged from the question formats, cycling programmes and confidence marking, were entirely based on the internal analysis of the tests, and accordingly, in the absence of external checks, would be suspect.

These criteria as mentioned on page 26 are perfectly satisfactory if the validity of multiple-choice testing is established, and this criticism would ignore the attention drawn to the yearly investigation of M.C.Q. validity in our hands at length in Chapter III to ensure the establishment of validity and hence justify the use of internal criteria for examination analysis.

This criticism would also ignore a major problem in the field of testing which is the manner in which external checks are to be made. External checks can only be made on the basis of scores obtained by other methods of testing. The problems in establishing reliable external criteria by external testing means are too well known to be discussed here and as mentioned on page 34 these external methods of testing may be less reliable or less valid than the test methods under survey and thus if used may result in incorrect results being obtained. Only when absolute and true external methods of testing are established – an impossibility in fact – would the argument that external criteria are necessary in analysing the data be acceptable.

Even were this possible the practical consideration of using external criteria would be formidable as every multiple-choice test would have to be accompanied by such an external test. For this reason as mentioned on page 26 the simpler method of internal criteria was used.

3. Manipulation/.. .. ..

3. MANIPULATION OF CONFIDENCE MARKED PROGRAMME

It may be asserted that a student, if he "figures" out the system, may perform better than he actually should. This apparently innocuous statement carries within it an assumption that cannot be justified. The assumption is that any method or methods of measuring performance in the past can indicate precisely how a student will perform in the future. It is well known that in the educational field this is not so as a student's work patterns, understanding, etc. vary from time to time and it must be remembered that any test is a test of a student's knowledge at that point in time and no previous results can give a precise indication of how any student will perform in the future – so the assessment of how a student "should actually perform" can only be conjectural. In confidence coding a student can only maximize his results by knowing when to put very sure, sure or guess – and this means he must have knowledge to do so i.e. the knowledge of knowing when the answer he has given is correct on the one hand or when on the other hand it may be incorrect. Since this is precisely what is required of the student better marks arrived at by manipulation of the scheme arise only on a basis of his knowledge which is precisely what this method of testing is designed to do.

In any event the striking correlation between confidence and knowledge are such to indicate clearly that the results of confidence marking clearly reward the more knowledgeable students which should be the basis of any examination rationale.

4. The Relation/.. .. ..

4. THE RELATION BETWEEN THE PHI COEFFICIENT AND ITEM DIFFICULTY

The curious relation seen between the phi coefficient and item difficulty as shown on pages 77 and 78 Book II, that is the loss of discrimination seen as an item deviates from an item difficulty of 50 percent is extremely interesting. It has been suggested that two explanations may account for this –

1) that the effect is inherent in the ϕ formula or that,

2) the easy questions discriminate only among the poor students and the difficult questions only among the best and the number of students in these tails are few.

This phenomenon was only observed late in the investigation that I undertook and then quite fortuitously while preparing graphs for publication. I am of the opinion that the emergence of this phenomenon as a result of my investigations is significant. Further work is now indicated to elucidate the reason for the phenomenon , especially in the light of recent questioning of the phi coefficient (Melzer C.W., 1978 -- personal communication).

The demonstration of this phenomenon represents an advance of our present knowledge, and it is of sufficient importance to be published at this time. I trust this publication will stimulate enquiry into the causes of the phenomenon and thus a further advance of knowledge. I did not consider that the presentation of my work, which is already of an extensive nature, should be delayed by a further inquiry into the causes of this phenomenon which enquiry can well be pursued as an independent investigation by myself or other workers in future.

5. Conclusions/.. .. ..

5. CONCLUSIONS OF CHAPTER IV, (PAGE 68).

Conclusion 4 states that individual item difficulty has not affected test reliability and Conclusion 5 that test difficulty appears to relate to test reliability. These two conclusions may well appear inconsistent, but due attention must be given to the way in which these statements differ and how they have been arrived at.

Conclusion 5 applies to test difficulty (as a whole and not to item difficulty), and was arrived at by the correlation of reliability and test difficulty as set out in Tables R4 and R5 – pages 40 and 41 (Book II), and refers to test reliability as seen in non-cycled tests. Conclusion 5 was formulated to express this relation. This relationship however is not seen when the positive step of increasing test reliability by cycling is undertaken when as it can be seen in Table R17 on page 58 (Book II) that there is no correlation between test difficulty and test reliability under these circumstances. In retrospect Conclusion 5 should be re-formulated as :

"CONCLUSION 5 – Test difficulty appears to relate to test reliability initially, but this relation is not seen when positive action to improve reliability by cycling out items of poor discrimination is carried out."

Conclusion 4 was formulated as discussed on page 66 on the fact that items are removed on the grounds of their discrimination ability alone, and this conclusion refers to items alone and not their combined test difficulty.

However, if Conclusion 5 is re-formulated as above the initial inconsistency disappears and certainly from a practical point of view there is no inconsistency in that an examiner in seeking enhanced test reliability can achieve this by attention not to test difficulty but only to a) individual item discrimination and, b) the number of items in the test.

6. Question/.. .. ..

6. QUESTION FORMATS

In the analysis of the performance of question types only 6 question formats were selected for investigation.

My work has concentrated upon the controlled observations of student performance when confronted with six differing formats. The difference of performance and their measurement has been the extent of my work.

Any examiner may question or reject any one or all of the question types that we are using in our department, but until this rejection is shown to be based on objective measurement of student performance the action of the examiner remains based on personal prejudice. He may eventually be shown to be correct – but until the proof is forthcoming the opinion is entirely personal and intuitive, and must be seen as such.

With reference particularly to the wrong-from-five format it may well be asserted that this calls for an inversion of thinking on the part of the candidate. If the M.C.Q. examination is being used for "once only" examinations such as the Fellowship Entrance examination in the United Kingdom where examinees do not keep their question papers, there may be some substance in this attitude. In the context of multiple-choice questions as given in our department however where students are permitted to retain their question papers and to establish the correct answers for themselves it may well be, as discussed in the text, page 72, that the student is being positively influenced by the three or four correct alternatives contained in these questions compared to the four false alternatives in the usual correct from five format.

However, as discussed in the thesis, on page 84, the results in our hands for the wrong-from-five questions have not been satisfactory as far as discriminative ability is concerned and as stated in the thesis consideration will have to be given as to whether it be dropped by the department, despite the arguments advanced for its intrusion.

A further/.. .. ..

A further test of this format is indicated in the light of my results as presented – but in my opinion the worth or otherwise of a single question format is a detail that can well be investigated using the methods of investigation presented in my thesis, to establish its reliability, or otherwise. In my opinion, the important work has been in initially setting up methods to investigate the performance of question types and the reporting of these innovative approaches.

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

The experiment of using multiple-choice tests in the Department of Anatomy grew from two concepts that had arisen in the consideration of the role of teaching in the department.

1. The first was the need for an instrument to measure the comparative worth of any changes that we might wish to make in the methods of teaching.

   It was observed by Professor Wells in discussion (Wells - 1967) that, "Any change of teaching method is accompanied by an apparent but transient beneficial effect in terms of student performance". This has been called the Hawthorn effect and has been observed by educationists, but Professor Wells went further in linking the beneficial effect to the enthusiasm of the teachers for the new methodology being communicated to the students who in turn became motivated towards performance.

   The transience of the effect was likewise related to the waning of the newness of the "new" methodology and the waning of enthusiasm for the "new" methods on the part of students and staff.

   A rider can be added to Professor Wells' observations in that the degree of improvement in performance will be especially marked if the examinations are conducted by those who were responsible for the planning and implementation of the changes in teaching under consideration.

   I had become interested in the possibilities of applying the method of programmed instruction to certain basic fundamental parts of the course in anatomy, and had been involved in the planning of a revised course in anatomy for physiotherapy students, certain aspects of which I wished to carry over to the teaching of anatomy to the students in the medical faculty.

   It was obvious that in view of the Wells-Hawthorn effect, and my rider, that unless some objective method of assessment could be devised, it would be impossible to allow of valid judgements to be made as the worth or otherwise of any innovations that were being considered in the departmental teaching methods.

   Two possible methods of objective assessment available are the setting of

one/.....

one-word answer papers or the setting of multiple-choice tests. The latter method is more objective but both methods are time consuming for teachers with reasonably heavy teaching loads as exist in the department.

At this stage that I became aware of a technique that Professor A. Davison, then lecturer in the Department of Physiology and Medical Biochemistry, University of Cape Town, was using to record student answers to practical problems set in the medical biochemistry course, which were then fed to a computer which printed a list of results (Davison - 1967). If this method of recording student answers could be adapted to a multiple-choice test then the time spent in marking scripts would be eliminated and an objective comparative evaluation of teaching methods would become feasible.

2. The need for such a system was underscored by the study of the functions of examinations.

There are at least four separate functions of any examination-

a) Certification,
b) Student grading,
c) Student evaluation of their own knowledge,
d) Evaluation of teaching.

These functions are distinct but in conventional practice it is not customary for an examiner to define precisely the objectives of his examination and the examiner may attempt to cover more than one of the functions set out above. It is customary, in fact, in most institutions for the examination at the end of any course of instruction to assign grades to students who are then certified in terms of these grades, i.e. using the first two functions.

It is felt by the author that, in fact, these two purposes of examinations should be distinct and that better measurements could be made if examinations were expressly designed for the purpose in view.

Most examiners are accustomed to grade students, and allocate them a numerical mark with a pass mark at some determined level with the construction of histograms of the distribution of marks in an examination that approximate

to/.....

to normal population distributions. (Fig. 1) (Book 11, Page 1). It could well be argued that this normal distribution applies only to physical measurements or to marks awarded prior to courses of instruction, and that they are completely false when applied to the measurement of knowledge after a course of instruction. In this respect if the requirements of certification were more closely linked to the parameters laid down for the validation of a programmed course of instruction namely, that after taking a course 90% of those entering must know 90% of the facts contained therein (Lysaught & Williams - 1963) - quite different curves of distribution would arise. An example of a curve of distribution that might be obtained under these circumstances is illustrated in Fig. 2, (Book 11, Page 2). Not only would different distribution curves be generated, but the possibility of direct comparisons of teaching methods might be more readily available.

As opposed to the type of distribution desirable for certification as illustrated in Fig. 2 a very different curve would be required for a test which had the sole objective of grading the students. It is unlikely that any two students would have precisely the same knowledge about any subject. Accordingly, the more precise the measuring instrument being used was made, the more apparent would these differences between students appear. If this was the intention of the examiner instead of trying to compress 90% of the student population into a small section of the marking scale it would be desirable to have a scale that extended for as many steps as there were students. An example of the type of distribution that would be most suitable for the grading of students is illustrated in Fig. 3 (Book 11, Page 3) where it can be seen that an effort is made to utilize the whole range of the scale. The scale need not have a zero as its ordinate, and can in fact be virtually open-ended.

Hence it is apparent that the type of tests the examiner should set should depend upon his objectives, i.e. whether he wishes to measure individual student differences for the purpose of student grading or whether he considers these relatively unimportant and has as his objective certification. Only by the clear separation of these functions can the <u>best</u> tests be devised as the desired end results in terms of marks, expressed as curves of distribution, are entirely different for these two objectives.

But/.....

But what of student self evaluation under these circumstances? It is obvious that any examination whatever its function will make the student aware of his knowledge or lack of it, and every teacher is aware of the anxiety and stress that occurs in students trying to evaluate accurately their knowledge on the results of class examinations. Would the separation of examinations into certifying and grading not confuse self-evaluation when the student is exposed to two different types of tests?

I believe to the contrary that if this separation of functions could be achieved in examinations self-evaluation would become meaningful to the student. In tests designed for certification, where basic knowledge and principals were stressed, he would obtain a clear picture of minimal levels for certification based on his performance in such an examination. If this were his sole objective he would not need to concern himself with the outcome of examinations for grading. Grading examinations would not consist of the same type of basic principles as certification examinations but would examine principles and knowledge to a far greater depth. A student in a grading examination may find himself unable to answer the majority of questions. This would occasion no confusion for him since he would already have a clear idea of his knowledge for certification requirements. The results of the grading test would allow the student to evaluate his own ranking in the class and the probabilities of his attaining a 1st class, 2nd class pass, etc. in the final examination.

The third function of an examination, namely that of student self-evaluation was the main reason for the adoption of multiple-choice in our department, because of the importance of furnishing a student with feedback as to his progress if he is in any way to be helped in the attainment of his learning goals. Lennox et al (1957) in discussing class examinations state that they consider the main function of an examination to be the provision of a measure of a student's progress during the course of instruction, and therefore consider that a good class examination should provide an accurate prediction of the results to be expected in the professional examination.

Obviously/.....

Obviously there will be limitations to the accuracy of this prediction - firstly because students do not procede at a uniform rate of learning - and secondly because even the final examination "cannot provide a completely accurate assessment of the student's knowledge and attainments". (Lennox et al 1957). However, the emphasis here is on student (and teaching) in-course evaluation, and validity of the predictive examination.

In the examination schedule by conventional methods that was in operation at the Department of Anatomy the first major self-evaluation opportunity afforded the student was the class examination that was held in April at the end of the first teaching term of twelve weeks. By the time the scripts had been marked and the results collated it was usual for another two weeks to have elapsed and the student received his marks and script after approximately fourteen weeks of instruction. In a thirty-week teaching year this meant that about 46 percent of the students' learning time in the department had elapsed before he was able to get an initial feedback as to progress.

It is true that the student was required to take vivas (or oral tests) on each section of the dissection work numbering three in the first term. These vivas are, however, held on a group basis involving six or more students lasting about one hour, and while fulfilling many useful purposes are not as a rule able to cover enough ground to give the student full and meaningful feedback, especially as he is not required to answer more than one-sixth of the questions.

In view also of the fact that we were at that time called upon to increase the number of students in the Anatomy class from one hundred and twenty students to one hundred and eighty students, with no increase of staff it was also a source of concern that the vivas might prove to be even less of a self-evaluation opportunity for the students.

It was felt that if it were possible to offer the students some sort of self-evaluation opportunity in the early days of their course remedial action by the students and teachers could be taken well before the halfway mark in terms of teaching time had been reached. It was essential that any such self-evaluation opportunities should create minimal demands on the teaching

staff/.....

staff who were already carrying severe loads imposed by the staff/student ratio of one to twenty.

There is another factor in student self-evaluation that arises when the hand marked script examination is considered. In order for self-evaluation to be meaningful speed of response in the educational feedback process is essential, so that mistakes can be quickly and easily corrected. This is the underlying principle used in the philosophy of programmed instruction where student mistakes are instantaneously ascertained and corrected (Lysaught & Williams - 1963). In a written essay examination the elapse of perhaps fourteen days will mean that the student will have forgotten the reasons for making his mistakes, and since the theoretical and practical instructional program will have moved on, the student will have lost interest in what has become a "stale" section of the work.

A further factor of concern in the suitability of the self-evaluation value of essays is the well known variability that occurs in the marking of examination scripts. This subject has been dealt with by many investigators, in particular Bull (1956) and Lennox and co-authors (1957).

It appeared therefore that the need for frequent self-evaluation opportunities in the early part of the course coupled with the need for speedy response to the students and the requirement of minimal time involvement on the part of the staff, could best be met by machine-marked multiple-choice tests.

Despite the fact that at the time of this decision multiple-choice examinations had been in use in the United States for about forty years (Sinclair - 1953) there was a dearth of reports of these techniques in the literature available to me. The majority of reports available suggested that the authors were in favour of the philosophy and techniques of multiple-choice examinations. In an editiorial (Ed. - 1967) of the British Journal of Medical Education the following appeared -

> "There is now clear evidence that multiple-choice examinations offer a means of measuring a range of an individual's knowledge at a particular time with greater accuracy

than/.....

any previous method".

The Editor went on to say -

" "Examinations in medicine have other (apart from licensing) and equally important purposes. Good or bad they exert a powerful influence on education and can be used as educational methods, a principle underlying programmed teaching. They can and should be used to tell both students and teacher what progress the individual is making, and to reveal his particular strengths and weaknesses, so that special attention can be paid to both".

Other favourable comments on multiple-choice included -

1) Cowles and Hubbard (1952).

"The objective examination when carefully and diligently prepared, appears to offer a more reliable and valid evaluation of the students knowledge and his ability to apply that knowledge to the situation in hand than can be obtained from the time honoured essay examination".

2) Sinclair (1953).

"The objective paper allows a more extensive sample to be taken of the students knowledge", (vis-a-vis essays).

3) Lennox et al (1957).

"We do not now maintain that the objective paper tests only practical knowledge. It has proved a matter of little difficulty to devise questions whose answers are readily deducible from a combination of elementary factual knowledge with a reasonable grasp of the principles".

4) and Stokes (1967).

"The most cogent argument in favour of substituting multiple-choice for the essay-type questions comes from the

great/.....

great explosion of medical knowledge over the last few decades. Most of those in charge of devising medical curricula are aware of the immense strain this throws on students and are beginning to recognize that it is no longer possible to achieve comprehensive instruction in all disciplines. If we accept this as a trend which cannot be reasonably resisted, it follows that the fewer subjects that are covered by questions in the written part of the examination the greater the disadvantage at which the student finds himself. This disadvantage is only fractionally compensated by offering him a choice of essay questions. Multiple-choice questions can test a much wider range of knowledge in a comparable time and have the advantage of being machine-scorable, thus avoiding the pitfalls of examiner correction.

A candidate will probably not be able to answer more than perhaps 50% of what he is asked, since gaps in knowledge are an inevitable accompaniment of training in contemporary medicine, but he will have a better chance of showing what he does know.

Questions can be designed to test not only factual recall but also association of ideas and, to some extent, judgement".

If the latter could be achieved many of the criticisms of multiple-choice could be answered.

2. ADVANTAGES OF M.C.Q.

The advantages that could be attributed to multiple-choice testings may thus be summarised -

1. In a given time it can sample the students knowledge over a greater variety of subjects than can be covered in the same time by other tests.

   A student who has missed a lecture or a couple of pages in his textbook stands the chance of being penalised in an essay examination with its limited subject coverage (Lennox et al - 1957). M.C.Q. with its width of coverage minimizes this chance.

2. The width of coverage of multiple-choice tests could also help in discovering defects in the teaching program.

3. It emphasizes recognition of facts rather than the art of recall.

4. It emphasizes the facility of thought and the storage of information - the latter two points being critical in the evolution of a trained observer.

5. Scoring is entirely objective and thus the variation inherent in the marking of examination scripts is eliminated.

6. Students are more certain about what is expected of them in multiple-choice questions (Cox - 1972).

7. Multiple-choice examination can be used to appraise the achievement of objectives in a course. (Cox - 1972)

8. The marking entails no involvement of time for the majority of staff in a department and the dull routine of scoring scripts is eliminated.

9. Staff would be able to devote this time to the more exacting and challenging task of devising appropriate items for future objective examinations.

10. It/.....

10. It has marked advantages in that results and scripts can be returned to students with 24-48 hours, which speed is an essential requirement of a meaningful self-evaluation opportunity.

11. Statistical data becomes available from machine-scoring which can be used for the analysis of the measuring instrument. Similar data is virtually impossible to obtain from the scoring of essay scripts.

12. The data that is obtainable can be accurately and intensively assessed leading to recurrent revisions of the questions permitting steady improvement of these examinations. (Cowles and Hubbard 1952).

DISADVANTAGES OF M.C.Q.

Naturally enough there are disadvantages to using multiple-choice questions, which may be summarised as follows -

1. It cannot be used to test the degree of skill acquired by a student. These skills are tested by the essay examination which "tests the ability of the candidate to comprehend simple written instructions at a moment of stress, and, having done so, to apportion his limited time judiciously. It tests his factual recall, and his power to select the most important relevant facts of which he is master. It tests the speed and precision with which he can arrange and set down these facts in logical order, and his capacity to develop with them a reasoned argument or a concise description. It tests his ability to write, spell and punctuate his own language legibly and fast, his capacity to use and understand the language of science, his neatness and tidiness, and, finally, the ingenuity with which he can conceal his own ignorance". (Sinclair - 1953).

2. Multiple-choice requires the ability to read quickly - which may place a candidate at a disadvantage if he is taking an examination in a language other than his native one. (Whether this would be more of a handicap to a student that the exhibition of linguistic skill required in an essay is debatable.)

3. It/.....

3. It trains a man to think in terms of correct answers rather than evidence. (Brooks - 1961)

4. Subjectivity on the part of the examiner may be masked. If set by an individual an item may reflect the bias of an examiner. Pullias (1937) found a low correlation between the marks of the same students in objective examinations designed to cover the same subject but set by different examiners. (The role of the panel of examiners discussed in Chapter 11 is of major importance in counteracting this effect).

5. A considerable amount of time is required for the preperation and selection of suitable questions.

6. The evaluation of statistical analysis is likewise tedious and requires the presence in the department of a teacher who is able and willing to undertake the necessary investigation of the statistical data presented by a computer for each test, and constantly revise material in a data bank.

7. If adequate statistical analysis and revision of material in a data bank is not undertaken and tests are not subject to rigorous examination, the possibility exists that multiple-choice testing could deteriorate to the extent that tests were giving students negative information for self evaluation and could thus be harmful rather than beneficial.

8. For any department with an active multiple-choice testing routine data must be accumulated in data banks which may become difficult to handle and keep up to date.

Criticism of the multiple-choice examination based on the ambiguity of questions and the desirability of deliberately formulating false facts about the subject, etc. have also been voiced. (Gibson - 1969 , Banesh Hoffman - 1962)

These/.....

These disadvantages, however, did not appear to preclude this type of testing and Karsner (1961) appeared to have no reservations in stating "the multiple-choice examination as now conducted by the National Board* is superior from every point of view to the essay examination".

In view of the fact that the advantages for our purpose appeared to outweigh the disadvantages it was decided to implement multiple-choice tests for student self-evaluation purposes in 1967.

* The National Board of Medical Examiners, Philadelphia, U.S.A.

3. THE SCOPE OF THIS INVESTIGATION.

After a preliminary test in 1967 of a computer-marking program that I had devised (See Chap. 11) the experiment of presenting the students with a multiple-choice test for self-evaluation purposes was commenced in the academic year of 1968.

Soon after the programme was put into effect Professor L. H. Wells, then Head of the Department, examined the results of these tests and their correlation with the marks obtained by students in essay class tests, and decided to include multiple-choice components in all class examinations, as well as presenting the students with multiple-choice self-evaluation tests. The factors that motivated this decision were -

1. The reasonably high correlation between multiple-choice tests and essay examinations, which varied between .587 and .774 and which were all highly significant for the number of students in the second year medical classes. (n = 120-180)

   and

2. the conviction that if a student acquired a skill during the year (in this instance skill at answering multiple-choice examinations) it was only fair that he should be given some opportunity to display this skill as part of his certifying examination. (Well - 1968).

The evolution of the marking programmes from 1968 to date are discussed in Chapter 11.

Since in staff discussion certain questions had been raised regarding the performance of multiple-choice tests, it was decided to investigate the results of the multiple-choice examinations available to answer the queries raised.

The questions to which this enquiry attempts to provide answers are:-

1. Is the examination of students by means of multiple-choice tests a valid measure of the students' knowledge in the subject of Anatomy?

2. Are/.....

2. Are multiple-choice tests a reliable means of examination?

3. Does the difficulty of a question relate to its ability to discriminate between the more knowledgeable and less knowledgeable students?

4. Does the format of a question affect the ability of the question to discriminate between the good and poor student?

5. Is the ability of the student to respond correctly to a question related to the format in which the question is presented?

6. Does confidence weighting of the students response increase the reliability of multiple-choice tests?

7. Is it possible that the results of confidence weighted scores can be influenced by student behaviour in the allocation of confidence codes?

The methods used in investigating these problems will be described in detail in the relevant chapters (Chap. 111 - V11).

Since the possibility of controlled experiment rarely offers itself in the field of examination it follows that the introduction of a new technique demands careful observations over a prolonged period. (Hobsley - 1976) The material used were the answers recorded by students who wrote the examinations set in the department in the years 1968 to June 1975, representing 7½ years of experiment. Results were available for 32 tests in this period of time and the student numbers approximated 170-180 per test. For some of these tests student answers were no longer available, but there was available in most instances data from the routine analysis carried out on each test.

In the expectation that the department was improving its ability to set questions as a result of experience gained, the investigation is weighted towards the more recent part of the sample when circumstances warrant, or when necessary.

CHAPTER 11

THE DEVELOPMENT OF THE MULTIPLE-CHOICE QUESTION MARKING PROGRAM.

1. THE FORMAT OF THE QUESTIONS.

The first steps in establishing M.C.Q. in the Department of Anatomy were taken, if not entirely in the dark, with very little illumination indeed. There was no department at the University of Cape Town, nor in South Africa, using this technique and we were forced to create our test model on the basis of published data.

Hubbard and Clemans (1961) had established a model for the National Board of Examiners based on a stem followed by five alternatives of which one was considered "best", the answer, and the others not so correct or wrong - called the distractors. The candidate's task is to select the best or correct answer from the alternatives presented and hence this format has become known as the one from five format. Hubbard and Clemans classified as many as ten varying ways in which the one from five format can be presented to the candidate ranging from the presentation of a stem followed by five completion alternatives to case histories, diagrams, etc.

This one from five format was also used by one of the pioneers of M.C.Q. testing in the United Kingdom - Professor Lennox (Lennox - 1967-a, 1967-b, Lennox, Anderson and Moorhouse - 1957, Anderson, Lennox and Low - 1964, Anderson, Dykes and Lennox - 1965-a, 1965-b) who calls it a simple system and who stated that while he had experimented with other systems, had "on the whole found the simplest the best" - (Lennox - 1967-b).

Among other alternative formats that presented themselves was the True/False format - one from two, in which the question is presented as a statement which is either true or false. This format tends to optimise the guessing or gambling tendency inherent in the students' approach to multiple-choice examinations which is present whenever the student encounters a don't know situation. For this reason it was felt that this format should not be presented to the students initially. There are additional problems associated with the scoring of this format which will be dealt with later.

Another format, which was at that time becoming increasingly popular in England, is the indeterminate question in which the number of correct alternatives is not limited to one. Theoretically this format may be more desirable in that the

candidate/.....

candidate is not assisted by the fact that there is only one correct answer. The main difficulty that presents itself in this type of question is that each of the five alternatives are unlikely to be right or wrong to the same extent and that to make this type of question fair to the student there should be different values of rewards or penalties for each alternative. While this commends itself and may be regarded as perhaps an ultimate in testing, the practical difficulties presented are virtually insurmountable.

Not only would the panel of examiners be requested to reach agreement on which of the answers were true or false, difficult enough on some occasions, they would have to reach agreement as to which were most important, to grade the correct answers in a positive order and the wrong answers in a negative order and finally come to agreement on the numerical scales to be assigned to these ascending and descending orders. The magnitude of the task would completely nullify one of the major requirements of M.C.Q. testing in that the setting of items should be as expeditious as possible so as not to cause unnecessary waste of the staff teaching or research time.

This scheme of positive or negative scaling might be feasible if it were decided to accept the class responses as the basis of the numerical scaling factor. The system could then be computer-linked and an automatic dynamic scale be adopted which would be set by the computer as a result of class responses. We have not experimented with this as yet.

It was for these reasons that the indeterminate format was initially rejected and the decision was arrived at to standardise the examination by presenting the student with 5 alternatives, one of which was the correct or best answer.

There are various formats by which this can be accomplished and examples of formats that have been used in our examinations are set out in Appendix A (Book 11).

2. THE FUNCTION OF THE MODERATOR PANEL

Earlier in this investigation it became obvious that if the contribution of multiple-choice items were left to individuals certain grave defects might occur -- namely

1. Ambiguity might occur in the content of the question,
2. Incorrect or controversial alternatives might be furnished as the correct answer,
3. There might be bias in the examiner's choice of facts or subjects to be examined,
4. The question might not reflect a reasonable expectation of what the student was expected to know.

Accordingly a moderator panel of teachers in the department was set up and it has become standard procedure that all questions to be submitted to students for any examination procedure should first be submitted to the moderator panel.

This procedure has proved more than warranted as the majority of questions submitted are altered or annotated in some way so as to render them a fair test of departmental consensus of expected student knowledge. In this way the defects of questions mentioned above have been kept to a minimum - but naturally still do occur and will be reflected by scrutiny of the analysis of the examination which is a by product of the computer program in use (vide infra).

3. SCORING THE EXAMINATION.

Since the questions to be presented were of the standardised one-from-five variety and in view of the fact that the indeterminate format was not to be used it appeared reasonable that all questions should rank the same and if correctly answered should score the same. A problem to be faced was whether a student should be penalised for a wrong answer. The conventional method of handling this problem is that when a candidate's answer is incorrect it is penalised by the statistical amount that might have been achieved by random chance for that question, which is governed by the number of alternatives in that question, that is the candidate's freedom of choice. This would be 1 in a True/False item, 2 in a 3 alternative item, 3 in a 4 choice item and the mathematical correction thus applied is $\frac{1}{n-1}$ where n equals the number of alternatives in that question.

While the penalty may appear correct for random guessing it in fact does not apply to the more probable situation in which a candidate has probably as a result of his knowledge been able to eradicate 3 out of 5 alternatives as being incorrect but is uncertain which of the two remaining alternatives is correct. He has, in fact, one freedom of choice, that is he has a 50% chance of being correct, and in the event of being incorrect he will be penalised by one-quarter of a mark. The odds in this situation are 4 to 1 in the candidate's favour and he would be ill-advised not to attempt a guess if he were in this position. To a lesser degree the same weighting of odds in the candidate's favour would operate whenever he was in a position to guess at the answer from a shorter list of alternatives than set by the examiner with the penalty for incorrect answers determined by the formula from the longer list of alternatives.

As it is anticipated that any candidate presenting himself for an examination should have some degree of knowledge the conventional model based on the statistics of chance does not apply, and it is to be anticipated that candidate's scores will be enhanced to some degree by the probabilities of scoring successfully presented by the basis of their knowledge. If no penalty is imposed for an incorrect answer guessing is encouraged. Experimentally it has been shown that when candidates are encouraged to guess significant increases in examination scores result. (Cooper and Foy - 1967 and Sanderson - 1973).

Despite/.....

Despite the loading of the formula in the candidates favour it was decided to penalize a wrong guess in order to discourage unwarranted guessing and using the conventional formula this would entail a penalty of -0,25 for a five-choice item.

Further in assessing the scoring system to be adopted we were faced with a decision as to whether the student was to be allowed to indicate whether he felt he did not know the answer to any question. Since the marking system included a penalty for a wrong guess, as discussed above, it was considered only fair that a student should not be forced into a response as this might have been regarded as unfair to the student. In view of the fact that in most cases a guess was more likely to be correct for the stronger candidates than for the weaker students, it was felt that this would weight the examination for the stronger student, if the students were forced to answer all questions. This impression has been borne out by the analysing of later examinations marked by confidence weighting in which it was found that students in the first quintile were correct 47% of the time they guessed compared to students in the fifth quintile who were right 38% of the times. (See Analysis of Confidence Marking, Chapter V11)

The report by Sanderson (1973) has confirmed not only the tendency of students to score higher marks when encouraged to guess but confirming our earlier impression that guessing, in fact, favoured the abler candidate.

Accordingly it was decided a correct answer would add 1 mark to a student's score, a wrong answer would be penalised by 0,25 marks, and no attempt by the student to answer would occur no penalty. This scoring schedule was used to mark the tests which have been analysed in the determination of the validity and reliability of multiple choice tests (discussed in Chapters 111 and 1V) and the relationship between the various item formats (discussed in Chapter V).

4. LANGUAGE AND M.C.Q.

It was our impression that students who were fluent speakers and whose home language was English were at a distinct advantage in that the setting of the M.C.Q. questions imposes a semantic restraint not only on the examiner but also the examinee. This appears to be in some measure confirmed by the studies of Young and Gillespie (1973) who observed a pass rate of between 50 and 60% in the M.C.Q. section of the primary fellowship examination in Glasgow among those candidates whose home language was English compared to a pass rate of between 10% and 20% among those whose home language was not English. But as the latter group included candidates from Asia, the Middle East, East and West Africa, as well as Europe - and it may well be that factors other than language alone might be operating to affect the pass rate of this group.

Of particular interest is the popularity of the examination found by Young and Gillespie among the candidates of whom 85% expressed a preference for this type of examination, and of whom 78,5% were in favour of the M.C.Q. examination being established as a standard, excluding candidates from further examination by way of orals.

It has not been possible to measure the effect of the home language of the student on the performance in multiple-choice tests in this series. Whether this factor plays a greater part in answering multiple-choice questions than essay questions, in anatomy at any rate, remains unanswered and might be worthy of subsequent study.

5. THE COMPUTER MARKING PROGRAM.

The essential requirement of a self evaluation opportunity which would prove of meaningful assistance to the student was speed in reporting back the results to the students as only then would they be possibly able to

i) recall their answers and more important

ii) the reasons for making these responses while the subject matter of the test was still fresh in their minds and they had not gone stale on the subject or proceeded to another section of study, and

iii) mistakes of detail or concept be corrected at a relevant stage of the learning process.

If a batch of multiple-choice questions from 150 students is marked manually marking time becomes an important factor and it is virtually impossible to process the test and return scores to the students under seven to ten days. (Kaplan - 1971)

If in addition any statistics regarding the performance of the questions is required, and in my opinion reliable multiple-choice tests cannot be constructed without the performance of items being recorded, the time required for adequate marking multiplies greatly and together with the time required for the construction and selection of items enlarges beyond the scope of any staff member who is also required to carry any sort of teaching load.

Initially therefore the marking program was devised and written for a first generation I.C.T. 1900 computer in M.A.C., (Manchester Auto Code) and was a simple program which

i) marked the students scores

ii) adjusted the marks for possible guessing (by the formula discussed above),

and expressed the corrected score as a percentage, analysed the questions very simply as the number and percentage of students who got it right, the percentage who chose a particular alternative and the percentage who made no attempt at all (see Scoring the Test), and finally printed a histogram of the distribution

of/.....

of class marks standardised for a class of 100 students.

As pointed out by Harris and Buckly Sharp (1968) in the design of a computer program the programmer at that stage in the development of computers was faced with two alternatives :- Either to limit the amount of data to be processed to the size of the internal core storage of the computer, or to make use of additional drum storage facilities which in the first generation computers had the effect of slowing down processing time enormously. However slow the computer it was still nevertheless immeasurably faster than manual marking and at no time in program design were the considerations of speed allowed to influence the amount of data, within reason, that was required. Harris and Buckly Sharp published a computer scoring program written in Fortran for a second generation computer which was for me of academic interest as we did not possess the facilities required.

In 1970 the university acquired an IBM 1130 second generation computer which had a Fortran compiler and this brought immediate benefits in that we were able to speed up the marking process and increase the amount of data that could be stored and processed in a computer run. This represented the second phase of our computer programming, which enabled us to produce a more sophisticated program which could not only handle the task of allocating student scores and basic item performance but could analyse item performance and enabled us to institute routines for the measurement of item performance and test reliability that are vital to the construction of valid tests.

At present all computing is done on the University's Univac 1100 third generation computer with data processing speeds up to one hundred times faster than the processing speeds of the 1st generation computers and an internal storage capacity over two thousand times as large. In addition the speed of transfer to ancillary storage discs is so rapid as to ensure that for practical purposes classes of upto 1000 students and tests up to 300 questions can be handled in less time than it took to read the data into the first generation computer. The present marking program - the U.C.T. Model Marking Program - for general use in the University was designed by a committee of interested departmental users on the model of the program that was initiated for use in the Department of Anatomy. The most

complicated/.....

complicated of the marking programs that we have yet devised takes less than five minutes for a complete run including marking, program analysis and re-marking in cycles dependent on parameters generated by the analysis as against nearly forty-five minutes required for the simple marking program devised for the first generation computer.

6. THE MEASURE OF QUESTION VALIDITY AND TEST RELIABILITY.

The question of test reliability is discussed in detail in Chapter IV and will not be considered at this stage.

Question Validity.

The validity of a test as a whole rests upon its ability to distinguish between the more knowledgeable and the less knowledgeable examinees or between good and poor students. This same test will apply to each individual question or item in the test and if most of the good students respond correctly to the item, and most of the poor students do not the item will be valid and have the ability to discriminate between good and bad students. The determination of what constitutes a good or bad student can be made either on student scores obtained in other testing circumstances (external) or on scores determined on the test being marked, (internal). It is a far simpler matter in analysing the validity of items to use the results of the test being marked than to key in separate data to establish good and bad students and for this reason we have used the internal students results (i.e. results from the test being analysed) in item analysis. The discriminative ability of items may differ if internal or external criteria are used - but if the test as a whole is valid (See Chapter 111) then the discriminative ability of items in that test even if based on internal consistency will likewise be valid.

The two most popular measures of item discrimination are:

i) biserial and point biserial correlation coefficient, and

ii) the phi coefficient.

i) The former are unsatisfactory since biserial correlation coefficient relies upon a normal distribution of student scores - which is seldom observed, and the point biserial correlation coefficient value is influenced by the item difficulty in that the maximum possible value of the point biserial correlation coefficient decreases as the item difficulty moves from 0,50 up to 1,00 or down to 0,00.

ii) For/.....

ii) For these reasons I have used the phi-coefficient as the measure of the discriminative ability of a question.

This coefficient is arrived at by comparing the performance in that item of the upper half of the students to the lower half of the students by the use of the following formula

$$\phi = \frac{pu - pl}{2\sqrt{pq}}$$

where pu = percent getting item correct in the upper half of score distribution.

pl = percent getting item correct in lower half of score distribution.

P = arithmetic mean of pu and pl

q = 1 - p.

7. THE RECORDING OF STUDENT ANSWERS.

Input of the data to the computer presents the major difficulty in the execution of any automated marking procedure.

Three alternative methods are available:

1. The simplest is for the student to record his answers on a sheet of paper bearing his name and number which is then manually punched by trained punch operators. The possibility of mispunching and recording incorrect data is a real one; but this is reduced by the practice of a second operator verifying the data punched by the first operator. Mistakes do occur and it is important to check the punched data against the students answer sheet in all borderline cases so that care is taken that no student might emerge with a fail mark due to punch errors. (Provision for this is included in the U.C.T. Model Marking Program)

2. The second alternative is for the student to mark his answers on a specially designed sheet of paper which is fed into a mark sensing machine which then punches the data cards. (Anderson, Wood & Tomlinson - 1968). Mark sensing machines are also available which record marks on specially prepared computer cards and repunch the data on to normal data cards. (Harris & Buckly Sharp - 1968)

Despite the apparent attractiveness of these two methods in removing the possible human source of punching errors in practice they have been found to give rise to a considerable number of errors, at a rate of between 0,11% and 0,39%. (Killcross - 1968) If this represents the published figure it might well be that actual errors are in considerable excess of this rate. The major danger of this system is that errors are likely to remain undetected as it becomes extremely tedious for the examiner to check the student's responses in this form against the recorded data.

Another factor which does not appear to have received recognition is that the

adoption/.....

adoption of either of these two methods is calling into effect another factor in that the examination now is measuring not only students knowledge but technical skill of the student in marking cards as well. If an examiner wishes to measure technical skill that would be a valid objective, but there are far better ways of doing it, and the problem is that the measurement of knowledge is now not clear - but will be masked as only those with absolute skill will be in a position to have their knowledge fairly tested. Any imperfections in a students technique in marking the sheet will intervene between the measuring tool and the student's knowledge and can thus only debase the quality of measurement made.

We would however recommend these methods for self evaluation tests where speed of response is more important than accuracy of mark at the 0,30% level.

3. The third alternative is to provide the student with a specially prepared prepunched card. This card is so designed that with the aid of a toothpick the student can dislodge a small rectangle of cardboard, equivalent to the size of cardboard removed by a punch machine, and the card when punched by the student can be fed directly into the card reading hopper of the computer.

This method was being used by Dr. A. Davison, then in the Department of Physiology, for recording answers to problems in biochemistry practical classes and home tasks. (Davison - 1967) Problems do occur in that it is impossible for the student to remove the rectangle of cardboard as cleanly as a punch machine and small projections might remain. These cards are slightly thicker than normal cards and thus these projections can cause a card jam in the computer reader. At low reading speeds this is not serious - as the card can be removed and repunched by the operator, but in modern high speed card readers up to twenty cards can be involved in a jam and the first few cards are so torn as to make it impossible for the data to be repunched.

A further argument against the use of these prepunched cards is that the examiner is creating an additional stress situation for the student, who as a result of being in an examination, is already carrying a sufficient load. Not only does the student have to cope with the task of reading and understanding the question,

considering/.....

considering the alternatives and choosing the answer, (as well as deciding upon the degree of confidence in his answer if this is a confidence weighted examination, see Chapter VI) all in the space of one minute, but he now has the additional task of correctly identifying the desired box in the card and punching it out cleanly. Both of these tasks are themselves a measure of skill, and again the situation arises when the measurement of the student's knowledge may be obscured by the degree of skill the student shows in exhibiting his answer. This skill can be greatly affected by a stress situation, especially should a student make a mistake in punching the last answer in his card and be forced to repunch another complete card with the resultant anxiety about the necessary loss of time and the possibility of making another mistake and being forced to repeat the whole procedure yet again.

When compared with the task of correcting a wrong answer by merely crossing it out in the requisite box and writing the correct answer the perspective becomes clear. For this reason we have standardised a simple form on which the student marks his answers and these answers are then captured on punch cards by the operators of the University's Computer Centre. This form is used on all multiple-choice tests.

If a mark scorer were available we would use it for self evaluation tests, but retain the form referred to above for certifying examinations where accuracy of marking is more important than speed of response.

CHAPTER 111

THE VALIDITY OF MULTIPLE-CHOICE EXAMINATIONS IN THE DEPARTMENT OF ANATOMY.

1. QUESTION TO BE ANSWERED.

The Question to be Answered is:

Is the examination of students by means of multiple-choice tests a valid measure of the students knowledge in the subject of Anatomy?

2. INTRODUCTION.

The initial problem that had to be solved when I started experimenting with Multiple-choice examinations was to establish whether they could be used to measure students' knowledge of Anatomy. We were fortunate in having in the department at that time a convinced sceptic who observed with conviction "You can't tell me that you can examine anatomy by means of multiple-choice tests."

Reference to the literature available at that time was of little help. Hubbard and Clemans (1961) mentioned the use of multiple-choice papers in the examinations in Anatomy for the National Board, but gave no references to validity studies.

In order to answer the question posed it became necessary for us to examine the validity of the tests given in our own department.

If any test is examined two qualities emerge (Ebel - 1972).

1. The consistency with which a test or a set of tests measure whatever they do measure. This is defined as the reliability of the test, and this is discussed further in a later chapter.
2. The accuracy with which a test or set of tests measures what they ought to measure. This is defined as validity.

Ebel (1972) presents an explanation of these concepts in the following terms "If the perforations on a target made by successive shots from a rifle are all clustered closely, the rifle is performing reliably. If those perforations are all clustered in the bulls eye, the rifle is also performing validly."

As will be discussed later (Chapter IV), statistical measurements of reliability are readily obtainable but validity does not lend itself to such quantitive analysis. In the first instance validity will depend to a large extent on what the objective of the test is, that is, what has it been designed to measure. Does it purport to measure knowledge at a given time, does it purport to measure what the level of performance at a later date will be, does it purport to measure the entire content of the subject , and so on.

Ebel/.....

Ebel (1972) has classified 10 types of validity from the extensive literature and discussions on the subject.

1. Concurrent. The relation of test scores to an accepted contemporary criterion of performance on the variable that the test is intended to measure. This is the definition accepted by the American Educational Research Association. * The problem here is that time-honoured comtemporary criteria of performance may be less reliable or even less valid than the newer testing method under survey. This aspect will be dealt with in greater detail in the discussion of the results of analysing the data as to the validity of M.C.Q. examinations.

2. Predictive. That is the ability of the test to measure performance in some other variable (the criterion) at a later stage. This type of validity is also referred to as criterion-related validity. (Brown - 1970) This aspect of validity has not been analysed in this work, but may be pursued at a later date.

3. Construct. This is concerned with what psychological qualities a test measures. This aspect of validity was considered to be outside the scope and experience of the author and was not pursued.

4. Content. Does the test content adequately sample the specified universe of anatomical content? Content validity has by design not featured in the tests given by the department. Self evaluation tests are set at regular intervals, and are limited to the content of the instructional course at that moment, and hence will not be validated if measured against the entire content of the instructional course in anatomy. In certifying examinations M.C.Q. tests were always given in conjunction with essay questions and it was policy to exclude those subjects which were being examined by essays from the coverage of the multiple-choice tests so that the students would not be cued by having the M.C.Q. alternatives displayed before them.

* American Educational Research Association - Technical Recommendations for Achievement Tests (Washington, D.C. A.E.R.A. 1955) Page 16.

During/.....

During the construction of a final certifying examination (which examinations have formed the basis of the study into validity) the essay portion of the examination was set first and the content not being tested was allocated to the M.C.Q. for examining purposes. An attempt was then made to obtain overall coverage of this remaining content by allocating a certain number of M.C.Q. questions to each section of the course, but on occasion satisfactory validated items were not available to the panel of examiners, with the possibility that certain content of the course was not adequately examined in our multiple-choice examinations. Accordingly no statistical evaluation of the M.C.Q. tests as to content validity has been attempted.

5. Curricular. This is determined by examining the test and evaluating whether it is a true measure of the important objectives of the course. All tests exhibited to the students, as discussed in Chapter 2, were first submitted to a panel of teachers in the department. Where specific objectives of the course had been laid down it was possible to test the items comprising the test against these objectives. Where specific objectives had not been laid down it was still possible for items to be discussed by the panel and in fact many items for any test were rejected by the panel as being too concerned with detail or not important. While the safeguarding of curricular validity has been an important function of the examining panel it has not been subject to statistical analysis, and is not dealt with in the later discussion of validity.

6. Empirical. This would refer to the relation between test scores and a criterion which is an independent and direct measure of what the tests purports to measure. Since there is no independent and direct measure of knowledge this validity cannot be assessed in the particular field of testing the achievement of knowledge.

7. Intrinsic/.....

7. Intrinsic. Involves experimental techniques, other than the correlation referred to under Empirical validity, to provide objective quantitive evidence that the test is measuring that which it ought to be. This technique again is not available for use in the field of testing the achievement of knowledge.

8. Face. Is a reference not to what a test necessarily measures but what is appears to measure. This distinction applies to testing in the psychological domain and does not enter into testing educational achievement.

9. Validity by definition. This aspect implies a corollary to the aspect of content validity in that it is an examination into the ability to handle a limited or defined area of the subject under test e.g. the practical test. I have not had the opportunity of studying this aspect as yet, but it appears a fruitful area for investigation, together with predictive validity, of the validity of the Self Evaluation M.C.Q. examination that we have been giving our students.

10. Factorial. Factorial validity is represented by a ratio between a test and the factor common to a group of tests measuring the same behaviour that the test measures. This aspect of validity is dealt with in that section of the discussion relating to multiple-linear regression analysis of our data.

All these types of validity group themselves into two major categories. (Ebel-1972)

1. Primary Validity. Where the tasks provide an operational definition of the achievement, or the achievement can be measured directly.

2. Derived Validity. Where the scores a test yields correlate with criterion scores that possess a direct primary validity.

Faced/.....

Faced with these many aspects of validity it was obvious that time would preclude a full and proper examination of each. Some of them are outside the domain of educational achievement testing and others such as predictive validity would have required additional data of student performance in their subsequent years of study which was not available to us. The reasons for the exclusion of these aspects of validity from my investigation nave been stated briefly under the classifications of validity presented above.

The validity that was of major interest to me and the one that perhaps in part answers the question that had been put was that of concurrent validity, (See (1) Page 34) and the investigation of this aspect forms the first part of the analysis of multiple choice tests in the Department of Anatomy.

In the investigation of concurrent validity there are however no means of establishing primary validity (i.e. a direct measurement of knowledge) and thus the establishment of primary validity and derived validity is not possible. It becomes necessary to postulate a third category of validity.

3. Inferential Validity. Where test scores are correlated against scores inferred to be adequate measures of whatever one is attempting to test.

3. METHODS AND MATERIAL.

As already defined concurrent validity would be evidenced by the performance of the multiple-choice examinations against accepted contemporary criteria of measurement.

For the purposes of certification, and of grading students at the end of the course into four grades of pass marks viz. First Class, Upper Second, Lower Second and Third Class, and two grades of failure, namely, supplementary grade and outright fail, the department had been examining using a battery of tests to which marks were assigned as follows:

1.	The average of class tests during the year	50
2.	A Practical examination	50
3.	Two Essay papers of 3 hours each of 3 to 4 questions per paper	200
4.	An oral examination of two 10 minute sessions with each of two different examiners	100
	Total	400

It was decided that, whatever the shortcomings of these examinations might in fact be, they had been accepted as reliable criteria on which to base student assessment at the end of the course in Anatomy, and that validity studies should be made on the performance of students in multiple-choice examinations against their performances in the conventional examinations.

Material. There was available for investigation data which I had been collecting since 1968 in the form of the marks obtained by the students for each of the components of the certifying examination in the years 1968, when multiple-choice was first introduced, up to 1974. This data was used as the basis of establishing the concurrent validity of the multiple-choice examination. There were 10 components of the examination in 1968 and 1969 and 11 components in the years 1970-1974. Student numbers in these years were 176, 183, 180, 173, 162, 174, and 178 respectively, for the seven years under review, thus allowing for statistical validity.

Method/.....

Method: The marks in the certifying examination were recorded for each component of the examination for each student. For the purposes of this investigation and all the investigations into validity each essay in both papers, as well as the practicals and orals have been treated as separate components.

Step 1.

A computer program was written which analyzed this data in the following way.

1. From the marks allocated to the question the score obtained by the student in that component was recorded. An example of this listing is given in Appendix B - CORCO L/O 1.

2. A histogram of the distribution of these marks as a percentage for each component of the examination was printed. An example of the histogram produced, that for the multiple-choice component of November 1974, is reproduced - CORCO L/O 2 (Appendix B).

3. The mean and standard deviation was recorded for each component and printed. - CORCO L/O 3 (Appendix B).

4. Each component was then cross correlated against each other component of the examination under survey. - CORCO L/O 4 (Appendix B).

5. Although this matrix enables one to see generally how the multiple choice is performing in relation to each of the other individual components it gives no information of the validity of multiple choice to the final mark that is obtained from the examination as a whole - which is the accepted criterion of performance, not the correlation to individual components. In order to obtain this information the marks obtained by the student for each component was correlated against the sum of marks obtained in the examination. In order not to influence this correlation the sum of marks obtained in the examination did not include the

component/.....

> component under survey, i.e. if the multiple choice component was being correlated to the sum of marks for the examination the sum of marks would be arrived at excluding the marks obtained in the multiple choice component. This was repeated in turn for each component of the examination. - CORCO L/O 5 (Appendix B).

In review it was considered that these correlations co-efficients might be weighted in favour of the multiple choice examinations as the class average mark was in fact made up in part of the marks obtained for multiple choice examinations given during the year in class tests, and in part by essay tests and practicals. In addition since 1971 the question in Neuro-anatomy had been examined by multiple choice and was also being included in the sum of components being correlated against the multiple-choice component of general anatomy.

Step 11.

> At this stage in the investigation the University had acquired a more powerful computer and as part of the software there was available the B.M.D. package containing the Biomedical Computer Programmes of the University of California's Health Sciences Computing Facility (BMD-1973), and it was possible using these facilities to correlate the multiple-choice component in general anatomy against all the other non-multiple-choice components.
>
> It was also possible to correlate the mark obtained in the Neuro-anatomy multiple-choice component since 1971 against the marks obtained in the non multiple-choice question components, and further to combine the marks obtained in both the multiple-choice components and correlate these against the marks obtained in non multiple-choice question components.

Step 111.

> The correlation obtained thus far are all simple correlations between one variable on one hand and the sum of a number of variables on the other hand. It was brought to my attention however (Juritz-1974) that better methods of investigation exist for determining the relation

between/.....

between one variable, defined as the dependent variable, (i.e. the one that is being investigated) and a number of others, and that more properly the relation of this dependent variable and the other variables should be investigated by the statistical methods demanded by multivariate analysis - that is, the analysis of multiple measurements made on several samples of individuals. (Cooley and Lohnes - 1971).

This technique consists of using each variable in a linear equation not in its absolute value but modified by a factor which will give the best possible relation between that variable and the dependent variable, and technique is referred to as multiple linear regression. If only some of the variables are used in the equation it is referred to as stepwise regression.

In our investigations all variables were used in obtaining correlation figures for the dependent variable (i.e. component of examination under investigation) and the other variables or components of the examination.

Using the BMD programs for multiple linear regression these results were obtained for the multiple-choice component alone (1968-1974) and both multiple-choice components (general and neuro-anatomy) on one hand and non-multiple-choice components on the other as well, (from 1971-1974).

This was repeated for the following categories of components in the examination:

i) practicals
ii) orals (combined with the marks for class vivas when they were included in the years 1971-1974)
iii) short answer questions - which was the format of the neuroanatomy examination until 1970
and
iv) essays. For this computation the essays were not considered separately but were combined in all years to form one variable.

4. RESULTS.

Step 1.

The mean percentage score and standard deviation of each component of the final examinations are set out in the Table V. 1-7 (Book 11, Pages 4-10), for the years 1968-1974. The mean percentage marks for 3 components, Class Average, Practicals and M.C.Q. are shown graphically in Figure V. 1 (Book 11, Page 11), for the years 1968-1974. Tables V.8 to V.14 (Book 11, Pages 12-18) show the correlation matrix that was obtained for each year when the variables of the final examination were correlated with each other. Table V.15 (Book 11, Page 19) shows the correlation coefficient obtained when each component was correlated against the sums of all other components in that examination. The highest correlation for each year being denoted by H in the table.

Step 11.

The results of correlating the scores obtained in the Multiple Choice component of the examination against only those components with no M.C.Q. element is set out in Table V.16 (Book 11, Page 20) where in the first column is set out the raw correlation of the Multiple Choice Question in General Anatomy to all other components of the examination (as determined in Step 1 above and set out in Table V.15 (Book 11, Page 19)) and the second column shows the correlation coefficient between the multiple choice question in General Anatomy and the sum of non-M.C.Q.components. For the latter the class average and M.C.Q in Neuro-anatomy (from 1971) were excluded from the sum of non-M.C.Q. components. In the third column are the correlation coefficients obtained when both the general anatomy question and neuro-anatomy question (both as Multiple-choice) are together correlated against the remaining non-M.C.Q. components.

Step 111.

When the scores of the components are subjected to the technique of multiple linear regression the correlation obtained between multiple-choice scores and non-multiple-choice are higher.

These/.....

These results for multiple linear regression are set out in Table V.17 (Book 11, Page 21) for both the general anatomy component (Column 1) and both components (Column 2) and compared to the raw correlation for these components to non-multiple-choice components, (in Columns 3 and 4) as obtained in Step 11 (Table V.16 (Book 11, Page 20).

For complete analysis of all components of the final examination, the results of applying the procedure of multiple linear regression to obtain the best correlation between the practical component, the oral component and the essay component of the examination as well as multiple-choice against the sum of the other components are set out in Table V.18 (Book 11, Page 22) for the purpose of comparison.

Analysis of Variance tests were carried out on the correlation coefficients of multiple linear regression of the components as set out in Table V.18 (Book 11, Page 22). These results for 4 variables are shown in Table V.19 (Book 11, Page 23) when Practicals, Essays, Class Average amd M.C.Q.'s were examined and the Table V.20 (Book 11, Page 24) when all 6 variables were examined.

5. DISCUSSION.

When the matrix of correlation co-efficients obtained in the years 1968-1974 is examined it is clear that the multiple-choice components have performed in a manner comparable to the other accepted criteria previously in use. (Tables V.8 - V.14, Book 11, Pages 12-18)

Over the years we had grown to accept the practical and the class average during the year as the best indicators of the students overall knowledge of anatomy. This relation can be seen in the matrices and it is noteworthy that the correlation between these components which was for:-

1968 - ,776
1969 - ,793
1970 - ,774
1971 - ,778
1972 - ,714
1973 - ,839
1974 - ,803

was in all years the highest correlation obtainable between any two variables, except in 1972 when it was second highest.

If these variables are taken as a yardstick it can be assessed that the M.C.Q. has slowly improved and over the last few years has been in the same order of correlation, namely:-

Year	M.C.Q.: r to Class Average	M.C.Q.: r to Practical
1968	,639	,647
1969	,647	,614
1970	,725	,706
1971	,019	,049
1972	,649	,581
1973	,769	,769
1974	,735	,738

The/.....

The notable exception is seen in the year 1971, when the multiple-choice portion of the examination in general anatomy correlated extremely badly. We have no explanation for this observation. The mean score and standard deviation of the M.C.Q. question in that year does not differ materially from values obtained in other years (Tables V.1 - V.7, Book 11, Pages 4-10 and Figures V.1, Book 11, Page 11), nor do the means and standard deviations of other questions vary markedly in that year. We had attempted to improve reliability by using questions at about the item difficulty of 0,50 as far as possible.

Notwithstanding the good results later claimed by Ebel (1972) for this procedure, and notwithstanding the data obtained later that is presented in Chapter V. (Item difficulty and discrimination) the manoeuvre proved highly unsuccessful in our hands. We were in fact not successful in raising the test reliability and certainly succeeded in reducing its concurrent validity considerably. In the subsequent years when we did not tailor the examination so severely test reliability and validity have been more satisfactory.

If we look at the correlation co-efficients obtained between any component and the sum of all other components it is again noteworthy that the class average and the practical have in all years with one exception headed the list of these co-efficients. (Table V.15, Book 11, Page 19).

In assessing these correlations the co-efficient obtained by the oral examination has been excluded since it has not been a strictly independent variable in that the examiner has had available the marks scored by the candidate or his grade achieved to that stage. (It is perhaps noteworthy that when in 1974 this procedure was altered in that the examiner at the oral had no information of the candidate's examination record to date the correlation co-efficient of the oral dropped relatively.)

The correlation achieved by the M.C.Q. to the sum of all other components has improved steadily and is now appearing with values approximating to those displayed by the class average and practical, and is significant in that $p = <,001$ (except for 1971). When the class average is removed and multiple-choice correlated only against non-M.C.Q. components, contrary to what might have been expected there is no significant change in the correlation co-efficients obtained

and/.....

and the values remain significant in that p = <,001 except for 1971. (Table V.16, Book 11, Page 20)

The effect of considering both M.C.Q. components together since 1971 when the format of the Neuro-anatomy question was changed to multiple-choice does not alter these correlations significantly.

When the figures of multiple correlation are inspected it can be seen that the correlations have been improved in all cases (except 1974) for the multiple-choice in General Anatomy and similarly for the two M.C.Q. components combined. (Table V.17 Book 11, Page 21).

Since all these figures of multiple correlation are significant at above the 99,9% level of confidence for the number of students concerned we can answer the question posed in this part of the investigation with this degree of confidence and can categorically state that in our experience, even in those early and possibly inept years, that the multiple-choice method of examination is a valid method of examining in Anatomy as judged by its concurrent validity with the other time honoured methods of examination. The improvement noted in M.C.Q. correlations bears out the prophecy mentioned when considering the advantages of multiple-choice questions as a method of examination (See Chapter1) in that this type of examination can be amended to give better results on analysing its performance.

When the correlations obtained by multiple linear regression for all components of the examination are examined (Table V.18 Book 11, Page 22) the practicals again emerge as that component with the highest correlation to all other components except in 1970 when it was supplanted by the multiple-choice component and in 1972 when it was displaced by the essay questions.

The orals have been ignored in this analysis for the reason advanced above in that they are not strictly independent variables. In 1974 when the oral was strictly independent it tailed all other components with r value of ,711.

When the tables for the analysis of variance are examined, namely for M.C.Q.'s, Practicals, Essays and Class Averages in Table V.19 (Book 11, Page 23) the F ratio is ,3063 for 3 degrees of freedom between sets and 24 degrees of freedom

within/.....

within sets. This is not significant at the 95% probability level. (Guildford & Fruchter - 1973). Similarly the F ratio for all 6 variables is ,8881 for 5 degrees of freedom between sets and 29 degrees within sets which again is not significant.

We can conclude that no significant difference is found in the performance of these variables in their performance in tests compared to other variables, and thus that all variables (or components) can justifiably be used to measure knowledge in Anatomy.

While not the object of this investigation the role of the essay questions in our examinations has been interesting. In the initial study when the essays were considered separately it was somewhat disconcerting to note the extreme variability with which the individual essay correlated

i) to the other components of the examination when correlated individually (Tables V.8 - 14 Book 11, Pages 12-18)

and

ii) to the sums of the other components of the examination (Table V.15 Book 11, Page 19).

When the essays are considered as a whole and the best possible correlation between them and the other components of the examination is obtained by the techniques of multiple linear correlation this concern is dispelled and the essay questions show correlations in line with the other components and in fact when tested by the analysis of variance did not emerge as significantly different in performing as validly as the other components. (Table V.18 Book 11, Page 21)

It has been suggested (Wells - 1976) that the explanation for this might lie in the fact that whereas irregularities in marking would appear if a single essay were being considered when all the essay marks were totalled that these irregularities would tend to concel each other out and that the sum of marks obtained for all essays would be a fairer measure than the marks obtained on any one essay. The same effect could be achieved by having more than one examiner score an essay question and aggregate the marks awarded to cancel out individual examiner irregularities (Fielding - 1973).

We/.....

We may therefore conclude that Multiple-choice tests, as well as Practicals, Essays and Orals, are valid and that their use is justified in the measurement of knowledge of Anatomy - at least under the conditions operating in our department during the years 1968 to 1974.

6. CONCLUSIONS.

1. If the conventional methods of examination are considered to be acceptable contemporary criteria of performance it is possible to conclude that the multiple-choice examination shows concurrent validity and its continued use in the examination of knowledge in Anatomy is justified.

2. There is no significant difference in validity between any of the components of the final examination as set in the Department of Anatomy, University of Cape Town - namely essays, practicals, multiple-choice and orals.

3. Figures for concurrent validity on the part of the multiple-choice component of the final examination have shown a rise over the years suggesting that departmental experience and the analysis of results is an important factor in the achievement of satisfactory valid results from the use of this technique.

CHAPTER 1V RELIABILITY

1. QUESTIONS TO BE ANSWERED.

Validity of multiple-choice having been established the next question that required answering was:

a) Are multiple-choice questions a reliable method of examining students in Anatomy?

b) A corollary of this question is what factors increase the reliability of multiple-choice testing?

2. INTRODUCTION.

Brown (1970) has defined a true score as a score that would be obtained if a test were perfect - that is if it could measure a given characteristic without error. But no test yet devised can do this, so an error exists in any test and this error plus the true score equals the obtained score.

i.e. $Xt = XT + Xe$

where Xt = test score

XT = true score

and Xe = error

The reliability of a test would be determined by the ratio of the variance of the true test scores to the variance of the total or obtained test scores. But the true test score and hence the error of a test can never be determined and the examiner has to make use of statistical tests devised to estimate either the error or true test score and arrive at an estimate of reliability. Since this estimation is usually arrived at by a series of formulae the fact that the result is an estimate is often lost sight of due to the hypnotic effect of the mathematical formula which tend to suggest that an absolute value has been arrived at.

There are three basic ways of estimating test reliability (Brown - 1966, Ebel - 1972).

1. The first is the test-retest method, either by giving the same test on two occasions or having a parallel or equivalent test available. The latter variation requires two forms of the same test which have been shown to be equivalent. One is given first and the other given later at the same or at a later time. The correlation co-efficient between these two sets of scores for any group of students is the measure of the reliability of either test. There is i) a difficulty and ii) a fallacy in this method.

   i) The difficulty is to devise two forms of any test that are equivalent. While the two tests may have been

equivalent/.....

equivalent in former years, they may not be equivalent for the student body that is being tested at any given time because of different emphasis in teaching normally present from year to year and the difference between the student bodies that occurs from year to year.

ii) The fallacy is that the student body doing the second 'retest' has the same knowledge as on the first test occasion. This concept entirely ignores the well known fact, referred to above in Chapter 1, that any examination is an important learning experience for the student, and the student will therefore be possessed of more knowledge in the "retest" situation.

In any event, this method of establishing reliability would prove to be far too tedious for routine use for any but a worker in a particular research project.

2. The second method is the Spearman-Brown "split-half" method which can be used for a single test. This relies upon the separation of the test into two halves, which can be arrived at in a number of ways, e.g. either sequential, or taking odd questions for one half and even questions for the other half, etc. The Spearman-Brown reliability is given by the formula Reliability $= \frac{2r}{1+r}$ where r = the original correlation between the two halves.

The estimate of test reliability will be dependant on which method of obtaining the two halves of the test has been chosen, and assumes that the variability of the two halves is equal. Lord (1956) has shown how variations of test reliability may occur when different methods of selecting the split halves of the test are chosen and the Spearman-Brown method has been supplanted by the Kuder-Richardson methods.

3. The most widely used estimates of reliability are the tests devised by Kuder and Richardson as a measure of the internal consistency of the test. Kuder and Richardson worked independently along

the/.....

the same lines and reached similar conclusions. They published jointly a series of formula, arrived at by slightly different methods and with slightly different characteristics (Kuder and Richardson - 1937).

It is not proposed to examine the derivation and characteristics of these formulae in detail. Examination of the performance of three of them however is of importance when consideration is given to which of them are suitable for general use.

i) The theoretically most exact formula published by Kuder and Richardson was the Kuder Richardson formula No. 8 (KR 8). This formula requires the computation of (a) mean score for the test, (b) standard deviation of scores, (c) correlation of coefficient between each item and the total test, (d) proportion of correct answers for each item of the test, (e) the product of each item-test coefficient and each corresponding item variance and the summation of them for all items, (f) the proportion of those getting the right answer and the product of this proportion and (g) the proportion not getting the item correct for each item, and the sum of this product for all test items. In tests conducted by the authors the KR 8 formula gave estimates slightly lower than estimates of reliability arrived at by the Spearman-Brown formula.

ii) Requiring far less computation and giving substantially the same estimate of reliability as the KR 8 formula is the Kuder Richardson formula No. 20 (KR 20) which was favoured by the authors as being suitable for most practical situations. The KR 20 formula has become the most widely used estimate of reliability in the reports from the journals of educational technology.

The/.....

The formula reads:

$$R_{tt} = \frac{n}{n-1} \times \frac{\sigma t^2 - \Sigma pq}{\sigma t^2}$$

where R_{tt} = Estimate of test reliability

n = number of items in the test

σt^2 = square of standard deviation of test scores obtained (variance of test scores)

p = proportion of correct answers to an item

q = 1 - p (i.e. proportion who got it wrong or made no response)

Σpq = sum of the products of p and q for all test items.

The KR 20 formula tends to give a low estimate of test reliability as compared to the Spearman-Brown estimate and other formulae of Kuder and Richardson (Richardson and Kuder-1939).

iii) Another popular formula is the Kuder Richardson formula No. 21 (KR 21) published in the same article which is based on a rigid assumption that all items have the same item difficulty. The KR 21 formula gives a much lower estimate of reliability and in any event since it is based on the rigid assumption of equal test difficulty of all items its use is precluded if test difficulty of the items varies greatly in any examination. Since the item difficulty in our tests varied considerably (See Chapter V) the KR 21 formula would not have been suitable.

In view of the small difference between the results using the KR 8 formula and the KR 20 formula it was considered that the latter, which requires less

computation/.....

computation, would be adequate for our needs. The disadvantages attributed to the KR 20 formula did not appear to be operative in the tests that had been given in our department.

Brown (1970) has pointed out that the KR 20 is applicable only to power tests and not to speed tests, as the values for p and q are to some extent reliable only is so far as each student has attempted the item in question. The tests that we had been giving were all designed as power tests - but it is perhaps true that to some of the weaker students, since they were followed by essay questions in class tests, there may have been some elements of a speed test in them. This might well be true of all power tests with a time limit.

Another objection to the KR 20 formula is that it applies only to tests scored by one point for getting the item correct (Ebel 1972), and not to tests where the questions are weighted, or where tests are corrected for guessing. Until such time as the trial of confidence testing the items in the tests in our department were unitary in value, and as in our experience the effect of correcting marks for wrong guessing (as discussed in Chapter 11) was to increase the reliability coefficient, this did not appear to contraindicate its adoption.

The major objection to using Kuder Richardson formulae has been that the only true measure of homogeneity would be by factor analysis in that if one factor alone was sufficient to account for the variation in performance on all items, the test would be considered homogenous in construction, but if more than one factor was required to account for this variation, the test would have to be considered non-homogenous. Measures of homogeneity have been devised (Lumsden-1961, Horst-1966, and Magnussen-1966), but none of these nor factorial analysis has gained general approval.

A further criticism of the coefficient of reliability has been that variability may arise not only from the quality of the test but from the variability of the group being tested, and for this reason it has been proposed that the standard error of the test, which is not dependant on the variability of the group being tested, be used as a measure of test reliability rather than the reliability coefficient. Lord (1957 and 1959) and Swineford (1959) have however shown that in a test using one type of item the standard error of measurement is almost entirely

dependent/.....

dependant upon the number of items in the test.

All reliability estimates have their drawbacks but since the KR 20 formula appeared to be the most widely used measure of the estimate of the reliability of a test it was decided to standardize our tests in relation to this formula and to use it as a comparative measure in assessing our tests. Within the limits that, at best, the coefficient of reliability as expressed by the KR 20 formula would:

i) give us a theoretical estimate only of test reliability

ii) would tend to read lower than estimates of reliability by other means

iii) might not be entirely accurate in tests weighted by items or corrected by guessing.

It was nevertheless felt by using the KR 20 routinely the error would be standardized and in the absence of any better alternative would at least serve as a means of comparing one test to others, not only within the department, but to other results using the KR 20 formula, which tend to form the bulk of published results.

One of the problems that we faced was that many of our questions were known to the students and in any examination it was necessary to introduce new items. Since we had no performance data available for these items we faced two possibilities.

Firstly that the items would not discriminate between good and poor students - and secondly that the effect of this would be to reduce the estimate of test reliability for the test in question by reducing the variance of test scores. (See formula) It was therefore important to note the effect on test reliability of removing items that did not discriminate effectively between students (as were it not for practical considerations these questions would not have been submitted in the first place). To this end (as described under Methods - this chapter) I examined the effect of the removal of these questions with poor discrimination on the KR 20 in the tests we had given.

3. METHODS AND MATERIALS.

A) In examining the factors that influenced reliability the following data was recorded from each multiple-choice test that had been given in the department and for which records were available.

i) the number of questions in the test
ii) the number of students writing the test
iii) test difficulty
iv) standard deviation of test difficulty
v) corrected mean score
vi) standard deviation of mean score
vii) Kuder Richardson 20 reliability coefficient

These variables were plotted against each other and cross-correlation coefficients were calculated for each of these variables. To obviate the possibility that student responses in self-evaluation tests might differ from their responses in certifying examinations this data was examined as well for class tests and final examinations only.

B) To investigate the factors that would increase reliability the effect of the removal of questions with a poor discriminatory power was examined. A program was written and incorporated in the Marking Program which successively eradicated questions below a required discrimination index, called the cycling program.

The cycling program marks and analyses the test and then, using a phi value chosen by the examiner, remarks and re-analyses the test ignoring those questions with a phi coefficient equal to or less than the value chosen. The Cycling Program will continue to do this, raising the operative value of the phi coefficient in steps, by a value which can again be chosen by the examiner or, in default of this choice, by a phi value of 0,02, which we found to be a convenient value for this incremental step. For this analysis the discriminatory index that was chosen for reasons discussed in Chapter 11 was the phi-coefficient. The program will continue to reiterate marking and analyzing

results/.....

results and excluding those questions whose phi coefficients are below the phi value given by the successive incremental steps of 0,02 (or the value chosen by the examiner) until no further rise in the Reliability Coefficient can be obtained.

The results of using this cycling program was examined for all tests from 1972 for which data was still available.

4. RESULTS.

A) The date and type of examination, the number of questions in the test, the number of students writing the test, the raw mean per cent of correct answers, the standard deviation of raw score, the corrected mean of students scores, the standard deviation of corrected scores, and the Kuder Richardson 20 reliability coefficient for all examinations are set out in the table (Table R1 (Book 11, page 25). The means and standard deviations of these variables for all tests are set out in Table R2 (Book 11, page 26). If the self-evaluation tests are excluded (denoted by SE in Table R1) the means and standard deviations obtained are set out in Table R3 (Book 11, page 27).

From these tables it can be seen that the lowest KR 20 value in all tests was ,62 for self evaluation tests in March 1971 and March 1972, while the lowest KR 20 value for class and final tests was ,72. The highest KR 20 obtained for SE tests was ,74 in April 1974, as compared to ,99 in the final tests that of 1974. A comparison of the KR 20 means in these two categories (Tables R2 and R3) reflects the difference, being higher for class and final tests than for all tests.

This difference of means for the KR 20 reliability coefficient between all tests and class tests only gave a t value of 2,47 which is significant at the probability value of $p = <,01$.

The scattergrams resulting from plotting these variables on the X axis against the KR 20 reliability coefficient on the Y axis are shown in Figures R 1-6, (Book 11, pages 28-33) for the data from all tests, and Figures R 7-12, (Book 11, pages 34-39) show the relationship of this data for class tests and final certifying examinations only.

The correlation coefficients between these variables are set out in

Table/.....

Table R4 for all tests and Table R5 for class tests and certifying tests only. (Book 11, pages 40-41). From these tables it can be seen that significant correlation exists between the KR20 reliability coefficient and (i) the number of questions, and (ii) the raw score or test difficulty for both all tests and class tests only. As can be expected significant correlation exists between the raw scores and corrected scores and the raw standard deviation and corrected standard deviation. Apart from these the variables examined are not correlated significantly with the KR 20 coefficient.

B) The changes that occur when the cycling program, referred to above, which eliminates items of poor discriminatory in successive steps was used is illustrated in Tables R6 - R15, (Book 11, pages 42 - 51). These tables set out the number of the cycle, the phi discrimination value for a question to be retained, the KR 20 value, the percentage of questions that (i) remained in the test and (ii) were deleted, the test difficulty, the corrected mean score and standard deviation for each cycle. These changes in each are shown graphically in figures R13 to R17, (Book 11, pages 52-56) for the following

i) KR 20 coefficient
ii) percentage of questions remaining in the test
iii) percentage of questions deleted
iv) the corrected mean score and
v) the corrected standard deviation.

From these results it can be observed that the KR 20 coefficient increased or remained fairly static in each cycle, Fig. R13 (Book 11, page 52) despite the fact that the number of questions in the test declined rapidly - Fig. R14 (Book 11, page 53).

Of significance is the fact that the corrected standard deviation increased with each successive cycle (Fig. R17 , Book 11, page 56) while no significant changes were seen in the other variables.

The/.....

The mean values, standard deviation and range of all variables available for study in the 56 cycles that were examined in the nine tests are set out in Table R16 (Book 11, page 57).

To examine the relationship between these changes the variables in 56 cycles for the 9 tests were correlated with each other and the results are set out in Table R17, (Book 11, page 58). From this table it can be seen that significant correlations exist between the KR 20 reliability coefficient changes and

i) the number of students
ii) the cycle
iii) the percentage of questions remaining in the test
iv) the number and
v) percentage of questions deleted and
vi) the corrected mean standard deviation

No correlation exists between the KR 20 coefficient changes and the number of questions, the test difficulty, the corrected mean score and the Chi-squared figure for the distribution of results.

5. DISCUSSION.

A) Conventional Program: From the KR 20 formula, namely

$$R_{tt} = \frac{n}{n-1} \times \frac{\sigma t^2 - \Sigma pq}{\sigma t^2}$$

it can be theoretically deduced that as the number of questions (n) increases the value of the reliability coefficient can be expected to fall if all else remains constant. As discussed by Ebel (1972) the converse in fact obtains and he has constructed a table, using a Spearman-Brown formula for the theoretical relation between test reliability and test length, illustrating the effect of doubling the number of 5 choice items in a test on the test reliability. Reference to the Table R1, (Book 11, page 25) and Figure R1 (Book 11, page 28) and Tables R4 and R5 (Book 11, pages 40-41) demonstrate this relation in the tests given in the department and our results are in conformance with the generally observed phenomenon that test reliability increases with the length of the test. The correlation coefficient between the number of questions in the test and the test reliability (KR 20) being ,824 for all tests submitted and ,691 for the class tests submitted in the department. Both these coefficients are highly significant.

From the KR 20 formula it can also be theoretically anticipated that for tests with an equal number of questions those with a greater variability of student scores (i.e. higher σt^2) will show greater reliability than those with a narrower spread. From Tables R1 and R3 (Book 11, pages 25 and 27) it can be seen that the standard deviation of the corrected scores was higher than the standard deviation of the raw scores for both all tests and class tests only, and it would be anticipated that the correlation between the reliability coefficient and the corrected standard deviations would be greater than between the reliability coefficient and the raw standard deviation. Table R4 and R5 (Book 11, pages 40/41) reflect this and show a higher figure of correlation between the standard deviation

of/.....

of corrected scores than standard deviation of raw scores. These differences are, however, not statistically significant. (Guildford & Fruchter - 1973). Ebel's remarks (1972) regarding the desirability of scores being arrived at not by weighting or correcting for guesses, as mentioned in the introduction to this chapter, are borne out by the fact that from Tables R4 and R5 (Book 11, pages 40/41), it can be seen that there is a higher correlation between the reliability coefficient and the raw score than between the reliability coefficient and the corrected score.

These differences are significant and tend to lead the examiner into a dilemma of whether to correct scores and thereby lessen the reliability of the test or to aim at high test reliability and allow guesses to go uncorrected. As discussed in Chapter 11 the prime function of the examination is surely to get as close to a true result for each student as possible and, as it has been shown that students can increase their scores considerably by guessing one ponders if the results might have been different had this been encouraged in our students.

(A feasible study for the future might be to investigate the effect of reliability of two situations, viz. (i) encouraging guessing, and (ii) penalizing guessing).

While test difficulty (as measured by raw mean score) correlates highly with the reliability coefficient (Tables R4 & R5, Book 11, pages 40/41) further investigations into our results (as discussed in Chapter V) has cast some doubt on this relationship. Test difficulty ranged in our tests from ,503 to ,738 and it may well be that this observation occurs only within this narrow range of test difficulty.

In any event the examiner may be constrained by certification requirements in choosing the test difficulty value that is desired, and it may not be possible to alter this variable in attempting to increase test reliability. This was in fact our experience in the general anatomy multiple choice paper of 1971 when the majority of items set were in the

middle/.....

middle difficulty range and a low coefficient or reliability was obtained. This theoretical concept is in contradiction to Ebel's statement that test reliability can be increased by the use of items of middle difficulty range. The percentage of items with an item difficulty of between 40 and 60 was ascertained for the tests of 30.10.69 to 05.11.74 and the results of these are set out in Table R18, (Book 11, page 59) and it can be seen that there is no difference between the reliability coefficient of these tests with a high number of these items (more than 30%) and the reliability coefficient of those tests with a large spread of item difficulty. (Table R19, Book 11, page 60).

B) Cycling:

The effect of variation of scores on reliability is most important when one considers the effects of cycling. Initially cycling was designed to allow of tests being given that included new items that had not been validated for their discriminatory ability in previous tests. If the discrimination ability of such an item was low it would not have been chosen for a test and accordingly the cycling program after initially marking the test would then remark the test excluding those items with poor phi coefficient of discrimination.

As the determination of the phi coefficient depends on the better students getting a question right and the poorer students getting it wrong the removal of such a question would adjust student scores by subtracting the mark of a correct response from the poorer students' score and adding back the penalty for getting a question wrong (usually ,25 of a mark) to the good students' scores. This would have the effect of broadening the distribution of scores - and while possibly decreasing the raw and corrected scores it would certainly increase the standard deviation of the scores. The extent to which this occurs can be seen from the Tables R6 to R15 (Book 11, pages 42-51) where the increase in the standard deviation of the corrected score can be noted in each successive cycle for all the tests examined.

From/.....

From the KR 20 formula this would increase the figure for σt^2 and so the right hand multiplicant. (See formula page 63). However, the number of items in the test would fall and so decrease the value of the left hand multiplicant. That the drop in this value is more than offset by the rise in σt^2 is well shown by the rise of the reliability coefficient that occurs with successive cycles (Tables R6 - R15, Book 11, pages 42-51).

Since items are deleted on their phi coefficient alone without regard for their item difficulty it is unlikely that the sum of pq's per item is being sufficiently changed. This aspect however has not been examined as from a practical aspect cycling was doing precisely what it had been designed for, i.e. crystallising a test in terms of the discriminatory ability of its items.

Since the KR 20 formula incorporates n (- the number of items) and σt^2 (- the variance) it is not surprising that high and significant correlations between these values and the values for reliability coefficient are obtained from cycled tests (Table R17, book 11, page 58). In those tests marked by cycling correlation is seen between the KR 20 coefficient and

i) the number of students
ii) the number of the cycle
iii) the number and percentage of questions deleted
iv) the percentage of questions remaining and
v) the mean standard deviation (Table R17, book 11, page 58)

as opposed to correlation between KR 20 coefficient and

a) the number of items in the test and
b) the raw score (test difficulty)

in conventional programs (Tables R4 & 5, Book 11, pages 41/42).

Since test difficulty may be conditioned by the requirements of

certification the examiner has in practice only two variables at his disposal for increasing test reliability, namely

1) the number of items in a test and
2) the discriminatory ability of any item - as revealed inferentially by previous testing or in reality by cycling the results of a test.

The mean test difficulty (mean raw score) or corrected mean score do not appear to relate to test reliability and the manipulation of these two factors would not of necessity result in an increase in test reliability. (Table R17, Book 11, page 58) Of interest is that the test variance appears to play a larger role in the establishing a value for test reliability than the number of questions, as shown by the fact that the percentage rise in standard deviation for each cycle is small when compared to the percentage drop in the number of questions - but is followed by a rise of reliability.

6. CONCLUSIONS.

1. The results of our observations have confirmed that the reliability of a test increases in relation to the increase in the number of items in that test.

2. A more powerful factor (as shown by the results of the cycling programs) in increasing the reliability of a test is by increasing the variability of student scores.

3. Raw scores correlate better with test reliability than do corrected scores in conventional tests, which may lead to a dilemma in the examiners mind regarding the desirability or not of penalising guessing.

4. In our hands individual item difficulty has not affected test reliability.

5. Test difficulty appears to relate to test reliability - but this finding may not assist an examiner to increase test reliability if faced with the constraints of marking for certification purposes.

6. For practical purposes the examiner has only two variables available to him to increase test reliability

   i) increasing the number of questions in the test - which may be limited by the time available for testing, and

   ii) increasing the variability of student scores.

   This latter can be achieved

   a) inferentially by using only items with high discriminatory ability as disclosed in previous examinations, or

   b) in reality on the performance of the item in the test in question by the use of a cycling program to eliminate those questions with poor discriminative ability.

CHAPTER V.

QUESTION FORMATS, ITEM DIFFICULTY AND DISCRIMINATIVE ABILITY.

1. QUESTIONS TO BE ANSWERED

The questions that were investigated in this section of the enquiry were:-

1. Is there any relation between the format of the question and its difficulty? In other words, do the students tend to cope more easily with one type of question than another?

2. Is there any relation between the format of the question set and its ability to discriminate between students?

3. Is there any relation between the difficulty of a question and its ability to discriminate between good and bad students?

2. INTRODUCTION.

1. Item Difficulty.

In any examination there will be a certain number of students who score the correct answer to any particular item. Some students will get it wrong and some will not attempt an answer. The easier the question the higher will be the number of correct answers, and conversely the more difficult question will present a greater number of students responding incorrectly or making no attempt.

The difficulty of a question can be measured by the ratio of correct answers to the total number of responses to that item (Hubbard & Clemans - 1961). I have related the number of correct responses not to the total number of responses actually made but to the total number of responses that could have been made if every student had attempted the item. This ratio, conventionally referred to as the Item Difficulty, is in fact inversely related to the difficulty of the question and the term Easiness Index (Lennox - 1974) is more apt, but in order not to create confusion the conventional term Item Difficulty has been retained despite the anomaly of this term.

2. Discrimination Index.

The measure of the ability of an item to differentiate between good and bad students has been discussed in Chapter 11 (Page 30). For the purpose of the investigation into the role of the discriminative ability of an item the phi index (denoted as ϕ in the tables and graphs) was used.

The determination of the phi index has been discussed in Chapter 11.

3. Question formats: (See Appendix A).

(i) Originally the type of questions set in the department were all in which the student had the task of choosing one correct answer from five possible alternatives. In its simplest form this is known as the one-from-five format in which one alternative is correct and the other four alternatives are incorrect.

This/.....

This represents the basic type of multiple-choice question and is used extensively throughout the world. In my analysis this type of question is referred to as the one-from-five format. (or 1/5 in graphs and tables)

However, this type of question has a disadvantage in that it is necessary to present the student with 4 incorrect distractors and the possibility exists that one of these distractors will become fixed in his mind, especially with questions in self-evaluation tests set during the year. If this were so it would be disadvantageous to the student.

(ii) Accordingly we have adopted an alternative method of posing the one-from-five choice to the student in presenting him with four statements, any one of which may be incorrect, but of which all four may be correct. The student is required to identify the incorrect statement or if he considers all four statements are correct he indicates this by choosing for his answer the fifth statement which in our examination is expressed as "5 - all four statements ARE correct".

It is considered by the staff that this is a more desirous type of question than the one-from-five format in that the student is presented with only, at most, one incorrect statement to identify and that a positive reinforcement of learning occurs when students encounter these questions in self-evaluation circumstances, since they see at least three correct answers. This format is referred to as the wrong-from-five format (W/5 in tables and graphs).

In addition to these two formats, we have also used to some extent some of the other formats presented by Hubbard and Cleman (1961) - namely:-

(iii) Correct-from-four (C/4 in tables and graphs) in which the student is presented with four statements, any number of which may be correct or incorrect. With the aid of an appropriate key the student is again presented with the task of choosing one from five

alternatives/.....

alternatives. (See Appendix A)

(iv) Causal.

The student is presented with two statements which may be independently correct or incorrect and which are linked with the word because.

Again a five choice opportunity is created for the student by the use of a key. (See Appendix A)

(v) Related-exclusion.

This format of question consists of two columns each with 5 alternatives. In the first column the student is required to identify one structure which is not of the same group as the other four and then identify this group from the five set out in the second column.

(vi) True-False.

In an attempt to extend the departmental bank of M.C.Q. questions we turned to the technique of a stem followed by a string of multiple-choice items, and while these are not one-from-five questions they were available for examination in this analysis.

3. METHODS AND MATERIALS.

The three questions posed were investigated using the same data. As a by-product of the computer marking program there was available for each question the phi-coefficient and the item difficulty.

On the assumption that the staff of the department were improving their ability to develop and select items, it was decided to sample questions retrospectively from the examination of November, 1974. Six questions types were selected for examination. These types of question format were:-

1. The correct answer out of 5 alternatives. (In the charts and figures that are presented this format type is denoted as 1/5.)

2. The incorrect alternative out of 4 or 5 if all alternatives were correct. (Denoted as W/5.)

3. The number which would relate to which combination of 4 given alternatives was correct. (Denoted as C/4.)

4. The correct answer to the question where two statements were linked with a Because. (Denoted as Causal).

5. The type of question referred to as Related Exclusion where the candidate has to select one of 5 alternatives which is not related to the others and denote the terms of this exclusion. (Denoted as RELATED.)

6. The true or false type of question where the candidate has to signify whether he considers a statement true or false. (Denoted as T/F)

The questions for each format were taken without selection from the most recent examinations set, starting with the examination of November, 1974 and working backwards in time until a sufficient number of questions for statistical purposes of each type of question had been sampled. For the related exclusion type of question it was not possible to find as many questions as might be considered statistically significant but sufficient numbers for the other types were available. The number of questions of each type was:-

1/5	=	130
W/5	=	107
C/4	=	110
Causal	=	82
Related	=	24
T/F	=	124
Total	=	577

The phi-coefficient and item difficulty was recorded for each question in the series and analysed as follows:-

1. The phi-coefficient and the item difficulty were plotted against each other for each type of question and for the series as a whole.

2. Following preliminary inspection of the results obtained from the above, and to test whether questions in the middle range of item difficulty were better discriminators than those at the extremes of the scale a computer program was written which calculated for each question the deviation of the item difficulty

   (a) from the mean item difficulty for that type of question
   and
   (b) from an arbitrary figure of item difficulty which for the purpose of this investigation was set at 50%.

   For each of the six blocks of question types that we were investigating, together with a seventh block which was made up of all the questions, data on the relationship between the phi-coefficient and the deviation of the item difficulty was obtained by means of this program:-

   (a) from the mean item difficulty for the block of questions of that particular question type, (or the mean item difficulty of all the questions in the case of the seventh block of all questions)
   and
   (b) from an arbitrary figure of item difficulty of 50%.

The/.....

The deviations of item difficulty so obtained were plotted against the phi-coefficient for each type of question and for the questions as a whole, and values were determined for the correlation between these figures again for each question type and for all questions.

3. The estimation of the analysis of variance was obtained for the means of the phi-coefficient and also for the means of the item difficulties for each type of question.

4. The mean phi was determined for each type of question and the means thus obtained tested against the means of other question types for significance of the variation obtained.

5. Similarly the mean item difficulty for each type of question was determined and the means thus obtained tested against the means of item difficulties obtained for the other types of questions.

4. RESULTS.

1. The Relation between the Type of Question format and the Difficulty of the Question.

The mean item difficulty of the question and standard deviation was determined for the questions as a whole and for each particular type of question format. These results are set out in Table QT 1 (Book 11, page 61). When arranged in descending order of item difficulty from the easiest to the most difficult types of questions these formats present themselves as set out in Table QT 2 (Book 11, page 62) with the Related exclusion type of question emerging as the type found most simple by the students and the causal that they found most difficult.

When the data was submitted to the Analysis of Variance test the F value obtained was 11,728 which indicates a significant difference between these means.

These means were then tested for the significance of the difference between each of the different formats. The results of this analysis are set out in Table QT 3 (Book 11, Page 63) from which it can be seen that the Related exclusion format was significantly easier than all other types of format. There was no significant difference between the 1/5 and T/F formats but these were significantly harder than the Related format and significantly easier than the other formats.

The last three types, C/4, W/5 and Causal are grouped together in not being significantly different from each other in difficulty but are significantly more difficult than the other types of format.

2. The Relation between Question Type and Discriminative Ability.

The mean discrimination index for each question type and the series as a whole is set out in Table QT 4 (Book 11, Page 64). For all 577 questions in the series the mean phi-coefficient was, ,1896 with a standard deviation of $\pm$,1112.

For individual question types the mean coefficient ranged from ,1325 to ,2588 as set out in the Table. The value of the phi-coefficient to

give/.....

give 99,9% and 90% reliability for 180 students is set out in Table QT 5 (Book 11, Page 65).

When arrayed in descending order of phi-coefficient or discriminative ability as set out in Table QT 6 (Book 11, page 66) it can be seen that the Related exclusion type discriminated best among students and the True/False type was the least efficient discriminator. It can be seen from Table QT 5 (Book 11, page 65) that all questions taken as a whole in the series were discriminating between good and bad students at better than the 98% level of confidence. The Related exclusion type discriminated to the level of 99,9% confidence, 1/5 and C/4 to the 99% level and W/5 and Causal to the 98% level - but the True/False format did not reach the 95% confidence level. (Table QT 6 Book 11, page 66).

The differences of means of phi-coefficients of the various question types was subjected to the Analysis of Variance - and as the F value of the test was 12,631 these differences are significant.

These means were then subjected to analysis as to the significance of the differences of means between each of the different formats. The results of the analysis are set out in Table QT 7 (Book 11, page 67) where the T values and degrees of freedom, in parentheses, are shown in the upper right half of the matrix with the statistical significance of this T value for the corresponding degrees of freedom shown in the lower left half of the matrix.

There were significant differences in discriminatory ability between the question formats, as measured by the comparisons of means, in that (expressed in rank order):

(i) The Related-Exclusion format did not differ significantly from the one-from-five but was significantly better a discriminator than all other formats

(ii) The one-from-five format did not differ significantly from the correct-from-four format but was significantly better than the

other/.....

other three formats.

(iii) The correct-from-four, wrong-from-five and causal did not differ among themselves but were significantly better than the True/False format.

(iv) The True/False format was significantly a poorer discriminator than all other question types.

3. The Relation between Item Difficulty and Discrimination Index.

Table QT 8 (Book 11, page 68) shows the number of questions, the mean item difficulty and standard deviation for the series as a whole and for each of the question types considered. For the series as a whole the mean item difficulty was 64,61% with a standard deviation of $\pm$ 20,17%, the mean item difficulty for question types varying from 56,59% to 76,08%, as set out in the Table.

The relation between the item difficulty of a question and its discriminative ability is shown for each of the 6 types of question investigated and the series as a whole in the graphs QT G-1-7 (Book 11, page 69-75) where item difficulty (plotted on the x axis) and discrimination ability (on the y axis) are plotted against each other. Graph QT G1 (Book 11, page 69) shows the results for all 577 questions examined and Graphs QT 2-7 (Book 11, pages 70-75) show the results for the specific type of format under review. Except in the case of the Related Exclusion type of question QT G-6 (Book 11, page 74) where there appears to be some degree of negative correlation between item difficulty and discriminative ability none of the graphs for the other question formats studied suggest any relation between these two factors.

Co-efficients of correlation were determined between item difficulty and the phi-coefficient and these results are set out in Tables QT 9 (Book 11, page 76) for the series as a whole and for each of the question formats individually.

The figures for correlation are not significant for the series as a whole nor for any of the question types examined except for the Related Exclusion type which is significant at the 98% level of conficence.

If/.....

If, however, the graphs plotted for the series are examined more closely it would appear, for some at any rate, that a relation exists which might be expressed in terms of a non-linear regression line.

This can be seen by reference to Graph QT 8 (Book 11, page 77) which represents the plot of all questions in this series on which curved lines have been drawn to represent this relation which can be expressed by a polynomial regression formula.

While a polynomial regression might fit the distribution it would not be easy to work with, and if, as an alternative two parallel lines are drawn instead of curved lines, as shown in Graph QT 9 (Book 11, page 78), along the upper and lower borders of the distribution towards the 50% level of item difficulty a suspicion emerges that a linear relation does exist between item difficulty and phi-coefficient, positive for those items with low item difficulty values and negative for those items with high item difficulty values.

This possible relation suggested further investigation and the item difficulty was then considered in two ways :-

i) as a deviation from the mean item difficulty obtained for that type of question, and for the series as a whole, and

ii) as a deviation from an arbitrary value of 50% for item difficulty.

The relation between the phi-coefficient and the item difficulty considered in these two ways was then examined.

The results of plotting these values for the deviation of item difficulty against the phi-coefficients are shown in Graphs QT 10-11 (Book 11, pages 79-80) for the series as a whole and in Graphs QT 12-23 (Book 11, pages 81-92) for each of the question formats under review. When these graphs are examined substantially the same picture emerges from both the plots for the deviation of item difficulty from the mean and from the arbitrary value of 50% Negative correlation can be seen for the series a

as/.....

as a whole (Graphs QT 10 and QT 11) and for Question types:-

W/5	:	Mean Deviation (Graph QT14) &50% Dev. (Graph QT15)
C/4	:	Mean Deviation (Graph QT16) &50% Dev. (Graph QT17)
Causal	:	Mean Deviation (Graph QT18) & 50%Dev. (Graph QT19)
Related	:	&50% Dev. (Graph QT21)
T/F	:	Mean Deviation (Graph QT22) &50% Dev. (Graph QT23)

No correlation is evident for question type one-from-five, either for mean deviation or 50% deviation (Graphs QT12 and QT13 (Book 11, pages 81-82) and for the related-exclusion question type for deviation from the mean (Graph QT20 (Book 11, page 89). This relationship is confirmed by the correlation co-efficients obtained between the phi-coefficients and the item difficulties expressed as deviations from the mean or from 50% as set out in Table QT10 (Book 11, page 93).

From this table it can be seen that with only three exceptions a significant negative correlation exists between the discriminative ability of a question and the extent by which it deviates from either the mean item difficulty or the arbitrary 50% value, thereby indicating that the discriminatory ability of an item falls off as it deviates from either of these mean values. The exceptions to this observation are the one-from-five and the related exclusion formats. The effect of ranking the question formats in order of their difficulty is shown in Table QT11 (Book 11, page 86).

(As discussed above the Item Difficulty is in fact a misnomer and really represents an Easiness Index, thus the more difficult questions will have low item difficulty values and the easier questions high item difficulty values.)

5. DISCUSSION.

1. The Relation Between the Format of a Question and its Difficulty.

From the results obtained (Tables QT 1-3 Book 11, pages 61-63) it is clear that there are significant differences in the degree of difficulty that these six question formats presented to the students, and they can be grouped into 3 categories which are significantly different from each other:-

i) Over 75% Mean Item Difficulty - related-exclusion significantly easier than all other types,

ii) 65 - 75% Mean Item Difficulty - one-from-five and true/false which did not differ from each other but were significantly less difficult than,

iii) those under 65% Mean Item Difficulty - correct-from-four, wrong-from-five and causal.

It is not surprising to find that the Causal format caused so much difficulty as this type of question was not all that enthusiastically viewed by the examining panel. It was, however, surprising to note that the wrong-out-of-five format appeared so difficult to the students since, as discussed above, this appeared to the panel to be a straightforward and praiseworthy type of question in that a minimum of wrong information was presented to the students. Nevertheless the students found this far more difficult than the traditional one-out-of-five format with its greater quota of false alternatives. This phenomenon might well spring from the inherent mistrust by students of their examiners and the misbelief that the examiners are on occasions presenting them with all four alternatives correct in an item. This aspect has not been examined in this study and might well justify a separate investigation.

That the related-exclusion type of question should prove the easiest type of question to answer is an unexpected finding and is perhaps due to the fact that it looks so frightening. Lennox (1974) who refers to it as the excluded-term type has castigated it unequivocally saying, "This is a terrible thing with which to face an unsuspecting candidate in the middle of a stressful

examination/.....

examination. This is likewise great fun to compose but the temptation should be resisted". In defence I may point out that we have invariably used this type of question only at the end of the multiple-choice paper, and it may be that the standard of knowledge required by the examiners has been less exacting in view of the apparent universal but undeserved ill repute that this format enjoys. This leniency may not, however, have been appreciated by those candidates with little knowledge, (to whom every question probably appears frightening), and they may well have left this type of question severely alone.

That the causal, wrong-from-five and correct-from-four formats of question were significantly found to be more difficult than the straightforward one-from-five or true/false format may be due to the fact that these former types of question require additional thought and concentration in achieving a correct response than do the latter two (1/5 and T/F) and it may well be that the less knowledgeable students in anatomy are not capable of either making the effort or making the effort correctly, or less inclined to make the attempt. These three formats therefore present themselves as admirable instruments for the purpose of grading a class as finely as possible - but might not be indicated if the examiner has other objectives such as the elicitation of basic knowledge in mind.

The one-from-five format which is the most straightforward of all formats occasion the student no problems in its presentation and is found to be significantly easier than all other formats. It is surpassed only by the related exclusion format and the reasons for this have been discussed above.

2. <u>The Relation Between Question Format and Discrimination of a Question.</u>

The results obtained in the analysis of our data indicate clearly that the type of question format chosen affects the ability of the question to distinguish between good and bad students. (Table QT 4-7 (Book 11, pages 64-67)).

From these figures it can be seen that the true or false format question fared as badly in our hands as suggested by other observers and discussed in Chapter 11. Were it not for the fact that the true or false format had

been/.....

been introduced for the purpose of allowing us to extend our bank of questions rapidly and allow our students to retain their question papers (which they appeared to be doing illegally anyway), and for the fact that we were using true/false formats only with confidence marking so as to offset the "noise" from guessing there would be little reason to continue with this format when judged by its discriminative ability, which was below the 95% confidence level.

To some extent the same fate appears to have overtaken the causal format and the wrong-out-of-five question format. That this lack of discriminatory ability should have appeared with the Causal format has not occasioned much surprise. There has been doubt in some cases among the examining panel as to when a causal relationship exists and in any event the candidates decision about a causal relationship is only called into question when both statements and reasons are true.

What is disconcerting, however, is the poor discriminatory ability of the wrong-answer-out-of-five format. As discussed earlier this format endeared itself to the panel in that there were the minimal number of false alternatives (and in some questions none at all) which appeared desirable in our self-evaluation testing (and student learning) opportunities. (Chapter 11). This effect cannot be explained in terms of our findings in the previous part of this chapter in that discriminative ability fell off as item difficulty deviated from a value of approximately 50%. The mean value of the items in this format of 57,54% with a standard deviation of 20,81% are all within normal limits (Table QT 1, page 61). It can also be recalled that this type of format showed a good correlation between phi-coefficient and deviation from 50%. (Table QT 10 page 93). It would appear nevertheless that in our hands so far this format of question has not fulfilled its promise and unless the examining panel can improve on its performance to date, further consideration will have to be given as to whether it is entitled to remain in our repertoire.

From the investigation it would appear that the related exclusion, the one-from-five and the correct-from-four are linked together as those types of

questions/.....

questions which are excellent discriminators of students, at above the 99% level of confidence. The W/5 and Causal formats are intermediate but still perform reasonably at a 98% probability level. The T/F format discriminates in the mean at a level of below 95% and is significantly different in its performance from the other formats. Unless justified by other reasons there appears to be no reason for retaining the T/F format if discriminatory ability is the sole criterion of questions to be used.

3. The Relation between Item Difficulty and Discriminative Ability of a Question.

When the mean item difficulties of the question formats under review are considered in relation to discriminatory ability certain salient features emerge.

Most question types, excepting 1/5 and Related, show that a question of middle range difficulty tends to discriminate more effectively among students than those questions which are too easy or too difficult. When the item difficulty of a question format deviates from the mean value the discriminatory ability of that question appears to become less marked. This was set out in Tables QT 10 (Book 11, page 93) and QT 14 (Book 11, page 94) and shown to be of significance for all question types other than the one-from-five and related exclusion formats where deviation from the mean I.D. value had no effect on the ability of these formats to discriminate between good and poor students.

The two formats with a mean of below 60% namely W/5 and Causal have the greatest value for correlation between the phi coefficient and the deviation of item difficulty from the mean, those between 60% and 70% namely C/4 and T/F, are still significant but those formats with over 70% mean item difficulty do not show significant correlation. The same observation can be made in respect of the ability of an item to discriminate as it moves from the level of 50% in that items of this difficulty discriminate better between good and poor students than items with greater or lesser difficulty. This is true of all question formats except the one-from-five format which again appeared able to discriminate

between/.....

between good and poor students at all levels of item difficulty.

The anomaly of the related exclusion format might be explained by the fact that the students had very little difficulty with the questions in this format and that there were virtually no questions under the 50% item difficulty in this category. If the mean of the format were lowered the related exclusion format might well behave in the same manner as the other formats.

Ebel (1972) and Cronbach (1946) have shown that test reliability is highest when items of middle difficulty are used as opposed to items of high or low difficulty. Their work has been empirical but it appears from the above results that it is a consequence of the enhanced discriminatory ability of items in the middle range of difficulty.

For the one-from-five format, however, there appears to be no relation between item difficulty and discriminative ability.

From these results it would appear that if a test objective were to test basic knowledge with student scores of a mean of over 75% the one-from-five format type of question, where item difficulty and discrimination do not appear to be related, is the only logical choice of question format for this purpose. The other formats with their tendency to show greatest discrimination in the middle ranges of item difficulty would be more suitable for tests which had as their objective the primary task of grading students.

6. CONCLUSIONS.

In the analysis of 577 questions set in the last few years to our students evidence has emerged that:-

i) Students answer certain types of questions more successfully than other types of questions (or that staff demand less ability from the students in certain types of questions than in others.)

ii) There is a difference between the discriminative ability of types of questions.

iii) There is a relation between item difficulty and discriminative ability as measured by the phi-coefficient.

These findings can be summarized as follows:-

1. Type of question and mean student response.

The correctness of student responses varied significantly with different question formats:

i) Despite its forbidding appearance students found the Related-exclusion type significantly easier than all the other formats studied.

ii) Next in order of easiness were the straightforward formats of 1/5 and True/False types of question,

iii) Significantly more difficult than the above were those types of format requiring some additional effort on the part of the student in making his response - namely the C/4, W/5 and Causal.

As these latter three have all proved to be good discriminators of good and bad students, and as their item difficulties approximate more closely to 50% than some of the other types, at which level the highest discriminating ability of items appears to be present, it would seem that they have a valuable contribution to make in the composition of multiple choice tests of required ability. They will also serve a useful function in allowing the examiner to achieve sufficient variation in item types so as

not/.....

not to blunt the students' performance by the boredom of having to wade through upwards of 100 items of the same type.

2. Discriminative Ability and Type of Question.

The discriminative ability of the differing types of questions vary significantly.

i) The T/F format did not reach the 95% level of confidence in discriminating between good and bad students;

ii) The other 5 question types examined all discriminated at the 98% level of confidence, or better;

iii) Outstanding discriminative ability was shown by the Related-exclusion type of format (despite its apparent easiness) which performed at a confidence level of 99,9% despite the relatively small sample (- or perhaps as a result of this.);

iv) The question types studied appeared to group themselves into

a) very good discriminators - Related-exclusion, 1/5 and C/4 - all discriminating with a confidence level of more than 99%

b) Good - the middle discriminators - W/5 and Causal at a confidence level of 98% and

c) Poor - the T/F format which did not reach the 95% level of confidence in our hands.

3. Item Difficulty and Discriminative Ability.

i) It appears that as the item difficulty of a question deviates from 50% (or from the mean item difficulty of this particular format of question as determined by past experience) so will the ability of this item to discriminate between good and bad students decrease;

ii) This effect is not apparent in the 1/5 question format.

This/.....

This may be related to the fact that all other questions formats (except T/F) require additional thought by the student before the response is made to the question, and that in an examination limited by time, as the difficulty of the question increases the student is less prepared to make the required effort;

iii) This effect is also noted in the Related-exclusion type of format, where instead there was a simple correlation in that the more difficult a question was the more it tended to discriminate between good and bad students. The number of questions of this format available for examination was, however, small and on the other hand there were few questions of great difficulty in this series.

iv) It would appear that in order to obtain the greatest reliability from a given test, unless the 1/5 and Related Exclusion formats are being used exclusively, attention should be given to the fact that middle difficulty questions in the other formats will result in better student discrimination than those at the upper or lower end of the scale of item difficulty.

If each type of question is considered in the light of the above findings, the summary as set out in Table QT 12 (Book 11, page 95) might be of assistance to test constructors. These characteristics of the different question types have not been suspected by our examining panel up till now. It would be most interesting to repeat this analysis in the future and see whether the pattern of behaviour exhibited by these formats to date will show changes reflecting the increased confidence in their use by the panel as a result of this knowledge.

CHAPTER VI.

CONFIDENCE WEIGHTING AND RELIABILITY.

1. QUESTION TO BE ANSWERED.

Does the use of confidence marking increase the test reliability of multiple-choice examinations?

2. INTRODUCTION.

General.

Confidence marking must surely be as nearly as old as civilised man himself. Certainly for candidates undergoing testing it has played a part in the successful outcome, or otherwise, since the inception of oral examinations which must have predated written examinations by many thousands of years. There is no examiner who would not justify the contention that the correct answer to a question given with assurance and confidence by a candidate justifies a greater mark than the same answer given with uncertainty or lack of confidence.

The assessment of this degree of confidence is usually a subjective assessment made by the examiner during the course of an oral examination. It may be assessed, again subjectively, during the marking of an essay test when an examiner will differentiate between a candidate who makes a positive correct assertion and one who hedges his answer unjustifiably. Both these assessments are empirically graded by the examiner when he makes the final and critical judgement inherent in his task - that of assigning a numerical value to the body of knowledge presented to him by the student. That the making and evaluation of this assessment is right and proper has long been accepted. Indeed one of the major criticisms levelled at the multiple-choice examination is that no such assessment is possible in the conventional multiple-choice test. (Banesh Hoffman - 1962)

A candidate may guess or inadvertently mark the correct alternative on his answer sheet. He will be rewarded by being credited with the same mark as a student who knows the correct answer and consciously chooses it. In a conventionally marked multiple-choice test this is undeniably true. Carrying the concept a step further it is obvious that there are grades of knowledge and even in the case of two candidates who consciously choose the correct item one candidate may have a stronger conviction as to the correct answer than another. Despite this they will still receive the same credit for their correct answer (as did the student who guessed or inadvertently chose the correct alternative). Although these three students may be graded differently by the other items making up the test as a whole or by procedures to penalise guessing, it is undeniably true that as far as this particular item is concerned all three students

are/.....

are credited with the same mark. Another criticism of the multiple-choice examination is that even if a student does not know the correct answer the format of the multiple-choice paper awards him the opportunity of blindly guessing at the answer with the possibility of a certain percentage chance of success.

These criticisms have been accepted by examiners using multiple-choice testing and, as discussed in Chapter 11 (see Scoring), various strategems have been adopted to overcome the enhanced score obtained in the main by the effect of guessing. While the conventional strategem to counteract this effect has been to penalise the student for any detected wrong answer, guesses or mistakes, it follows that no penalty is incurred by guesses which are correct, and hence non-detectable.

If it were possible to measure the confidence that a candidate may have in the "correctness" of his response the above arguments might well be met. This procedure has been variously termed confidence weighting or confidence testing, and in brief amounts to a scoring system whereby additional weight is afforded a more confident answer than a less confident one. The problem in multiple-choice tests is that neither the computer nor for that matter the manual scorer, is able to make this assessment of confidence. It becomes necessary therefore for the candidate to assess his own confidence in the correctness of his answer and indicate this degree of confidence to the marking agent in the same manner as he indicates his answer to that item. Naturally as a candidate receives a higher mark for a correct answer chosen with great confidence than for a correct answer chosen with less confidence it will be necessary for the respondent to incur a greater penalty for an incorrect answer chosen with great confidence than for an incorrect answer that was chosen with less confidence or guessed, and it is necessary for this corollary to be built into the marking scheme to prevent the exhibition of unwarranted confidence on the part of the candidate.

This procedure has been challenged by the assertion of experts in the field of testing that "confidence doesn't exist" and this procedure has made a negligible impact in the field of educational testing. Shuford (1969 - a) deals with these arguments in an article and proposes that the dialestical exercise in semantics be avoided and the operational results be studied, and states:

"For/.....

> "For this reason, I want to propose an operational test for the validity of confidence measurements, however obtained, and for the meaningful existence of confidence. I will illustrate its use by applying it to some data that we obtained by asking students to state their confidence for each of the alternatives in a multiple choice test. The logarithmic admissible scoring system was used in all cases. Now, to the existence test itself. What would we like to happen when we ask a student to allocate confidence among the possible answers to a question? The student should recall relevant information and then evaluate the quality of this information in terms of confidence. The higher the quality of information, the more assurance is justified on the part of the student. Now, if the student's information bears much relation to reality and if the student is discriminating good from bad information, then we should expect that the more confidence a student places in an answer, the more likely it is that the answer will be the correct one".

Shuford has further provided evidence that, allowing for sampling variation, as more confidence is placed on an answer, the more likely it is to be the correct answer. Shuford's initial experiments were performed on a range of twenty-six (26) grades of confidence. He has subsequently constructed a mathematical model (Shuford 1969-2) whereby a student in a test of knowledge or skill can express his confidence with any degree of precision he desires, and has published this methodology under the name of SCORULE TM. Of interest though this is the application of this method would have posed many problems to us and Shuford's methods were in practice not applicable for us to use in testing a large class of nearly two hundred students in anatomy and I choose a method of capturing confidence which would dovetail with the multiple-choice examination program that I had devised and was using routinely at this stage.

Application of Confidence Scoring to the U.C.T. Model Program.

In presenting multiple-choice examinations to our students one of the criticisms made by the students had been one of the arguments against the validity of objective testing mentioned above, namely, that a student who guesses an answer

correctly/.....

correctly gets the same mark as a student who really knows the answer. I felt accordingly that we should try the effect of confidence marking in the multiple-choice examinations we were presenting to the students of anatomy. In order to present the test to the student two decisions were required:

(1) How many degrees of confidence did we expect our students to recognise and,

(2) What marks would be allocated for each progressive step of confidence.

(1) Degrees of Confidence.

In determining the degrees of confidence he intends using in the test construction the examiner is at the beginning of an open-ended road. He may regard confidence as existing in only two degrees. That is, the candidate thinks he either knows or does not know the answer. This has led to the two-level confidence testing by Sandborgen in Holland as reported by Ahlgren (1969). This dichotomy is fundamentally true, but the essence of confidence marking is the attempt not only to measure the ability of the student to distinguish between these two states of mind and recognise them, but also the attempt, when a student recognises that he knows the answer, to measure how much knowledge he has of the correctness of his answer. The latter follows Shuford's demonstration that the more confident he is in his answer the more likely he is to be correct. In essence we are attempting to measure an opinion - the student's opinion of his knowledge.

Measurement of opinion can be made in two basic ways:

a) by semantic coding which enables the examinee to choose between phrases such as "uninteresting", "slightly interesting", "quite interesting", "very interesting", if, for example, a lecture was being assessed, and "guess", "a little sure", "sure", "pretty sure", "very sure", and "absolutely sure", if an assessment of an answer is being called for. Whatever semantic guide is used it is important that either the size of the

step/.....

step between each grade of confidence be equal (so that marks can be allocated evenly in progression of these steps), or if not equal that the relation of each step of confidence to all other steps of the scale be precisely determined. If one considers the different meaning that words convey from person to person the impossibility of arriving at a precise stepwise semantic scale for a class of about 200 students becomes obvious.

b) The other method of capturing the examinee's estimate of his confidence would be the technique adopted in the main by the psychologists which involves semantically anchoring the two extremes of opinion which would "guessing" at one end of the scale and "positive" at the other end of the scale. These extremes are numerically demarcated and the examinee is then required to assess his confidence on the numerical scale between the two extremes. This can be visualised (Fig. X1) where guessing is represented by 1 and positive by 9 and the candidate's assessment of his answer is circled as 4.

Fig. X1

1 2 3 (4) 5 6 7 8 9

guess positive

By using the scale method the degrees of confidence open for choice can be made very numerous indeed, for example from 0-100, in fact Shuford (1969 -a) used this scale for his investigations. I was not able to find references to any examination as to whether opinions expressed on the numerical scale would correspond to those on the semantic scale. I had already set up a 7-point semantic scale for the evaluation of lectures and in a casual test a

fellow/.....

> fellow lecturer of the department and I scored our evaluation of a lecture we had both attended by these two methods - semantic and numerical. In both cases the results of each evaluation differed from each other and it appeared that scores by the two methods would not correspond, and that it would be obligatory to decide which method to incorporate into the investigations.

While the presentation of a large scale for choice might present no difficulties for those persons accustomed to working in this field and with time to consider their choice it had to be borne in mind that I was devising a test for second year medical students who would have had no experience of these numerical scales, who would be under stress owing to the fact that they were scoring an examination and most important, who would be pressed for time by the M.C.Q. requirements. It was accordingly decided that the slight help that might be afforded the students by a semantic scale was desirable and accordingly such a scale was decided upon.

The next decision was as to how many degrees of confidence a student could be expected to recognise. Obviously a semantic scale of 100 was out of the question. Since the students were novices at confidence assessment it was considered that a small scale was desirable and initially a scale of 3 steps was adopted, (i.e. guess, sure or very sure).

The format of multiple-choice current at that time was that a student marked his answer if he felt he knew it, but if he felt that he did not know the answer he was encouraged to mark 0 or leave it blank, so as to avoid a penalty for guessing incorrectly. If he marked an answer he was then required to signify whether this was a "guess" - confidence level 1 (low), whether he had moderate confidence in this answer, that is "sure" (moderate) confidence level 2, or whether he had strong confidence in his answer, that is "very sure" - confidence level 3 (high).

It was not considered that the appraisal of his answer into these three

levels/.....

levels of confidence would prove a task beyond his capabilities or a burdensome task, even in examination situations. After initial hesitation on the part of the students they were in a main fully capable of coping with these demands, and the difficulties that have arisen appear mainly to be those of knowledge rather than assessment.

Reports of using three levels of confidence (Ahlgren (1969) and A.I. Rothman (1969)) did not suggest that difficulties were encountered due to the task of confidence assessment at these three levels. When Professor E.N. Keen assumed headship of the Department of Anatomy in 1974 he questioned whether the recognition of three levels of confidence was possible for the student (Keen - 1974). He suggested that a student could easily come to a decision as to whether he would give an answer or not, and if he ventured an answer all he could assess was whether he had guessed at this answer or had some confidence, or varying degree, in this answer. Since we had already embarked on the system using three grades of confidence this was continued and the investigation is in respect of these three grades.

(2) The next decision was that of scoring the responses. Ahlgren had proposed a scoring system adopted by Rothman as follows (Ahlgren - 1969):-

Correct	very sure	4/3 +
	fairly sure	1 +
	guess	2/3 +
Incorrect	guess	1/3 +
	Fairly sure	0
	very sure	1/3 -

The existing format of M.C.Q. answer was at that stage - $\frac{1}{n-1}$ for an incorrect answer (which in a 5 alternative question equals - 1/4) and 0 if a student made no attempt. This scoring procedure as explained in Chapter 11 was adapted in order to minimise the effect of guessing.

If/.....

If Ahlgren's scheme were adopted this would have the effect of rewarding the student who guessed an incorrect answer compared to the student who honestly assessed that he did not know the answer. Secondly the fact that a wrong answer of moderate confidence would score the same amount as a non-response would suggest that Ahlgren's scoring scheme would encourage guessing. Ahlgren (1970) has justified the adoption of this scoring scheme on the grounds that student scores on tests marked by this confidence scoring scheme approximate to scores obtained by conventional methods making it more easily understood and hence more readily accepted. He conceded that if one is examining "on the curve" the scale of weighting for confidence levels would make no difference (Ahlgren - 1972).

Scoring of confidence tests hence will depend to some extent on the objectives of the test (as discussed in Chapter 1). Since the major objective of our testing has been to give students full and adequate self-evaluation whenever possible it was considered that the factor of grading should be considered as more important than certification and I adopted a marking scheme were no answer carried no reward or penalty and the levels of confidence were arranged to either side, viz:-

	very sure	3 +
Correct	sure	2 +
	guess	1 +
No answer		0
	guess	1 -
Incorrect	sure	2 -
	very sure	3 -

and these instructions were then conveyed to the student who was required to mark his answer sheet, and if he made a response to indicate his confidence in one of the three grades.

Paton (1971) proposed a scheme somewhat similar to the above namely:-

Correct	very sure	+ 5/3
	sure	+ 4/3
	guess	+ 1
No answer		0
Incorrect	guess	- 1/3
	sure	- 1/2
	very sure	- 1

and in support of his proposed scoring scheme had calculated the consequences of different students strategies using (a) Ahlgren and Rothman's scoring and (b) his own (presented above) Tables CR A1 and CR A2 (Book 11, pages 96/97).

After due consideration of our own marking scheme Professor Keen and I had come to the conclusion that the grades of confidence did not present equal steps. The task facing the student was initially did he know anything about the question or not. If he felt that he did know he would then be required to assess whether he was guessing or had a greater depth of knowledge - and he was required to assess this greater depth into two grades - moderate-sure, or more significantly - very sure.

It was felt that the step between sure and very sure was less than the step between guess and sure, and of about the same magnitude as the step between don't know and guess. It was decided tentatively to allow 5 marks for very sure, 4 for sure and 1 for a guess correct, and -1, -3, -5 for a wrong guess, sure and confident answer respectively, and the results using these scales were tested against the results allocated +3, +2, +1, 0, -1, -2, -3 and also scales using +10, +8, +3, 0, -3, -8, -10 and Ahlgren and Rothman's scale. The highest figures for reliability were obtained using the first of these alternatives and accordingly our revised scoring

scale/.....

scale was as follows:-

Correct	very sure	+5
	sure	+4
	guess	+1
No answer		0
Incorrect	guess	-1
	sure	-3
	very sure	-5

The theoretical consequences of using this scale are set out in Table CR A3 (Book 11, page 98). From this table it can be seen that a student will add to his score only when he knows the correct answer. If he cannot distinguish between two alternatives, by the laws of chance, he will neither score nor be penalized (except in the sure category where the scoring formula is slightly in his favour). If his knowledge is less than the above, by the same laws of chance, he will be penalized to a varying extent dependent on the confidence he has placed in his wrong answers.

The opportunity for the student to enhance his marks by the adoption of any strategy (apart from his true knowledge) is reduced to the absolute minimum and this scoring method allows the most sensitive measurement of knowledge uninfluenced by student strategy. Using this scoring system a large body of the class will score minus marks, and the majority of the class will score below the conventional 50% pass mark. In a grading examination this fact is of no importance. In a certifying examination it is necessary to adjust the student marks. I have adopted the procedure of determining the student confidence Z-score (or standard score), that is the score expressed as a proportion of the standard deviation for that test, above or below the mean. The obtained Z-score is then converted to a score for that student based on a desired mean and desired standard deviation. The desired mean originally adopted was 62,5% and the desired

standard/.....

standard deviation was 10% which means that a student who scores a mark in excess of 1,25 standard deviations above the mean will be awarded an honours mark whereas the student who scores less than 1,25 standard deviations below the mean will be regarded as failing that test.

These criteria operate for the grading system at the University of Cape Town where a pass mark is 50% and the required mark for honours is 75%. Naturally these parameters can be adjusted to suit the requirements of any certifying body.

Payne (1968) considers that this procedure for establishing standard as Z-scores should be adopted for any component of an examination in student evaluation - but we have not put this into effect in any of our examination components other than confidence-marked multiple choice papers.

It was recognized that the student had now been given a task additional to choosing the correct alternative and it was necessary to extend the time allocated for the test. Whereas we had been submitting conventional tests at the rate of 60 questions in 40-45 minutes the time for confidence marked tests was extended to 60 seconds per 5 alternate questions.

Weighting test scores by the evaluation of confidence has practically without exception raised the reliability of test scores since the first published study by Kate Nevner in 1932. (Hevner - 1932). Ahlgren (1969) in a review of 25 studies in the literature records that 24 of these reported an increase in reliability when marked by confidence weighting compared to conventional marking. It was also noted by Ahlgren that greater gains were noticed in less reliable tests than in the more reliable tests. Since reliability is expressed as a ratio with a theoretical upper limit of 1,0 the closer that a test approaches this level in the first instance the more difficult it becomes to increase the reliability ratio.

In/.....

In evidence supporting confidence marking a measure that is often used is that of effective test length. The concept underlying this measure is that it is well known - and as has been demonstrated again in my experimental series- that the reliability of a test increases proportionately as its length increases. (See Chapter IV) (As discussed in Chapter IV, the effect is inherent in the Kuder-Richardson 20 formula).

The effective test length is a ratio of how much longer a conventional test would have had to be to give the same reliability as that given by the confidence marked test under review, and is arrived at by the application of the Spearman-Brown formula. Ahlgren (1969) has challenged the use of effective test length as a measure of the efficiency of confidence weighted tests on the following grounds.

i) Effective test length magnifies the effect of small inaccuracies which may be present in the estimates of test reliability.

ii) The assumption is made that the items that would be added to extend the test would have the same measuring characteristics of the original test.

While it is theoretically reasonable to anticipate that the number of added items performing more efficiently would equal the number performing less efficiently, Ahlgren claims that it is hard to construct good items and implies that as the construction of a test proceeds the efficiency of the added items decreases, and concludes that effective test length ratios underestimate the advantages of confidence marking. While this may well be the impression of any author who has been involved in the construction of multiple choice tests I have not examined this in detail in my study and have no informed opinion on this aspect.

Finally, Ahlgren points out that the time measure of comparison should be based on the time taken to administer the confidence test

to/.....

to the time that would be required to administer the lengthened conventionally marked test, which would effectively lessen the apparent advantage of confidence marking. If the estimated time of the hypothetical length of a comparative conventional test is related to the actual testing time of a confidence test as suggested by Ahlgren to give "test efficiency", a fairer measure of the comparative value of the two methods of testing can be arrived at. This measure appears to be reasonable and has the advantage of presenting for each test a readable and comparative measure of efficiency of test performance.

3. METHODS AND MATERIALS.

All tests with confidence coding were marked by the two different schemes - (i) by using the conventional scheme with a penalty of $\frac{1}{n-1}$ for a wrong answer, and (ii) by using confidence marks of 1, -1, +4, -3, +5, -5 for low, moderate, and high confidence right and wrong respectively.

The reliability coefficient of the same tests marked by these two different methods was determined.

For each confidence test a hypothetical test length for a comparative test length was determined. The estimated time of this hypothetical test length was then calculated, allowing 45 seconds per five alternatives per item and 15 seconds per True or False item. The actual confidence test length was then compared to the estimated time of the hypothetical test length to arrive at a figure for Ahlgren's suggested measure of test efficiency.

This analysis was applied to 10 tests for which data was available to me, i.e. from 271072 to 190475.

4. RESULTS.

The KR 20 formula for the estimation of test reliability for the tests examined as arrived at by conventional scoring and by confidence weighted marking are set out in Table CR 4 (Book 11, page 99).

The difference between the KR 20 estimate obtained by conventional marking and by confidence marking can be represented graphically as in Fig. CR 1. (Book 11, page 100).

The hypothetical test length of a conventional test which would be required to give an equivalent reliability coefficient together with the actual length of the confidence marked multiple-choice paper given and the figure for test efficiency as suggested by Ahlgren are set out in Table CR 5 (Book 11, page 101).

5. DISCUSSION.

Inspection of the Table CR 4 (Book 11, page 99) shows without exception that the same test marked by confidence invariably gave a higher reliability coefficient than when marked by conventional means, and the results as shown in Fig. CR1 (Book 11, page 100) show the elimination of the relation between the reliability coefficient and the number of questions noted in the analysis of conventionally marked examinations and discussed in Chapter IV.

This finding is of great significance to the teacher as by this means it is possible to have a series of short tests, either for self-evaluation or for certification purposes, and be assured that they are reliable. Unless confidence weighting is used the examiner cannot possibly have the assurance, as, even using the technique of cycling (as discussed in Chapter IV), it may not be possible to obtain reliability coefficients of the required level to ensure adequate predictive or concurrent validity. Confidence weighting has accordingly made it possible for us to give multiple self evaluation tests to our students with the assurance that the results are reliable and has enabled us to approach our original aim in establishing multiple-choice examinations in the first place. The degree to which confidence weighted tests have performed more efficiently than conventional tests is set out in Table CR 5 (Book 11, page 101) where the (i) effective test length represents the number of questions that would have to have been set in a conventional test to give the same reliability, (ii) time of equivalent conventional exam is the time in minutes this would have taken and (iii) test efficiency - the time of the equivalent test compared to the actual time of the confidence weighted test actually given. It is of interest to note that as the number of questions in a test approaches 100 the efficiency of confidence weighting falls off. In all tests however, it would have required a greater number of questions marked by conventional means to obtain the same reliability for the test marked by confidence weighting, and the test efficiency as shown in Table CR 5 (Book 11, page 101) has in all cases been well above unity. If tests are grouped according to the number of questions set, test efficiency means are obtained as set out in Table CR6 (Book 11, page 102) where it again can be seen that the confidence weighted tests are more efficient than conventional tests and that the effect is more marked for the shorter tests. It can however be seen

that/.....

that even for the longest tests set the use of confidence marking increases the efficiency of a test by approximately seventy per cent, and accordingly it is established that confidence weighting effectively increases the reliability of multiple-choice tests. These results would answer the question raised by Palva and Korhonen (1973) as to the desirability of using confidence weighting in preference to conventional tests.

The validity of confidence weighted examination has not been examined in this study. It is noteworthy however that the examination of November 1974 which was confidence weighted has given higher correlation figures than previous M.C.Q. exams for raw correlation (Tables V 15 and V 16, Book 11, pages 19/20) and for multiple linear correlation (Tables V 18, Book 11, page 22). This is contrary to the opinion of Hopkins et all (1973), and further study of this aspect is indicated, and this will be undertaken when sufficient results have been accumulated for confidence tests in the final examinations in future years.

6. CONCLUSIONS.

1. The test reliability coefficient for confidence marked tests has on all occasions been extremely high.

2. Confidence weighting has on all tests examined given greater reliability for a test than the same test marked by conventional methods.

3. When tests are compared on a basis of test efficiency, or the time that the two comparative tests would have taken, confidence weighted tests are more efficient in a given time than conventional tests.

4. The shorter the test the more efficient is a confidence weighted test than a conventional test in terms of test reliability.

5. The test reliability coefficient obtained in short tests marked by confidence weighting enables the examiner to set numerous short self-evaluation or class tests with no doubt as to their reliability.

6. Even with lengthy tests confidence marking is significantly more efficient than conventionally marked tests.

CHAPTER V11

ANALYSIS OF CONFIDENCE MARKING

1. QUESTIONS TO BE ANSWERED.

Three questions presented themselves for investigation in this part of my investigation:-

1. Is confidence related to knowledge?

2. Can a student improve his results by the exhibition of an unwarranted degree of confidence?

3. Are the results of confidence marking related to knowledge?

2. INTRODUCTION.

Despite the increase of reliability obtained by testing with confidence marking the procedure would only be justified if it would be shown that whatever was being measured by confidence was in fact related to knowledge.

Shuford (1969 -a,b) claimed that, because the number of times a candidate was correct was greater when he had high confidence in his answer than the number of times he was correct when he had no or low confidence in his answer there was evidence of the relation between confidence and knowledge. To my mind this did not present a clear proof of the relation which could only be considered inferentially, and thus was not beyond doubt. It was felt that the data being collected from confidence marking should allow of more definite proof one way or another than Shuford's inferential conclusion.

If one considered the factors operative in a test it appeared that one could extract at least three independent variables and investigate the relationship between these. The first independent variable is knowledge on the part of the student, the second independent variable is the degree of confidence exhibited by a student in the test, and the third independent variable is the change of marks experienced by that student when marked by confidence weighting as opposed to conventional marking.

If it could be shown that knowledge and confidence measured independently were related we would be able to confirm Shuford's contention that they were in fact related. If confidence and the change in results, again measured independently, were shown to be related we would have an answer to the second question, and if the relationship between the change of results and knowledge could be established we would have an answer to the third question. Precisely how these variables were defined is discussed below. In our analysis of confidence we have ignored the possible effects of personality on the students confidence response. Ebel (1965) has drawn attention to this aspect of confidence weighted tests and the subject remains a problem to be studied.

3. METHODS AND MATERIALS

The initial step in analyzing the changes that occurred in confidence marking was to examine as closely as possible the performance of the individual student. To effect this a computer program was devised to examine the performance of the individual student in detail, to obtain mean scores for performance based on class quintiles and to examine the performance of the questions submitted in the test. An example of the print-out arising from this program is shown in Appendix C (Book 11). In the First Table (CPO 1) the print-out lists

a) the unweighted marks assigned for the confidence marks in the run viz. +1,00 for a right guess, -1,00 for a wrong guess, etc. up to -5,00 for a very sure wrong answer;

b) the number of blocks of questions, the weight for a question in that block and the number of questions in that block; (in this example (180675) 78 questions were of the 5 alternative type and allocated three marks and 22 questions were True or False and allocated one mark each);

c) the numbers of those questions deleted from the marking programme;

d) the lower figure for quintile ranges for conventional marking, designated old, and for confidence marking, designated new.

The next Table (CPO 2) of print-out shows

1. The student number, followed by data for that student;
   - i) the number of questions for which he was very sure and right, - a
   - ii) the number of questions for which he was very sure and wrong, - b
   - iii) below this is given the percentage right and percentage wrong for those questions for which the student was very sure, - c,d,
   - iv) the number of questions sure and right, - e
   - v) the number of questions sure and wrong, - f

vi)/.....

vi) and the percentage of each out of questions answered as sure on the line below, - g; h.

vii) the number of right guesses, - j

viii) the number of wrong guesses, k

ix) the percentage of each of these, -l,m.

x) the number of questions answered correctly by the student, - n.

xi) the number of questions answered wrongly by the student, - o.

xii) on the line below the percentage of questions correct, - p

xiii) and the percentage wrong out of all questions answered, - q is set out,

xiv) the number of questions for which the student was very sure, - r

xv) expressed on the line below as a percentage of all questions set and, - s,

xvi) as a percentage of the answers given by the student, - t

xvii) the number of questions sure, - u

xviii) below as a percentage of questions set and, - v

xix) as a percentage of questions answered, - w

xx) the number of guesses, -x

xxi) as a percentage of questions set and, - y

xxii) as a percentage of questions answered, - z

xxiii) the number of no attempts or no guesses and, - aa

xxiv) below as a percentage of all questions in the test, - bb

xxv) and as a percentage of questions answered, - cc

xxvi) the raw score of the student by confidence marking, - dd

xxvii) and the percentage score of the student by confidence, - ee marking - both designated new i.e. new score and new score %,

xxviii) the percentage score obtained by conventional marking - designated old score %. - ff

xxix) the new quintile (i.e. by confidence) - gg and the, -

xxx) old quintile of the student (i.e. by conventional marking) - hh

xxxi).....

xxxi) the rank of the student by confidence marking (designated N), - jj

xxxii) and the rank of the student by conventional marking (designated 0) , - kk

From the data for each student the performance of the class in quintiles was established. (See Table CPO 3 Appendix C). In the table for each quintile the following data has been recorded:-

1. Very sure and right questions.
   i) as a percentage of those answered very sure, - a
   ii) with the standard deviation of the percentage on the line below, -b,
   iii) total number of questions very sure and right for that quintile, - c
   iv) the percentage very sure right out of those questions answered correctly for that quintile, - d

The same data is recorded for:-

2. Questions answered Sure and Right; - e,f,g,h
3. Questions answered Guess and Right; - i,j,k,l
4. All questions answered correctly; - m,n,o,p
5. Questions answered very sure and wrong; - q,r,s,t
6. Questions answered Sure and Wrong; - u,v,w,x
7. Questions answered Guess and Wrong; - y,z,aa,bb
8. All questions answered wrongly; - cc,dd,ee,ff

Then for each quintile the program records:-

9. Questions answered Very Sure (i.e. both right and wrong)
   i) as a percentage of total questions answered by that quintile, -gg
   ii) the standard deviation on the line below, - hh
   iii) the raw number of questions answered Very Sure (both right and wrong), - ii
   iv) the number as a percentage of all questions answered, - jj

(The/.....

(The slight discrepancy between the figures for (ii) and for (iv) (which should be the same) are due to the different ways these two figures have been arrived at.)

The data is repeated for:-

10. Questions answered Sure (right or wrong); kk,ll,mm,nn
11. Questions answered Guess, and - oo,pp,qq,rr
12. Questions not attempted. - ss,tt,uu,vv

Finally for each quintile:-

13. The total number of questions answered is set out; ww,xx,yy,zz
14. The mean confidence score of that quintile (and on the line below the standard deviation) and, ab, ac
15. The number of students in each quintile, - ad

The program also analyzes each question answered (Table CPO 4 - Appendix C), and prints out for each question:-

i) The question number, - a
ii) the percentage of answers for that question that were guesses, - b
iii) the percentage of those guesses that were correct, - c
iv) the percentage of answers that were assessed as sure, -d
v) the percentage of those answers that were correct, -e
vi) the percentage of answers that were assessed as very sure for that question, - f
vii) the percentage of those very sure answers that were correct and, - g
viii) the percentage of no responses for that question, - h

At the end of the table the means of all these values for all questions is shown - k

From these results a computer program was devised to measure the 3 independent variables of knowledge, confidence and change of results.

1. The measurement of confidence.

This program allocates a confidence score to each student based on the number of questions answered very sure, sure and guess - irrespective of whether they are right or wrong. The confidence score is arrived at by allocating 1 mark for a guess, 2 marks for a score and 3 marks for a very sure answer (all right or wrong). This confidence score is reduced to an index by dividing it by the number of questions answered by the student. The mean confidence score and mean confidence index is determined for the class as a whole and for each quintile. It is therefore possible to obtain a confidence rating for each student either:-

a) Raw - his confidence score,
b) as an Index - the confidence index of the student,
c) the Class Confidence Ratio, i.e. the confidence index of an individual student over the mean confidence index of the class.
d) related similarly to the mean quintile confidence index - the Quintile Confidence Ratio.

While the raw confidence score may be influenced by the number of questions answered by the student the Confidence Index of a student would be a measure of his average confidence for all the questions that he answered. When related to the mean class confidence and mean quintile confidence index it would provide a measure of whether the student is more confident or less confident than the class mean or quintile mean and thus provide some measure of his over or under-confidence.

Table CPO 5 (Appendix C) shows these results from the analysis of the class test of 180675. The mean confidence score and confidence index is expressed for the class as a whole, and for each of the five class quintiles. The student class confidence ratio is obtained by dividing his confidence index by the mean class confidence index, and the quintile confidence ratio by dividing by the mean quintile confidence index.

As/.....

As independent measures of confidence for each student the following were then used to correlate against the variables for knowledge and change in results.

1. Confidence Score
2. Confidence Index
3. Class Confidence Ratio

The quintile confidence ratio was discarded for comparative purposes of measuring as it was considered that the division of the class into quintiles could create some anomalies in that two candidates ranking in succession would be measured for purpose of comparison by two different parameters if in different quintiles.

2. The Measure of Knowledge.

The measure of knowledge presents many problems. We have noted earlier that there are no absolute measures of knowledge and there is inherent in any measuring instrument or test a certain error - the test error. (See Chapter 111).

Another problem was to decide whether one would use external measurements of knowledge, i.e. other tests, or an internal measurement, i.e. arising from the test itself under consideration. The results obtained in investigating the validity of multiple-choice tests (Chapter 111) were convincing enough to allow me to decide that an internal measurement, which is far simpler to obtain, would be adequate in establishing a student's knowledge. The results obtained for reliability in multiple-choice tests (Chapter 1V) was further evidence that the test error in an internal assessment of knowledge was minimal and it was considered that external assessment would not materially reduce this error, if indeed the error would, in fact, be less.

Accepting then the concept that internal measurements of knowledge would be satisfactory the problem still remained as to what precisely could be defined as knowledge. In the first instance knowledge may be measured by the number of questions that the student obtains correct. Similarly the raw number that the student gets wrong may be considered as non-knowledge or a "negative knowledge" score. The question of not

attempting/.....

attempting an answer poses a problem. Can a non-attempt be considered to display partial knowledge on the part of the student in that at least he knows that he doesn't know the correct answer?

The definition of a measure for establishing knowledge is arbitrary and in the final analysis five measures of knowledge were used:-

1. The number of answers correct.
2. The number of answers wrong.
3. The number of answers correct less the number wrong.
4. The number of answers correct plus half the number of the no-attempts (assuming that a non-attempt was in some way partial knowledge).
5. The number of answers correct less the number wrong plus half the number of no attempts.

Each of these five variables was correlated against the variable considered to be independent measures of confidence on one hand and change in results on the other.

3. Measuring of Change of Results.

From the data we had assembled there were 3 different changes that could be recorded:-

1. Absolute: The actual change in the percentage mark attained for a conventionally marked test and the percentage obtained in a confidence marked test. Since the figure for the confidence marked percentage is invariably lower than the percentage obtained in a conventional test the change is always in a negative direction but for the sake of convenience the change has been expressed as positive, and used as one measure of the change of results, (i.e. the greater the change the less the score).

2. Relative: When percentage marks are obtained for confidence marked tests it is not uncommon to find negative scores, and the means are too low to allow these scores

to/.....

to be used for certification without some adjustment. It has been our practice to ascertain the Z-score or standard score of a candidate and for the purpose of certification to convert this score to one based on a desired mean (usually 62,5 per cent) with a desired standard deviation (usually of 10 per cent). This has meant that a student who has a Z-Score of +1,25 will be an honours student and a student with a Z-score of -1,25 will fail in that examination being marked. In order to compare the relative performance between conventional tests and confidence tests, the Z-score was determined for each candidate for each marking method in that test and the difference in Z-score determined. The difference has been used as the second measure of the change of results in using confidence marking.

3. Comparative: The final measure of the change in results has been the change in the ranking of the student when marked by the two methods. In arriving at the measurement of the difference in rank the new rank has been subtracted from the old rank so that a gain in rank is denoted by a positive (+) sign and a fall in rank is denoted by a negative (-) sign. The difference in rank so obtained has been the third measure of change in results which has been examined against the independent variables for knowledge and confidence. Table CPO 6 (Appendix C) shows part of the printout of the results of the analysis of an examination (180675) and for each student the information is set out as follows:-

i) number, - a

ii) number of questions right, - b

iii) number of questions wrong, - c

iv)/.....

iv) number of questions very sure (right or wrong), - d
v) number of questions sure (right or wrong), - e
vi) number of questions guessed (right or wrong), - f
vii) number of questions not attempted, - g
viii) raw confidence score, - h
ix) confidence index, - i
x) class confidence ratio, - j
xi) quintile confidence ratio, - l
xii) confidence percentage score, - m
xiii) conventional percentage score, - n
xiv) difference in percentage scores, - o
xv) new quintile - i.e. by confidence marking, - p
xvi) new rank - i.e. by confidence marking, - q
xvii) old rank - i.e. by conventional marking, - r
xviii) difference in rank, - s
xix) new Z-score - i.e. by confidence marking, - t
xx) old Z-score - i.e. by conventional marking, - u
xxi) difference in Z-score, - v

The data as obtained in these programmes for each student was then correlated and the correlation coefficients for each of the variables defined as independent measures of knowledge, confidence and change in results were recorded. This procedure was followed for each examination for which we had confidence coding data in the department from 1972-1975. Finally the raw confidence score obtained by a candidate in each examination was converted to a percentage score of the maximum confidence score that could have been obtained for that examination, and the number of questions answered correctly, wrongly or not attempted was similarly converted to a percentage for each student per examination. When these variables had thus been standardised they were totalled for each of the nine (9) examinations that were under review, and the correlations between the variables described above as representing confidence, knowledge and change in results were correlated for the series as a whole. The decision as to the statistical significance of coefficients of correlation in this series was taken from the values published by Snedecor and Cochran (1967).

4. RESULTS.

The results of the analysis of all confidence marked tests for which we had data are summarised as follows:-

The percentage of correct answers were noted in each test for each of the confidence categories, guess, sure and very sure. These results are set out in Table CAR 1 (Book 11, page 103).

For each quintile 1 then noted the percentage of questions answered by that quintile as very sure - representing high confidence (whether they were right or wrong) and the percentage of questions answered with low confidence (right or wrong). The percentage of questions answered with high confidence in each quintile is set out in Table CAR 2 (Book 11, page 104), and the percentage of questions answered as a guess (low confidence) in each quintile is set out in Table CAR 3 (Book 11, page 105).

The percentage of questions correct for each response from very sure to guess was also examined per quintile and these results are set out in Table CAR 4-8 (Book 11, pages 106-110).

The mean confidence score for each quintile, the percentage of that mean quintile score out of the maximum confidence score that would have been attainable in that test and the class means are set out in Table CAR 9 (Book 11, page 111).

The mean confidence index per quintile and the mean class confidence index is set out in Table CAR 10 (Book 11, page 112) for all tests analyzed. The next step is analyzing the results obtained was to correlated the independent variables for knowledge, confidence and change in results. Tables CAR 11-15 (Book 11, pages 113-117) show the correlation coefficient obtained between the measure of knowledge:-

1. Knowledge as number of questions correctly answered (Table CAR 11)

2./.....

2. Knowledge as number of questions wrong (Table CAR 12).
3. Knowledge as number right - wrong (Table CAR 13).
4. Knowledge as number right plus $\frac{1}{2}\hat{}^{no-guess}$ (Table CAR 14).
5. Knowledge as number right plus $\frac{1}{2}\hat{}^{no-guess}$ - number wrong (Table CAR 15).

and three measures of confidence as:-

a) Raw Confidence (Confidence Score)
b) Confidence Index, and
c) Class Confidence Ratio. (See Page 117)

The correlation between these variables for the measure of confidence namely,

a) Raw Confidence
b) Confidence Index, and
c) Class Confidence Ratio,

and the variables measuring change namely,

1. Percentage difference in scores (Absolute);
2. Difference in Z-scores (Relative);
3. Difference in rank (Comparative);

are set out in Tables CAR 16-18 (Book 11, pages 118-120).

As a possible measure of over-confidence or under-confidence the Quintile Confidence Ratio was also examined against the change in results and the correlation coefficient obtained as set out in Table CAR 19 (Book 11, page 121).

Three variables for the measurement of knowledge namely:

1. the number of questions right,
2. the number of questions wrong, and
3. the number of questions right less those wrong,

were/.....

were correlated against the variables representing the change in results for the series as a whole. The figures obtained for these correlations are shown in Tables CAR 20-22 (Book 11, pages 122-124).

Finally, the cross correlation obtained for the series as a whole when these variables were standardised as described in the methods used (page 127) are set out in Tables CAR 23-25 (Book 11, pages 125-127).

5. DISCUSSION.

1. The Relation between Knowledge and Confidence.

In analysing the relation between knowledge and confidence from the results in Table CAR 1 (Book 11, page 103), which shows the mean percentage of answers that were correct for each of the tests, it can be seen that when students were most confident (i.e. very sure) they were correct significantly more often than then they were sure or when they guessed the answer. Similarly there is a significant difference between the percentage of correct answers when students were sure of the answer than when they were guessing.

If the confidence responses of the students in each quintile are examined - Table CAR 2 (Book 11, page 104) it can be seen that subject only to three exceptions (starred in the Table) that students in quintile 1 were more confident in their answers than students in Quintile 2 and students in Quintile 2 more confident than students in Quintile 3, and so on. This is seen for each of the nine examinations analysed.

Since Quintile 1 represents the top 20 per cent of the class by mark these results would corroborate the relation between knowledge and confidence by the corollary that when students were scoring well they were more confident than students scoring less well. The converse of this is shown in Table CAR 3 (Book 11, page 105) where it can be seen that, the better students had minimal confidence in their answers (i.e. were guessing) less frequently than the weaker students, and this effect is repeated for each quintile again save for a few exceptions starred in the Table, where the strict sequence is interrupted.

When the relationship of correct answers to confidence is examined for each quintile as set out separately in Tables CAR 4-8 (Book 11, pages 106-110) for Quintiles 1-5 it can again convincingly be seen in each quintile that when students were sure of their answer they were correct significantly more often than when they were sure or guessed.

These/.....

These results alone would appear to provide ample evidence that knowledge is related to confidence as asserted by Shuford (1969 - a,b).

Corroborative evidence however can be obtained when the performance of the class in choosing the categories of confidence for their answers is examined by the quintiles of class performance. Table CAR 2 (Book 11, page 104), which indicates for each examination the percentage of questions answered (correctly or incorrectly) in that test as very sure per quintile, shows clearly that the more knowledgeable students were significantly more confident in their answers than the less knowledgeable students. Conversely Table CAR 3 (Book 11, page 105) shows the percentage of questions guessed at, again correctly or incorrectly, per quintile and shows convincingly that less knowledgeable students were guessing significantly more frequently than the more knowledgeable students. This effect can again be observed when the confidence of the students is examined per quintile.

From Table CAR 9 (Book 11, page 111) which sets out per quintile the mean raw confidence score and mean confidence percentage and from Table CAR 10 (Book 11, page 112) which sets out the confidence index per quintile it can be noted that confidence is higher as a rule in the upper quintiles than the lower quintiles (except for some results denoted by an asterisk in Table CAR 9).

These observations, convincing though they may be, do not furnish a precise answer to the question whether confidence is related to knowledge, and it becomes necessary to examine more closely the relation between knowledge (as measured in the five ways defined in Methods). ~~The numerical relationship between the variables of knowledge and confidence (as measured in the three ways defined in Methods).~~ The numerical relationship between the variables of knowledge and confidence as defined can be examined in Tables CAR 11-15 (Book 11, pages 113-117).

When knowledge is defined as the number of correct answers a highly significant correlation is noted to confidence - measured in each of

three/.....

three different ways - in each of the tests investigated and for the series as a whole (Table CAR 23 (Book 11, page 125)). When the measure of knowledge is based on the number of questions wrong a significant negative correlation is obtained with confidence as measured by the Confidence Index and Class Confidence Ratio - again for each of the examinations investigated and the series as a whole - but only in three of the last four examinations does this relation emerge for the raw student confidence score (Table CAR 12 (Book 11, page 114)), but the latter has disadvantages. (See page 117) When knowledge is measured by the number of questions correct less those wrong (Table CAR 13, Book 11, page 115), or by the number of questions correct plus half the number not attempted, (Table CAR 14, Book 11, page 116) or by the number of questions plus half the number not attempted less the number wrong (Table 15, Book 11, page 117) the same significant results are again observed for each of the examinations investigated and for the series as a whole. (Table CAR 23, Book 11, page 123). From these results we can draw the following conclusions:-

1. When students have a high degree of confidence in their answer examined either for the class as a whole or per quintile, they are more likely to be correct in their answer than when they have a low degree of confidence, and
2. more knowledgeable students exhibit a higher degree of confidence in their answers, as evidenced by their mean raw confidence score and confidence index, than less knowledgeable students.
3. A significant correlation is shown between the independent measures of knowledge and confidence.

From these above conclusions, in answer to the first question investigated, Shuford's contention that confidence is related to knowledge can unequivocally be confirmed.

2./.....

2. The Relation between Confidence and Change in Results.

In investigating the second question, namely does confidence per se influence the results of the examination, the investigational results are not so striking and clear cut.

1. Confidence Score. Table CAR 16 (Book 11, page 118) shows this effect when confidence is measured as a raw confidence score. Since this score is influenced in part by the number of questions answered or omitted it is not felt that this is a good measure of confidence - and it is presented for completeness but will not be discussed further. When the raw confidence score is standardised as a percentage of questions answered for the series as a whole (Table CAR 24, Book 11, page 126) it can be seen that a significant negative correlation exists between confidence, when measured as a standardised percentage and the difference in scores between a confidence marked and conventionally marked multiple-choice examination.

2. Confidence Index. The relation between the confidence index (that is mean confidence per question) and the three variables for the change in results are set out in Table CAR 17 (Book 11, page 119). When the absolute change, that is the column for the difference in percentage marks, is considered it appears from the negative correlations that the more confident students showed less change in marks. As confidence weighting has in our hands invariably given lower marks than conventional marking this would suggest that confidence per se is beneficial to the student. When, however, the column of relative change (the difference in Z-scores) is examined the results are not so conclusive. In two examinations, those of 230672 and 271072, confidence and a rise in Z-scores correlate significantly in favour of confidence being beneficial to the student. In two examinations, those of 230474 and 180674, confidence correlates significantly with a fall in Z-scores (i.e. the converse of the

above/.....

above) and the five examinations remaining show no significant correlation between the confidence and the Z-score. When the series is considered as a whole (Table CAR 24, Book 11, page 126) the correlations between the degree of confidence exhibited by the student measured by the confidence index gives a correlation coefficient of -,252 with the difference in Z-scores and thus for the series as a whole a student exhibiting unwarranted confidence, as measured by the confidence index, would receive a significantly lower mark than that obtained in a conventional test.
When the confidence index is examined in relation to the difference in rank (namely relative change) as set out in Table CAR 17 (Book 11, page 119) the results are again equivocal in individual tests. Confidence results in a rise in rank in the examination of 230672, 271072 and 041174, a fall in rank in the examination of 180674 and no significant change in rank in the five remaining examinations. However, when the series is examined as a whole (Table CAR 24, Book 11, page 126) a high confidence index is associated with a significant fall in rank.

3. Class Confidence Ratio. If confidence is measured by the class confidence ratio the results again are variable. When correlated against the difference in percentage score (Table CAR 24, Book 11, page 126) it would again appear that the more confident students are faring better in their results. This is supported by the significant correlation between confidence and Z-scores as shown in the examination of 230672, 271072 and 041174 but the reverse relation is shown in the examinations of 230474 and 180674 while the results in the later examinations

are/ .....

are equivocal for the relation between confidence and Z-scores. For the series as a whole (Table CAR 24, Book 11, page 126) the correlation of -,220 between class confidence ratio and Z-scores is significant and confirms the suggestion observed above that students are faring worse in a relative way as a result of confidence alone.

When confidence in terms of the class confidence ratio is examined against the difference in rank (i.e. the comparative results) (Table CAR 18, Book 11, page 120), the same contradictions as observed in terms of the relative results are seen (i.e. better results for three examinations and worse for one examination).

For the series as a whole the correlations between change in rank and confidence index of r = -,140 and between change in rank and class confidence ratio of r = -,115 and between change in rank and percentage confidence score of r = -,250 again are significant and thereby suggest that confidence per se means that the student will fare worse.

The difficulty in the interpretation of these correlation figures is twofold. In the first instance confidence per se has been shown to be related to knowledge as best as we have been able to measure the latter. So that in reality the more confident students are in the main the more knowledgeable students and as such should have been able to perform better.

The second difficulty arises due to the fact that the yardstick being used to compare the results of confidence weighting is the results of the conventional test which may be imperfect and thus all the results of changes may well be misleading. If one considers as an analogy a class arranged in ascending order of height being rearranged in descending order of height an orderly relation of change

would/.....

would occur. If however the class had been arranged previously in alphabetical order random change would occur.

Since the relationship between confidence and change in results in individual examination has not given an unequivocal measure of being correlated it would appear that at best we can only conclude in general terms, viz. It does not appear that confidence per se is associated definitely with an improvement in student performance measured in absolute, relative or comparative terms.

We can accordingly answer the second question posed by observing that according to the results in our series in individual examinations it would appear unlikely that a student could influence the results to his advantage by the exhibition of an unwarranted degree of confidence.

If the series as a whole is taken as the determining observation the fact emerges that the student by exhibiting unwarranted confidence will achieve the opposite effect to that intended, and achieve a lower score and rank as evidenced by the significant correlation between the drop in % score, Z-score and rankings on one hand and confidence index and class confidence ratio on the other in the analysis of the series as a whole (Table CAR 24, Book 11, page 126).

3. The Relation between Knowledge and Change in Results.

When one examines the results of correlating knowledge (as defined above) and the change of results expressed in absolute terms for the tests individually it is noticed (Tables 20-23, Book 11, pages 122-124) that the more knowledgeable a student the less change occurs between the per cent

scores obtained in the two marking methods, and that the less knowledgeable, judged from questions wrong (Table CAR 21, Book 11, page 123) the lesser the score he receives.

When knowledge is related to relative and comparative changes anomalous results emerge in that knowledgable students are emerging with lower Z-scores and hence lower ranking positions, and this is observed from the series as a whole (Table CAR 25, Book 11, page 127).

The explanation of this paradox namely, that knowledge correlates well with the absolute change in scores - but not with the comparative or relative change in scores may well be due to the fact, (as discussed under the relation between confidence and change in scores) that the relative and comparative scores observed in a conventional test bear little relation to true knowledge and that changes observed in confidence marking would not be orderly change.

We can therefore conclude from our observations that when the changes in marks obtained by confidence weighting compared to conventional marking are examined there is evidence that knowledge and the change in absolute marks are related, but there is a significant and opposite correlation between knowledge and the change of marks in relative and comparative terms.

It is suggested that this effect may be due to the fact that the conventional examination is deficient in ordering students in terms of their knowledge. The precise reason for this has not been determined by my analysis but may in fact be due to rewards in guessing differentials. If one refers to the percentage of questions correct for each quintile when students were guessing (Tables CAR 4-8, Book 11, pages 106-110) it can be observed that when students guessed they were

correct:-

i)	in Quintile 1	43,0%	
ii)	in Quintile 2	42,2%	
iii)	in Quintile 3	41,9%	
iv)	in Quintile 4	38,3%	and
v)	in Quintile 5	32,8%	of the time.

In a marking scheme wherein an equal mark is awarded for an informed answer and a guess it is apparent that guessing will enhance the scores of the more knowledgeable students. This advantage will still further be enhanced by the fact that the better students are getting fewer answers wrong and are being subject to less penalties than the poorer students. Under these conditions the conventional scoring system is exaggerating the relative difference in knowledge between these two extremes of students.

When the more precise scoring mechanisms of confidence coding are brought into effect the changes occurring in relative (Z-scores) and hence comparative student results (rank) might be due to this difference in favour of the better students inherent and not measureable in the conventional marking system.

In answering the third question we can only record that knowledge is related to change in students' results in that the more knowledgeable students show less of a drop in their marks than the less knowledgeable students, that is, there is a relation in absolute terms. In relative and comparative terms the relation between knowledge and change of results is the opposite and more knowledgeable students are faring less well, and it is queried that this may be due to a reward for better students inherent in the conventional marking system, which cannot be recorded and hence corrected.

6. CONCLUSIONS.

1. Knowledge and Confidence.

i) Students who have high confidence (very sure) in their answer are correct significantly more often than when they have less confidence (sure) in their answer or when they have minimal confidence (guess) in their answer.

ii) Students who score well in an examination answer questions with high confidence (very sure) more than students who do not do so well.

iii) Students who score well answer less questions with minimal confidence than do students who fare less well.

iv) Significant correlations occur between independent variables which represent knowledge on one hand and confidence on the other furnishing absolute proof as to the interdependence of knowledge and confidence.

2. Confidence and Change in Student Results.

v) No evidence exists that the student can influence the results of the examination by the exhibition of unwarranted confidence.

vi) When the series of examinations is considered as a whole there is evidence that the exhibition of unwarranted confidence (or an attempt to manipulate the confidence code without corresponding knowledge) will result in the student obtaining a lower standardised or Z-score and a lower ranking.

3./.....

3. Knowledge and Change in Results.

vii) There is a significant correlation between knowledge and change in results on confidence marking in that knowledgeable students fare better in confidence tests in terms of absolute change. (i.e. actual marks)

viii) Knowledge is significantly inversely related to relative and comparative change of results between confidence and conventionally marked examinations.

ix) It is suggested that this paradox may be due to the fact that the conventional multiple-choice examination is less efficient in ordering students according to their knowledge than a confidence multiple-choice examination, and in fact carries undue rewards for the better students.

CHAPTER V111 CONCLUSIONS.

As discussed earlier I embarked upon M.C.Q. examinations in the department initially as an opportunity for self-evaluation by students. The initial correlations of results with essay questions were encouraging enough for us to include M.C.Q. questions in the final certifying examinations from 1969. Initially the M.C.Q. counted 10% of the written examination but this has been increased over the years and it now counts 50% of the written paper.

Despite the satisfaction within the department in our results it was important that we obtain finite answers to the questions that were being asked and this study was undertaken, which has enabled us to answer these questions.

From a practical point of view a study such as this would have been a mammoth task even a few years ago and has been made possible only by having access to the computing facilities of the University. For the analysis of confidence and results in Chapter VII alone I examined 2,779,000 student responses and recorded and tabulated the results, a task that would have been impossible without computer assistance.

In the other aspects of the analysis at all times sufficient data has been available to ensure statistical validity and where clearcut conclusions have emerged from this analysis they are supported by statistical validity.

It is not claimed that all multiple-choice examinations will exhibit the results that we have obtained, but it is desirable that all departments using multiple-choice examining methods should examine their results to ensure that these results are not inimical to student interest. It should not ever be forgotten that the multiple-choice examination is a powerful tool and care must constantly be taken to safeguard the position of the examinees in such a situation. The procedures adopted by our department are set out later in this chapter.

1. RESUME OF CONCLUSIONS

The conclusions of this investigation - which were discussed in the relevant chapters may be summarised as follows:-

1. Multiple-choice examinations are a justified means of testing knowledge - as evidenced by concurrent validity with conventional methods of examining. (Chapter 111)

2. Correlation coefficients for validity have shown a slow but steady rise over the years suggesting that with experience and the regular analyses of items multiple-choice tests can be improved. (Chapter 111)

3. Multiple-choice tests are also reliable methods of examination, and the reliability of a test is proportional to the number of items in a test. (Chapter 1V)

4. Reliability can be markedly improved by the use of the technique of "cycling" - or the removal of items with poor discriminatory power, based on the actual results in the test prior to the allocation of student marks. (Chapter 1V)

5. The type of question format influences the ability of students to answer the item correctly. (Chapter V)

6. The ability of items to discriminate between good and bad students varies

   i) with the format of the question, and

   ii) with the difficulty of the question in most formats.

   (Chapter V)

7. Confidence weighting of students' answers increases the reliability of multiple-choice tests considerably, and even the longest tests show a significant increase in reliability when marked with confidence weighting. (Chapter V1)

8./.....

8. The validity of multiple-choice examination appears to be enhanced when marked with confidence weighting.

9. Confidence and knowledge when measured independently are highly correlated. (Chapter V11)

10. The display of confidence alone does not enable a student to improve his performance in a test. Indeed the converse appears to hold in that the display of unwarranted confidence will result in the student obtaining poorer scores and ranking. (Chapter V11)

11. The relation between knowledge and the difference in results between conventionally marked and confidence marked examinations varies:

    i) in absolute terms knowledge is correlated with better scores in confidence marked tests than in conventional tests, but

    ii) the more knowledgeable students fare worse in terms of ranking;

    iii) an explanation is advanced to explain this anomaly. (Chapter V11)

2. SAFEGUARDS IN USING MULTIPLE-CHOICE EXAMINATIONS

The findings as summarized above are explicit justification for the continued use of multiple-choice examinations in Anatomy, at least in our department. Their use would also be justified in any other department where the performance of items in tests is analysed and items of poor quality removed. The parameters of performance will be individual and vary from department to department but what matters is that parameters be set, be observed and be acted upon, to ensure reliable and valid examinations for students.

The investigation of confidence weighting in marking examinations has been most rewarding as the use of this technique has enabled us to use short tests for self-evaluation purposes with confidence as to the reliability and validity of these tests. Self-evaluation opportunities are of great help to students, not only allowing them to honestly appraise their success in achieving the objectives of the course without fear of punishment when they are not succeeding, but also in allowing students to acquaint themselves with the format of questions in use and to familiarise themselves with the scoring systems in use. If they so desire they have the opportunity to test their scoring strategies under "battle" conditions, again without being penalised should these strategies not prove successful.

I have been more than gratified i) by the relation between confidence and knowledge which would support the use of confidence weighted scoring methods, and ii) especially the evidence that the exhibition of unwarranted confidence without the requisite backing of knowledge not only does not allow the student to improve his marks but acts in the opposite manner. This finding is especially important in the Medical Faculty, as when faced with problems in medical practice it is vital that not only does a doctor know what he knows, but knows what he does not know. For this reason alone confidence weighting in examinations in the Faculty of Medicine would appear to play an important role. The implications of the findings in improving our results are numerous and while some have been put into operation as part of the ongoing experiment in the department others have not been suspected and will have to be considered in the further construction of tests given in the future.

In/.....

In the main the results of this investigation have confirmed the advantages claimed for computer based examinations, and have confirmed the validity and reliability of this form of testing. In addition the department now has at its disposal an efficient tool, in which it can have full confidence, for evaluating all degrees of cognitive knowledge (Bloom et al - 1956) and which tool can be precisely adjusted in a manner impossible with conventional methods of examination, and which will allow for objective evaluation of differing strategies of educational methods, again a facility unavailable in conventional examining methods.

The results that emanated from the analysis of the performance of differing question formats, which had not been suspected by us, will have to be taken into account in the selection of those questions used for routine testing in the department for the purposes of certification and grading.

The final implication is that any department which sets multiple-choice examinations must be aware of the vast amount of information in respect of the performance of the items that emerges from an examination and must be prepared to devote time not only to the setting of an examination but to its analysis as well. This task is necessary not only to monitor its own success but to ensure that examinations are set which are just and fair to the students concerned. The procedures that we have adopted towards this end may be summarized as follows:

1. In setting an examination paper the departmental question bank is scanned and questions with satisfactory past performance selected.

2. The staff are encouraged to write new questions covering those aspects of the work that we are testing. At no time however do we include more than 30% of new questions in a test, as it is necessary to have a core of established questions to ensure that the Cycling Programme will operate correctly.

3. All questions are submitted to the moderator panel. This panel consists of all members of the department and all new questions are discussed before being submitted to the students. This

is/.....

is one of the most important steps in setting multiple-choice examinations as by this procedure questions which are ambiguous, which are too preoccupied with detail, which display a personal bias of the item writer, or which contain wrong information or answers are eliminated.

4. After the test has been scored the analysis of the performance of the questions is scrutinised and all questions which have not performed adequately in respect of the following four parameters are scrutinised as to clarity, ambiguity and false answers or information. The parameters inspected are:-

a) The percentage of no-attempts: We regard any question in which more than 15% of the class made no attempt as suspect;

b) the percentage of students getting the answer correct. If the item difficulty is below 30% it is apparent that the item is returning minimal information for the time it occupies in the test, and only a few questions of this magnitude of difficulty, used to sort out the top students, have a place in any examination. Similarly questions with an item difficulty of 80% or over provide the examiner with little information. There is a place in an examination which is testing the student body's ability to grasp basic facts for this type of question, but in a grading paper few of this sort of question are called for:

c) the discriminatory ability of the question: Any question with a low discriminative ability is reviewed, as it will have been removed by the cycling programme and will not in fact have been contributing to the examination;

d) Performance of the distractors: If we find a distractor

is/.....

is consistently ignored by the students as being obviously incorrect then the item is no longer a five-choice item and the distractor should be replaced by another alternative which would attract at least some of the students as a possible correct answer.

5. Questions which have been suspect in any of the above parameters are either discarded or if the subject matter is of importance rewritten and resubmitted to the moderator panel.

6. Questions which are satisfactory are passed into the question bank and are available to form a core of questions of proven reliability for subsequent use. We have accummulated many questions which with slight variations cover the same subject but this is not a problem as it enables us to use these different questions in a student year.

7. The stock of questions in the department is now sufficiently large to enable us to allow students to retain the question papers issued to them during the year so that by comparing the correct answer to their own they can derive an enormous learning benefit from the tests.

8. Finally at the end of the year we submit all components of the final certifying and grading examinations to a validity test to ensure that the multiple-choice component is performing adequately in relation to the conventional testing methods.

Above all the department must be frank with students and explain to them fully the scope of the test, the type of questions that will be set and inform them as fully as possible about the methods of scoring that will be used. If applicable the students should also be informed whether any manoeuvres will enhance their score, and what manoeuvres may lower their score.

In our experience entering students in our department have been suspicious of multiple-choice tests initially but the majority have accepted the fairness and objectivity of the method during the course and most express a preference for the method towards the end of their stay in the department.

It/.....

It is fitting that this investigation ends with a tribute to those students of the department who have borne with us, albeit not always patiently, during the development of multiple-choice testing in our hands and who have co-operated, albeit not always willingly, in providing me with the extensive data on which my findings are based.

REFERENCES.

Ahlgren, A. (1969) Reliability, Predictive Validity and Personality Bias of Confidence Weighted Scores. Mimeograph - Centre for Curriculum Studies, University of Minnesota, Minneapolis.

Ahlgren, A. (1970) Personal Communication.

Ahlgren, A. (1972) Personal Communication.

Anderson, J.R., Dykes, J.R.W. and Lennox, B. (1965 - a) 'Recognition and recall questions in 'objective' examination for medical students.' Lancet, 1, 953-955.

Anderson, J.R., Dykes, J.R.W. and Lennox, B. (1965 - b) '"objective" examinations in Medicine.' Lancet, 1, 1333-1334.

Anderson, J.R., Lennox, B. & Low, A. (1964) 'Medical Students Performance.' Lancet, 1, 96-100.

Anderson, J., Wood, H. and Tomlinson, R.W.S. (1968). 'Examination Marking by Computer.' Br. J Med. Educ. 2, 210-212.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956) The Taxonomy of Educational Objectives, Part 1 - Cognitive. Longman Group, New York.

Brooks, C.Mc. (1961). Quoted by Hubbard and Clemans (1961).

Brown, F.G., (1970) Principles of Educational and Psychological Testing. The Dryden Press, Hinsdale, Illinois.

Brown, J. (1966). Objective Tests: Their Construction and Analysis. Longmans, London.

Buckley-Sharp, M.D. and Harris, F.T.C. (1972). 'Methods of Analysis of Multiple-Choice Examinations and Questions.' Br. J Med. Ed. 6, 53-60.

Bull/.....

Bull, G. (1956). 'An Examination of the Final Examination in Medicine.' Lancet, 2, 368-372.

Cooley, W.W. and Lohnes, P.R. (1971). Multivariate Data Analysis. John Wiley and Sons, New York, N.Y.

Cooper, B. and Foy, J.M. (1967). 'Guessing in Multiple-choice Tests.' Br. J. Med. Ed. 1, 212-215.

Cowles, J.T. and Hubbard, J.P. (1952). 'A comparative Study of Essay and Objective Examinations for Medical Students.' J.M. Ed. 27, 14-17.

Cox, R. (1972). 'Value of Objective Examinations.' Nature 237, 489-492.

Cronbach, L.J. (1946). 'Response Sets and Test Validity'. Ed. Psych. Meas. 6, 475-492.

Davison, A. (1967) Personal Communication.

Ebel, R. (1965). 'Confidence Weighting and Test Reliability.' J. Ed. Meas. 2, 49-57.

Ebel, R. (1972) Essentials of Educational Measurement. Prentice Hall, New Jersey, N.Y.

Editorial (1967) Examinations. Br. J Med. Ed. 1, 314-315.

Fielding, M. (1973). Personal Communication.

Gibson, A.L. (1969). 'Second Thoughts on Multiple-Choice Examinations.' Br. J. Med. Ed. 3, 143-150.

Guildford, J.P. and Fruchter, B. (1973). Fundamental Statistics in Psychology & Education. McGraw-Hill, Kogakusha, Japan.

Harris, F.T.C. and Buckley-Sharp, M. (1968). 'Automation of Multiple-Choice Examination Marking.' Br. J Med. Ed. 2, 48-54.

Hevner, K. (1932). 'Method for Correcting for Guessing.' J. Soc. Psych. 3, 359-362.

Hobsley, Michael (1976). 'Assessment of Anatomy in the Primary F.R.C.S.' Ann R.C.S., 58, 382-384.

Hoffman, B. (1962) The Tyranny of Testing. Crowell-Collier, N. York.

Hopkins, K.D., Habistian, A.R. and Hopkins, B.R. (1973). 'Validity and Reliability Consequences of Confidence Weighting.' Ed. Psych. Meas. 33, 135-141.

Horst, J.L. (1966). 'Some Characteristics of Classroom Examinations.' J. Ed. Meas. 3, 293-295.

Hubbard, J.P. and Clemans, W.V. (1961). Multiple-Choice Examinations in Medicine. Lea and Febiger, Philadelphia.

Juritz, J. (1974) Personal Communication.

Kaplan, E. (1971) Personal Communication.

Karsnev, H.T. (1961) - quoted by Hubbard and Clemans (1961).

Keen, E.N. (1974) Personal Communication.

Killcross, M.C. (1968) 'The Use of Machine Readable Sheets with Multiple-Choice Examinations.' Br. J. Med. Ed. 2, 297-300.

Kuder, G.F. and Richardson, M.W. (1937). 'The Theory of the Estimation of Test-Reliability.' Psychometrika, 2, 151-160.

Lennox, B. (1967 - a) 'Marking Multiple-Choice Examinations.' Br. J. Med. Ed. 1, 203-211.

Lennox, B. (1967 - b) 'Multiple Choice.' Br. J. Med. Ed. 1, 340-344.

Lennox, B. (1974) Hints on the Setting and Evaluation of Multiple-Choice Question of the One from Five type. Booklet 3, Ass. for the Study of Medical Education, Dundee.

Lennox/...

Lennox, B., Anderson, J.R. and Moorhouse, P. (1957). 'A Comparative Trial of Objective Papers and Essay Papers in Pathology and Bacteriology Class Examinations.' Lancet 2, 396-402.

Lord, F.M. (1956) 'Sampling Error due to Choice of Split in Split-Halves Reliability Coefficients.' J. Exp. Ed. 24, 245-249.

Lord, F.M. (1957) 'Do Tests of the same Length have the same Standard Error of Measurement.' Ed. and Psych. Meas. 17, 501-521.

Lord, F.M. (1959) 'Tests of Same Length do have the same Standard Error of Measurement.' Ed. and Psych. Meas. 19, 233-239.

Lumsden, J. (1961) 'The Construction of Unidimensional Tests.' Psych. Bull. 58, 122-131.

Lysaught, J.P. and Williams, C.M. (1963). A Guide to Programmed Instruction. John Wiley & Sons, New York, N.Y.

Magnussen, D. (1966) Test Theory. Addison-Wesley, Reading, Mass.

Palva, I.P. and Korhonen, V. (1973). 'Confidence-testing as in Improvement of Multiple-Choice Examinations.' Br. J Med. Ed. 7, 179-181.

Paton, David M. (1971) 'An Examination of Confidence Testing in the Multiple-Choice Examination.' Br. J Med. Ed. 5, 53-55.

Payne, David A. (1968) The Specification and Measurement of Learning Objectives. Blaisdell, Waltham, Mass.

Pullias, E.V. (1937) quoted by Sinclair (1953).

Richardson, M.W. and Kuder, G.F. (1939). 'The Calculation of Test Reliability Coefficients based on the Method of Rational Equivalence.' J. Ed. Psych. 30, 681-687.

Rothman, A.I. (1969), 'Confidence Testing an Extension of Multiple-Choice Testing.' Br. J Med. Ed. 3, 237-239.

Sanderson, P.H. (1973) 'The don't know option in M.C.Q. Examinations.' Br. J Med. Ed. 7, 25-29.

Shuford, E.H. (1969 - a) Systems of Confidence Weighting: Theory and Practice. Mimeograph - APPA order No. 833, Advanced Research Projects Agency, Los Angeles, California.

Shuford, E.H. (1969 - b) Confidence Testing: A New Tool for Measurement. Mimeograph - Shuford-Masengill Corporation, Lexington.

Sinclair, D.C. (1953) 'Objective Examinations in the Medical Curriculum.' Lancet 2, 947-951.

Snedecor, G.W. and Cochran, W.S. (1967) Statistical Methods. 6th Ed. Iowa State University Press, Ames, Iowa.

Stokes, J.F. (1967) 'Examining in the United States: The National Board of Examiners.' Br. J Med. Ed. 1, 320-329.

Swineford, F. (1959) 'Notes on "Tests of the Same Length do have the Same Standard Error of Measurement".' Ed. and Psych. Meas. 19, 241-242.

Wells, L.H. (1967) Personal Communication.

Wells, L.H. (1968) Personal Communication.

Wells, L.H. (1976) Personal Communication.

Young, S. and Gillespie, G. (1973) 'Experience with the Multiple-Choice paper in the Primary Fellowship Examination in Glasgow: Phase 1V of an Educational Study.' Br. J Med. Ed. 7, 16-20.

BOOK 11.

FIG. 1. DISTRIBUTION OF SCORES IN AN EXAMINATION

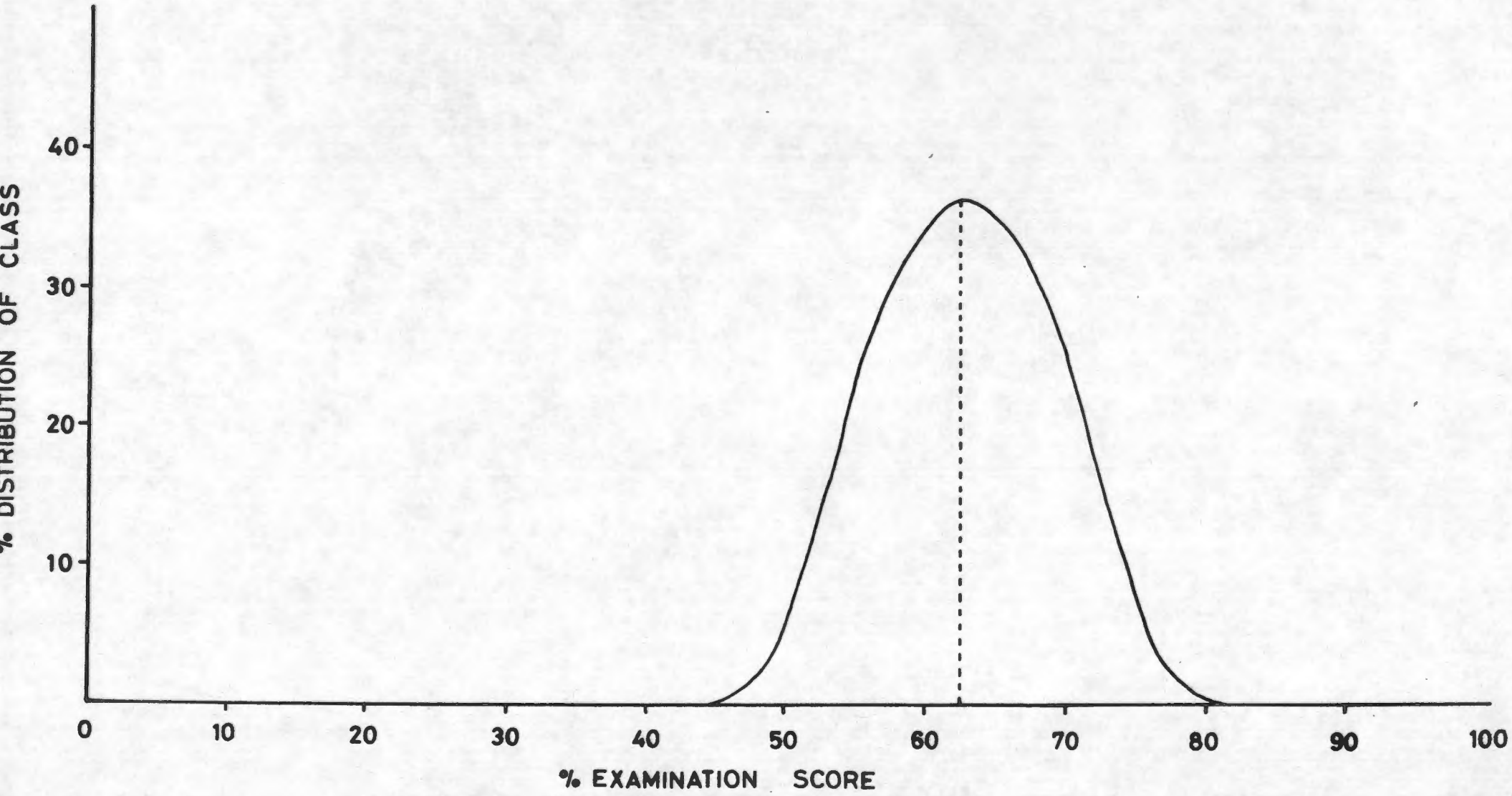


FIG. 2. THEORETICAL DISTRIBUTION OF SCORES FOR CERTIFICATION

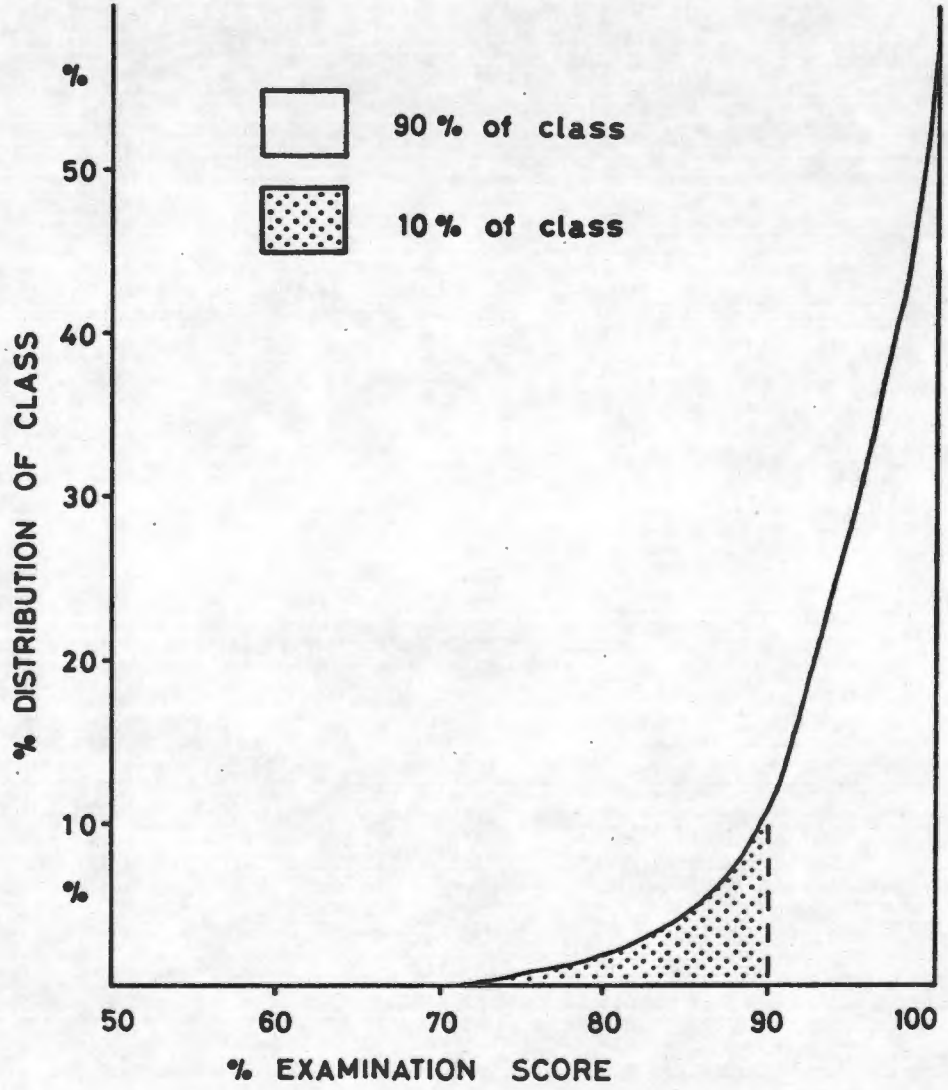


FIG. 3. THEORETICAL DISTRIBUTION OF SCORES IN A GRADING EXAMINATION

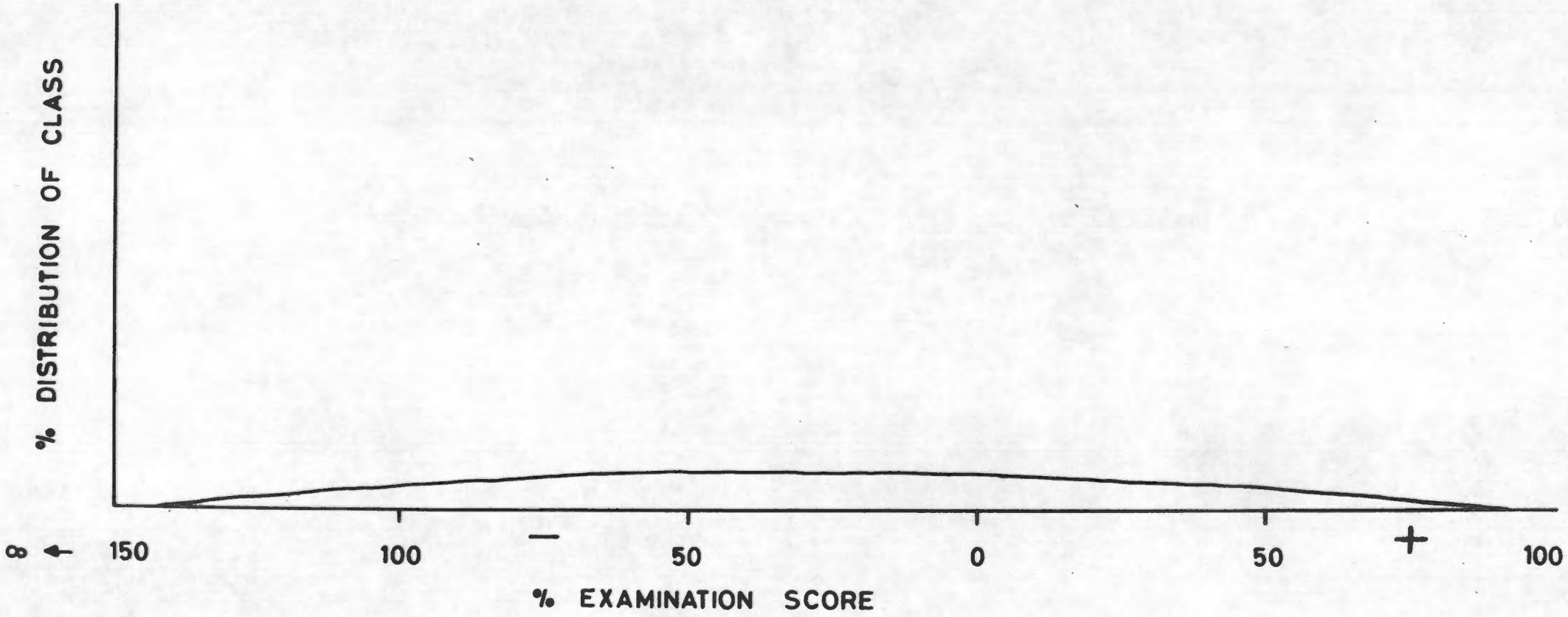


TABLE V 1.

VALIDITY

Mean & Standard Deviation of Components of Final Examination.

Year 1968

n = 168		Mean	S.D.
Class Average		56,3	10,7
Practical		74,2	14,0
Paper 1 Essay	1	57,0	14,7
	2	68,0	10,4
	3	65,4	16,5
Paper 2 Essay	1	75,5	15,7
	2	65,6	9,6
	3	60,9	15,0
M.C.Q.		54,5	15,0
Oral		30,8	13,1

TABLE V 2.

VALIDITY

Mean & Standard Deviation of Components of Final Examination.

Year 1969

n = 183

		Mean	S.D.
Class Average		56,6	9,7
Practical		67,3	12,2
P1. Essay	1	58,4	11,6
	2	56,7	9,5
	3	66,0	13,0
P2. Essay	1	59,0	10,6
	2	49,8	12,6
	3	60,6	9,8
M.C.Q.		59,6	13,7
Oral		58,3	7,3

TABLE V 3.

VALIDITY

Mean & Standard Deviation of Components of Final Examination.

Year 1970

n = 177

		Mean	S. D.
Class Average		56,5	10,0
Practical		69,2	11,2
P1 Essay	1	38,3	14,2
	2	54,1	12,9
	3	61,2	12,9
	4	66,1	13,3
P2 Essay	1	59,7	13,8
	2	55,1	17,0
	3	59,2	15,3
M.C.Q.		58,5	13,0
Oral		59,4	9,0

TABLE V 4.

VALIDITY

Mean & Standard Deviation of Components of Final Examination.

Year 1971

n = 172

		Mean	S.D.
Class Average		63,5	8,3
Ave. Vivas		67,3	8,3
Practical		64,7	11,4
P1 Essay	1	69,4	12,0
	2	68,6	14,4
	3	64,0	11,6
M.C.Q. NA		76,8	9,7
P2 Essay	1	69,4	13,2
	2	47,0	14,0
M.C.Q.		58,4	18,1
Oral		62,1	7,7

TABLE V 5.

VALIDITY

Mean & Standard Deviation of Components of Final Examination.

Year 1972

n = 159

		Mean	S.D.
Ave. Vivas		67,7	4,0
Class Average		60,8	10,0
Practical		68,6	9,8
P1 Essay	1	65,6	10,1
	2	61,3	14,4
M.C.Q. NA		76,7	12,4
P2 Essay	1	53,7	14,5
	2	59,5	16,9
	3	50,8	18,8
M.C.Q. GA		68,7	12,7
Oral		61,9	8,2

TABLE V 6.

VALIDITY

Mean & Standard Deviations of Components of Final Examination.

Year 1973

n = 171

		Mean	S.D.
Class Average		62,5	11,2
Vivas Ave.		71,9	6,0
Practical		76,9	10,7
P1 Essay	1	47,8	14,7
	2	64,8	14,8
M.C.Q. NA		77,0	12,2
P2 Essay	1	81,7	14,8
	2	74,1	12,2
	3	72,4	19,1
M.C.Q. GA		63,5	13,4
Oral		64,4	7,7

TABLE V 7.

VALIDITY

Mean & Standard Deviations of Components of Final Examinations.

Year 1974

n = 177

		Mean	S.D.
Class Average		57,4	9,9
Average Vivas		68,9	6,8
Practicals		69,5	13,0
P1 Essay	1	60,5	8,3
	2	54,2	14,7
M.C.Q. NA		62,4	10,0
P2 Essay	1	78,0	10,3
	2	61,4	10,7
	3	50,3	19,6
M.C.Q. GA		62,3	9,9
Oral		63,7	7,6

FIG. V_1. MEAN % MARK OF 3 COMPONENTS OF FINAL EXAMINATIONS 1968 - 1974.

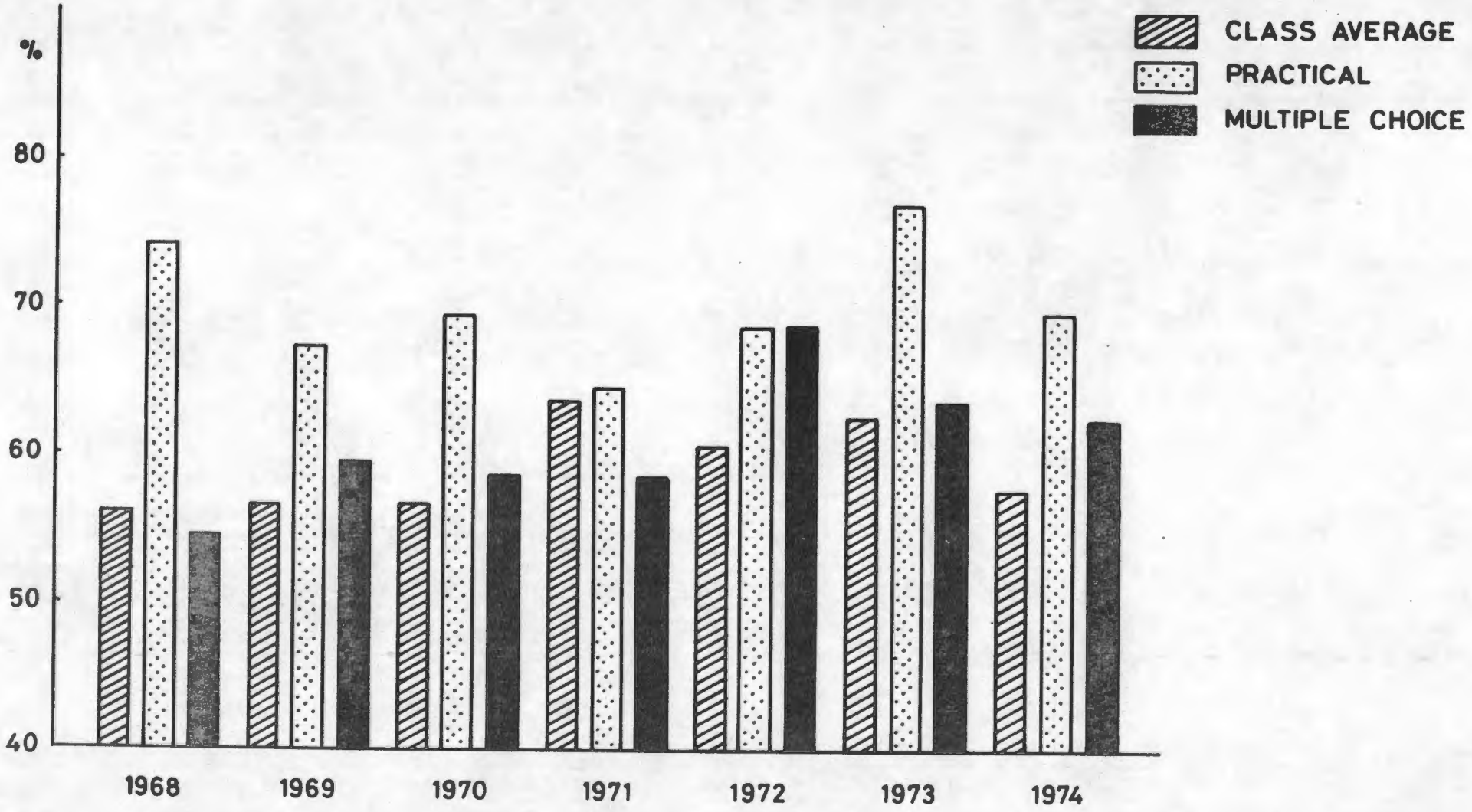


TABLE V 8.

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1968

n = 168

Class Ave.	1,00									
Practical	,776H	1,00								
P1 Essay 1	,540	,535	1,00							
2	,449	,397	,224	1,00						
3	,652	,621	,399	,291	1,00					
P2 Essay 1	,511	,508	,338	,405	,433	1,00				
2	,365	,398	,381	,188	,252	,178	1,00			
3	,598	,526	,442	,351	,632	,455	,314	1,00		
M.C.Q.	,639	,647	,504	,375	,541	,393	,392	,507	1,00	
Oral	-,320	-,333	-,139	-,244	-,351	-,300	-,083	-,197	-,307	1,00

H = Highest Correlation

TABLE V 9.

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1969

n = 183

Class Ave.		1,00								
Practical		,793H	1,00							
P1 Essay	1	,347	,339	1,00						
	2	,570	,437	,321	1,00					
	3	,499	,471	,166	,350	1,00				
P2 Essay	1	,436	,371	,230	,305	,294	1,00			
	2	,466	,442	,362	,382	,327	,348	1,00		
	3	,341	,280	,238	,303	,185	,290	,190	1,00	
M.C.Q.		,647	,614	,376	,354	,513	,427	,493	,320	1,00
Oral		,752	,727	,419	,525	,511	,452	,416	,420	,574	1,00

H = Highest Correlation

TABLE V 10

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1970

n = 177

Class Average		1,00									
Practical		,774H	1,00								
P1 Essay	1	,283	,277	1,00							
	2	,495	,436	,240	1,00						
	3	,281	,271	,243	,210	1,00					
	4	,544	,470	,225	,347	,302	1,00				
P2 Essay	1	,310	,332	,157	,173	,140	,255	1,00			
	2	,395	,395	,189	,337	,226	,201	,274	1,00		
	3	,492	,482	,182	,347	,219	,285	,211	,258	1,00	
M. C. Q.		,725	,706	,292	,453	,309	,489	,334	,384	,462	1,00
Oral		,745	,706	,315	,486	,345	,504	,359	,451	,500	,750	1,00

H = Highest Correlation

TABLE V 11.

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1971

n = 172

Class Average		1,00									
Av. Viva		,544	1,00								
Practical		,778H	,402	1,00							
P1 Essay	1	,415	,225	,471	1,00						
	2	,479	,266	,489	,333	1,00					
	3	,255	,278	,204	,298	,380	1,00				
M.C.Q.	NA	,594	,374	,639	,319	,559	,223	1,00			
P2 Essay	1	-,008	-,035	,044	-,017	-,080	-,119	,017	1,00		
	2	-,126	-,049	-,038	-,130	-,111	-,040	-,037	,469	1,00	
M.C.Q.	GA	,019	,013	,049	-,069	-,063	,076	,083	,458	,435	1,00
Oral	,	,656	,394	,694	,382	,492	,202	,642	,066	,035	,068	1,00

H = Highest Correlation

TABLE V 12.

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1972

n = 159

Average Viva	1,00										
Class Average	,610	1,00									
Practical	,461	,714	1,00								
P1 Essay 1	,329	,481	,512	1,00							
2	,402	,581	,535	,380	1,00						
M.C.Q. NA	,455	,609	,594	,419	,441	1,00					
P2 Essay 1	,431	,594	,523	,488	,491	,484	1,00				
2	,386	,507	,420	,386	,428	,346	,386	1,00			
3	,304	,511	,449	,356	,444	,342	,391	,346	1,00		
M.C.Q. GA	,496	,649	,581	,425	,448	,580	,499	,342	,306	1,00	
Oral	,582	,743H	,683	,511	,622	,665	,605	,504	,478	,675	1,00

H = Highest Correlation

TABLE V 13.

VALIDITY

Correlation Matrix of Component of Final Examination.

Year 1973

n = 171

Class Ave.	1,00										
Av. Viva	,529	1,00									
Practical	,839H	,462	1,00								
P1 Essay 1	,581	,288	,486	1,00							
2	,460	,179	,501	,216	1,00						
M.C.Q. NA	,679	,437	,667	,448	,409	1,00					
P2 Essay 1	,415	,099	,365	,259	,353	,430	1,00				
2	,544	,266	,464	,401	,315	,348	,188	1,00			
3	,317	,169	,254	,175	,051	,319	,281	,073	1,00		
M.C.Q. GA	,769	,406	,769	,433	,408	,697	,368	,397	,291	1,00	
Oral	,664	,468	,655	,506	,346	,578	,406	,411	,305	,557	1,00

H = Highest Correlation

TABLE V 14.

VALIDITY

Correlation Matrix of Components of Final Examination.

Year 1974

n = 177

Class Average		1,00									
Ave. Viva		,677	1,00								
Practical		,803H	,631	1,00							
P1 Essay	1	,363	,255	,397	1,00						
	2	,634	,389	,688	,378	1,00					
M.C.Q.	NA	,671	,480	,726	,215	,606	1,00				
P2 Essay	1	,268	,234	,344	,211	,174	,282	1,00			
	2	,464	,291	,419	,255	,544	,383	,189	1,00		
	3	,352	,225	,404	,228	,367	,334	,172	,253	1,00	
M.C.Q.	GA	,735	,516	,738	,343	,657	,660	,229	,458	,417	1,00
Oral		,598	,521	,635	,383	,529	,554	,205	,422	,327	,614	1,00

H = Highest Correlation

TABLE V 15.

VALIDITY

Correlation of Component to sum of all other Components in Final Examinations 1968 - 1974.

Year		1968	1969	1970	1971	1972	1973	1974
n =		168	183	177	172	159	171	177
Class Average		,435	,830 H	,823 H	,662	,883 H	,875 H	,828
Vivas		N/A	N/A	N/A	,446	,626	,510	,651
Practicals		,369	,769	,779	,717 H	,763	,850	,869 H
Essay	1	,408	,430	,345	,387	,573	,567	,422
	2	,209	,563	,541	,477	,640	,474	,720
	3	N/A	,552	,378	,288	N/A	N/A	N/A
NA	4	,269	,495	,555(M.C.Q.)	,649	,696	,750	,744*
Essay	5	,247	,541	,387	,191	,667	,462	,310
	6	,319	,399	,452	,111	,536	,511	,521
	7	,453 H	N/A	,528	N/A	,521	,332	,428
M.C.Q.		,363	,708	,790	,146	,710	,783	,803*
Oral		-,369	,800	,821	,673	,847	,717	,687+

N/A = Not set in that year
H = Highest Correlation

* = Confidence Z Scored
\+ = Mark of candidate unknown to examiner conducting oral

TABLE V 16.

VALIDITY

Correlation of Multiple-Choice to Non Multiple-Choice Components

	G.A. to Sum all other components	G.A. to non-M.C.Q. components only	G.A. & N.A. to non-M.C.Q. components.
	r =		
1968	,363	,683	
1969	,708	,698	
1970	,790	,805	
1971	,146	,193	,430
1972	,710	,701	,780
1973	,783	,728	,786
1974 *	,803	,842	,823
Mean	,615	,566	,704

* = M.C.Q. Confidence Z Scored

GA = M.C.Q. Question in General Anatomy

NA = M.C.Q. Question in Neuro-anatomy

Non-M.C.Q. = Essays & practicals & orals

TABLE V 17 VALIDITY

Correlation of Multiple-Choice to Non Multiple-Choice questions Multiple Linear Regression and Raw Correlations.

	Multiple Linear Regression Correlation		Raw Correlation	
Year	GA only	GA+NA	GA only	GA&NA
	r =			
1968	,702		,683	
1969	,727		,698	
1970	,825		,805	
1971	,549	,577	,193	,430
1972	,731	,801	,701	,780
1973	,791	,831	,728	,786
1974*	,798	,848	,842	,823
Mean	,732	,764	,566	,704

* = M.C.Q. Confidence Z Scored

TABLE V 18.

VALIDITY

Correlation of Components by Multiple Linear Regression to sums of all others

Year	MCQ	Practicals	Essays	Orals	Vivas	S.A.
	r =					
1968	,702	,796H	,769	,833	N/A	,753
1969	,727	,773H	,715	,828	N/A	,599
1970	,825H	,443	,765	,847	N/A	,651
1971	,577	,766H	,555	,764	,574	N/A
1972	,801	,773	,814H	,855	,838	N/A
1973	,831	,842H	,729	,656	,570	N/A
1974*	,848	,875H	,777	,711	,723	N/A

S.A. = Short Answers

H = Highest value that year

N/A = Not applicable

* = M.C.Q. Confidence Z Scored

TABLE V 19

VALIDITY

Correlations of Coefficients by Multiple Linear Regression for Four Components of Final Examination 1968 - 1974

	M.C.Q.	Essays	Practicals	Orals
Mean =	,7587	,7320	,7511	,7849
S.D. =	,0971	,0845	,1461	,0769
n =	7	7	7	7

Analysis of Variance

	Sum of Squares	DF	Mean Square
Between Groups	,0101	3	,0034
Within Groups	,2629	24	,0110
	,2730	27	

F Ratio = ,3063

TABLE V 20

VALIDITY

Correlation Coefficients by Multiple Linear Regression for 6 Components of Final Examination. 1968-1974.

	M.C.Q.	Essays	Practicals	Orals	Short Answers	Vivas
Mean =	,7857	,7320	,7511	,7849	,6677	,6762
S.D. =	,0971	,0845	,1461	,0769	,0783	,1292
n =	7	7	7	7	3	4

Analysis of Variance

	Sum of Squares	D.F.	Mean Square
Between Groups	,0498	5	,0100
Within Groups	,3253	29	,0112
	,3741	34	

F Ratio = ,8881

TABLE R.1. RELIABILITY

Summary of Results of Multiple-Choice Tests - 1969-1974 - All.

Date	Type	No Q	NS	Mean	Sd	Corr. Mean	SD	KR20
301069	FT	60	183	66,4	11,5	59,7	13,6	,82
220470	CT	40	183	53,7	13,1	45,2	14,7	,76
170670	CT	60	186	56,8	12,5	48,5	14,4	,83
051070	SE	34	88	64,1	16,0	56,0	12,1	,65
051170	FT	75	182	65,3	11,8	58,7	13,8	,86
290371	SE	30	171	62,1	11,2	54,8	12,6	,62
200471	CT	45	176	64,5	12,6	58,1	14,5	,78
140671	SE	30	158	56,9	12,0	48,2	14,2	,66
280671	CT	60	177	65,7	10,5	58,9	12,0	,79
021171	FT	75	183	66,6	14,5	58,2	18,1	,92
030372	SE	30	145	50,3	10,8	40,9	12,8	,62
170472	SE	30	113	60,9	11,9	53,7	13,3	,66
260472	CT	50	169	66,3	13,9	59,8	16,1	,84
120672	SE	30	116	50,4	12,5	40,5	14,9	,70
230672	CT	75	166	62,7	10,8	55,1	12,6	,84
271072	FT	90	163	73,8	10,8	68,5	12,8	,88
060373	SE	24	173	61,3	14,2	54,0	16,5	,70
100473	SE	30	172	67,5	12,5	61,3	14,5	,73
240473	CT	45	178	66,6	13,7	59,8	15,7	,85
110673	SE	30	168	55,0	12,8	40,6	15,2	,72
270673	CT	85	178	57,3	11,6	48,4	13,2	,86
071173	FT	90	176	69,5	11,6	63,1	13,7	,92
050374	SE	30	178	63,0	12,0	57,3	13,7	,61
090474	SE	30	179	61,6	14,1	54,2	16,6	,74
230474	CT	60	181	60,1	11,3	52,3	12,7	,76
180674	CT	100	177	58,0	13,7	39,6	14,7	,72
041174	NFT	149	179	72,3	11,5	67,2	13,4	,99
051174	GFT	134	179	64,0	10,0	48,6	13,1	,99

FT = Final Test
CT =Class Test
SE = Self Evaluation Test

TABLE R.2. RELIABILITY

Means and Standard Deviation of Variables - All Tests.

n = 28	Mean	SD
No. of Questions	57,89	32,77
No. of Students	167,04	23,67
Raw Mean Score	62,24	5,89
St. Dev. Raw	12,34	1,40
Corrected Score	53,97	7,85
St. Dev. Corr.	14,13	1,48
KR20	,779	,108

TABLE R.3. RELIABILITY

Means and Standard Deviation of Variables - (Class Tests)

n = 17	Mean	Standard Deviation
No. of Questions	76,06	30,27
No. of Students	177,41	6,22
Raw Mean Score	64,10	5,48
St. Dev. Raw R.S.	12,08	1,32
Corrected Score	55,86	7,78
St. Dev. Corr. C.S.	14,07	1,51
KR 20	,847	,076

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - NUMBER OF STUDENTS

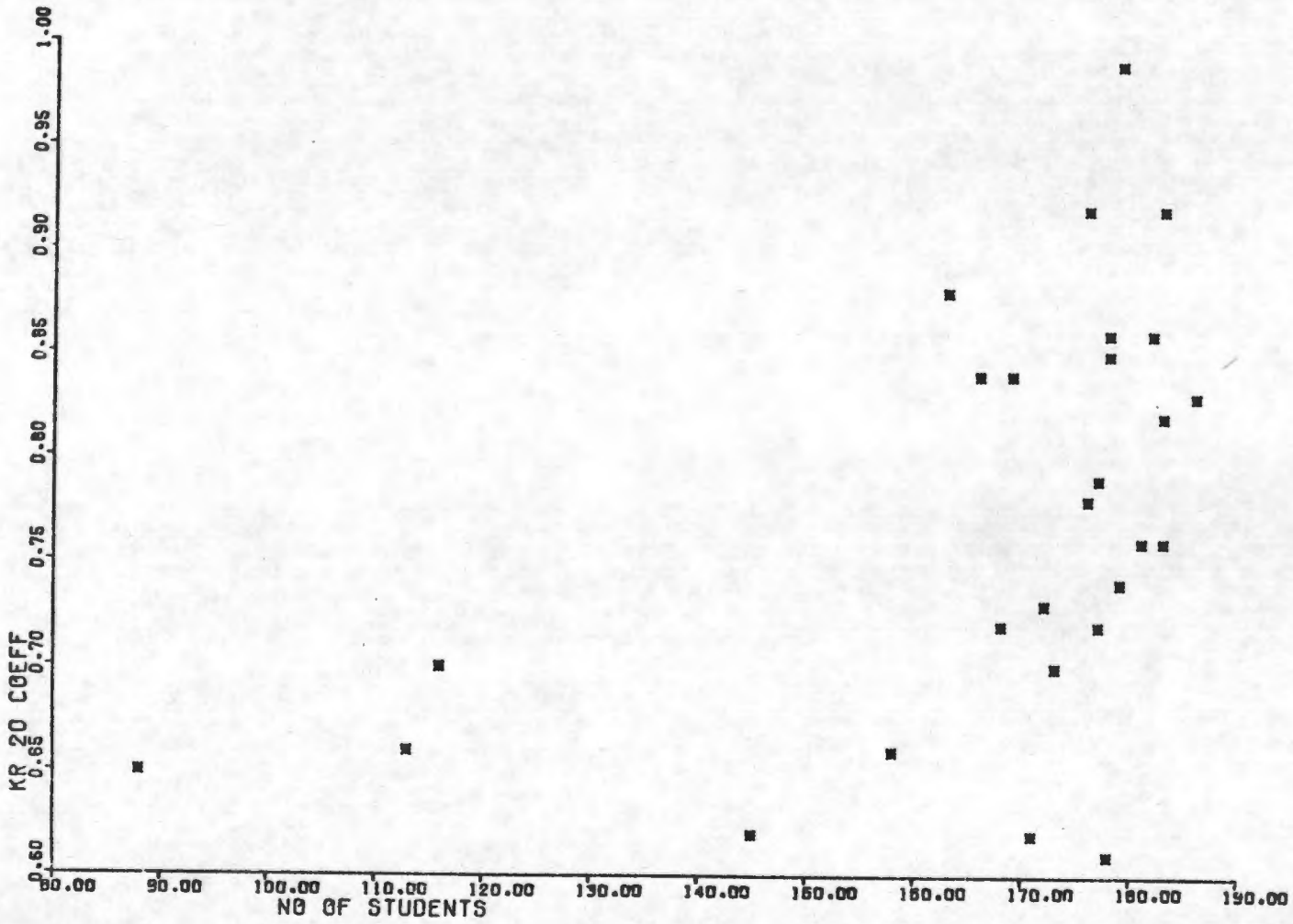


Fig. R1

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - NO OF QUESTIONS

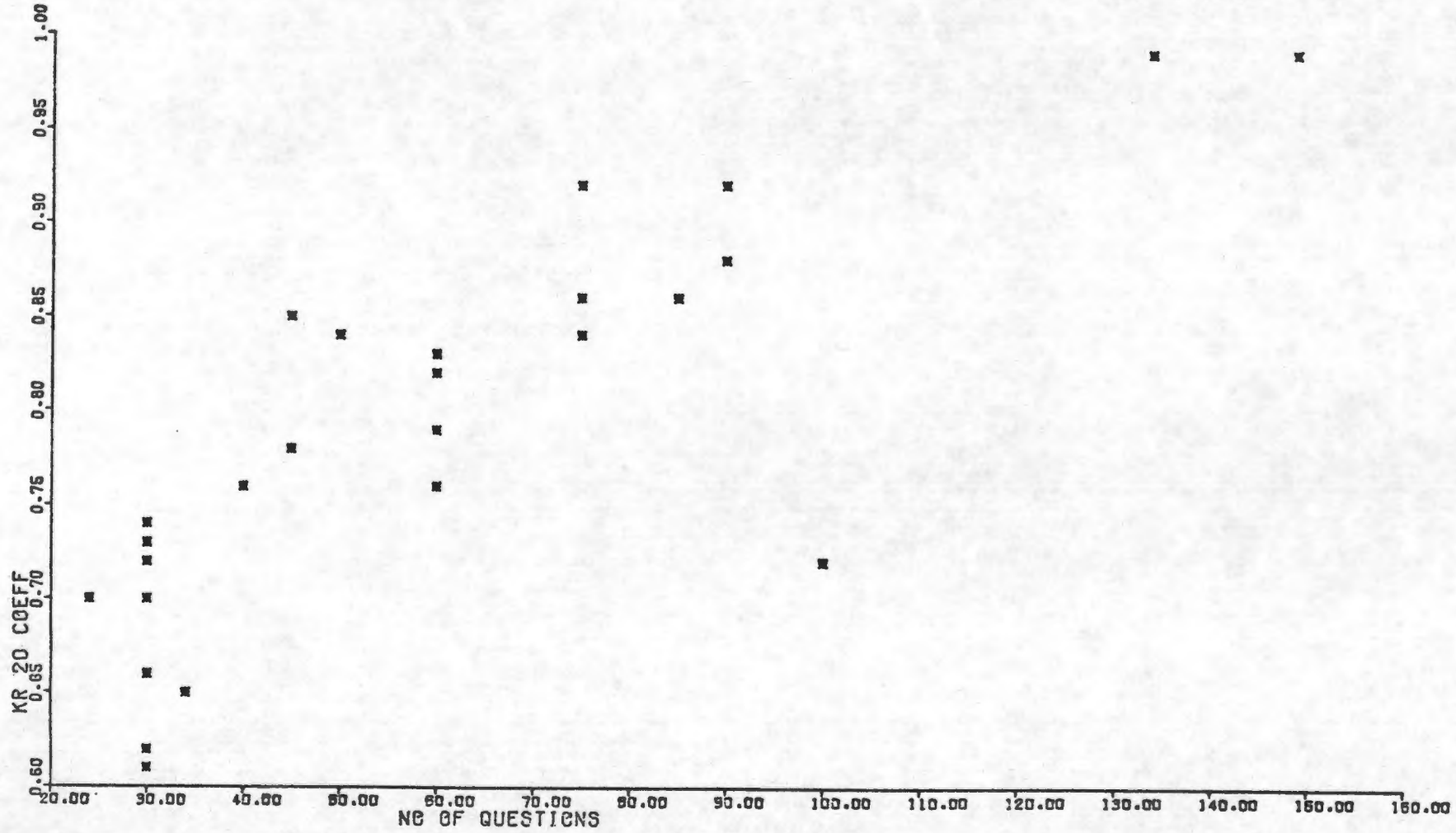


Fig. R2

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - RAW MEAN SCORE

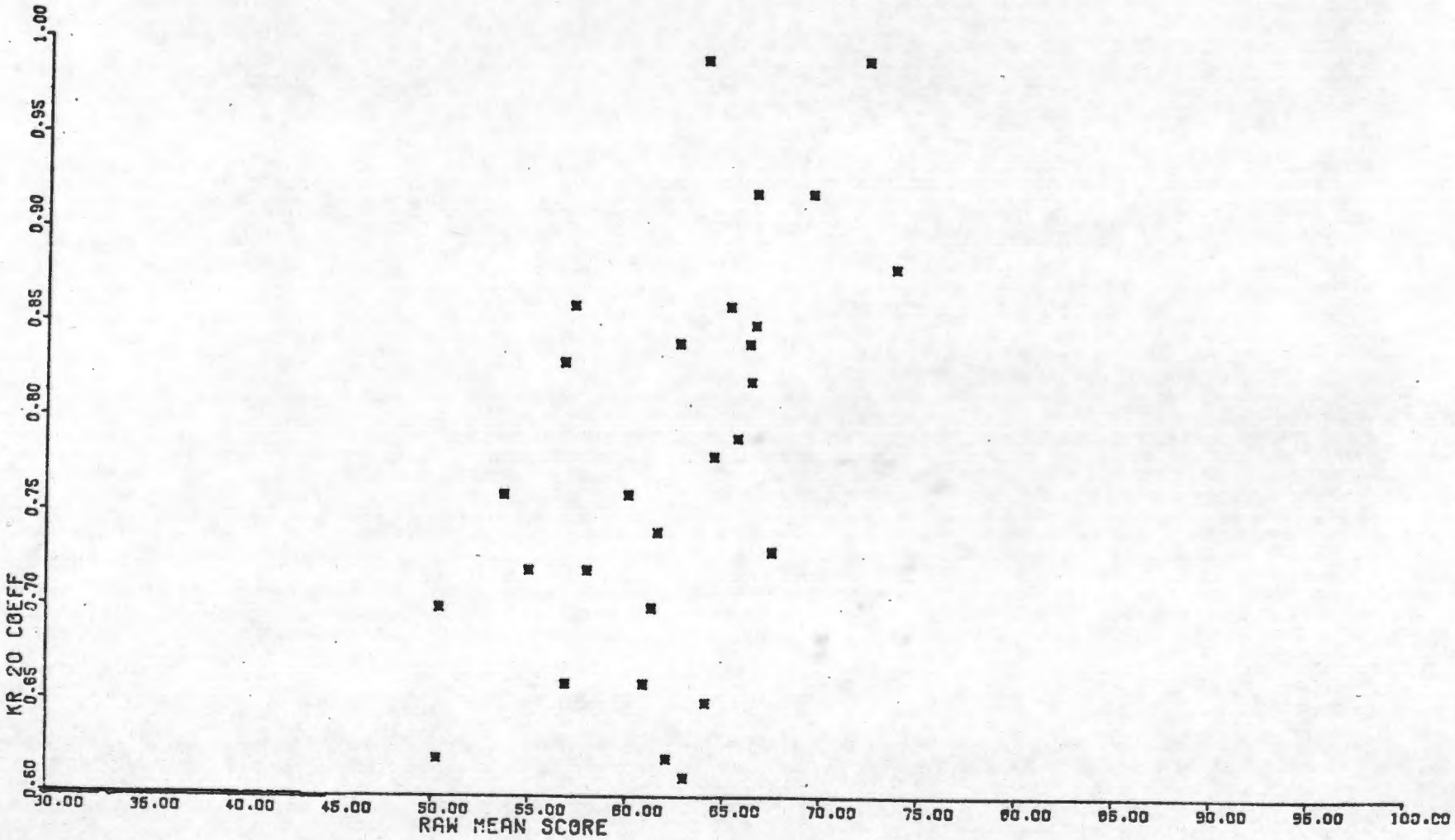


Fig. R3

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - RAW STANDARD DEVIATION

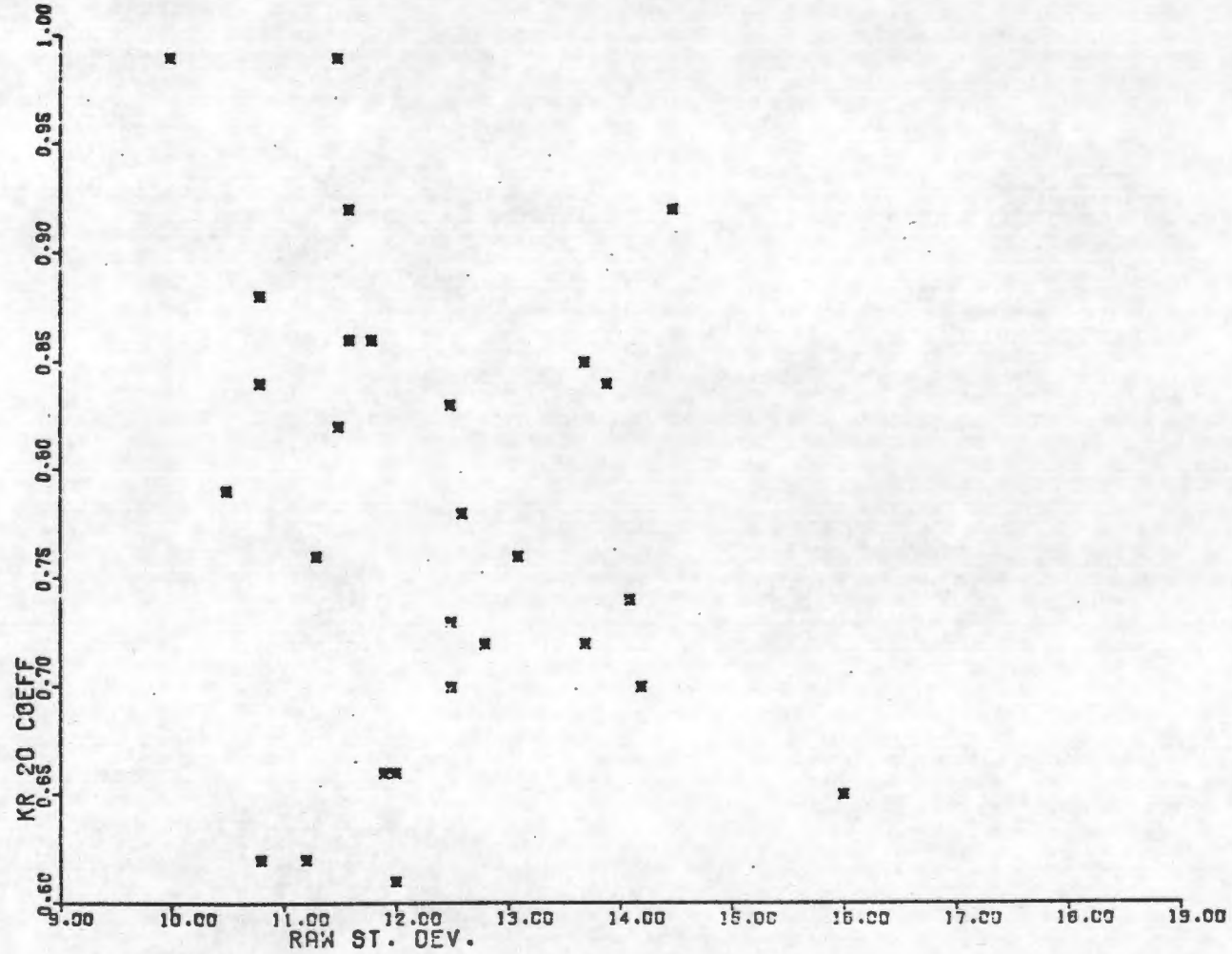


Fig. R4

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - CORRECTED MEAN SCORE

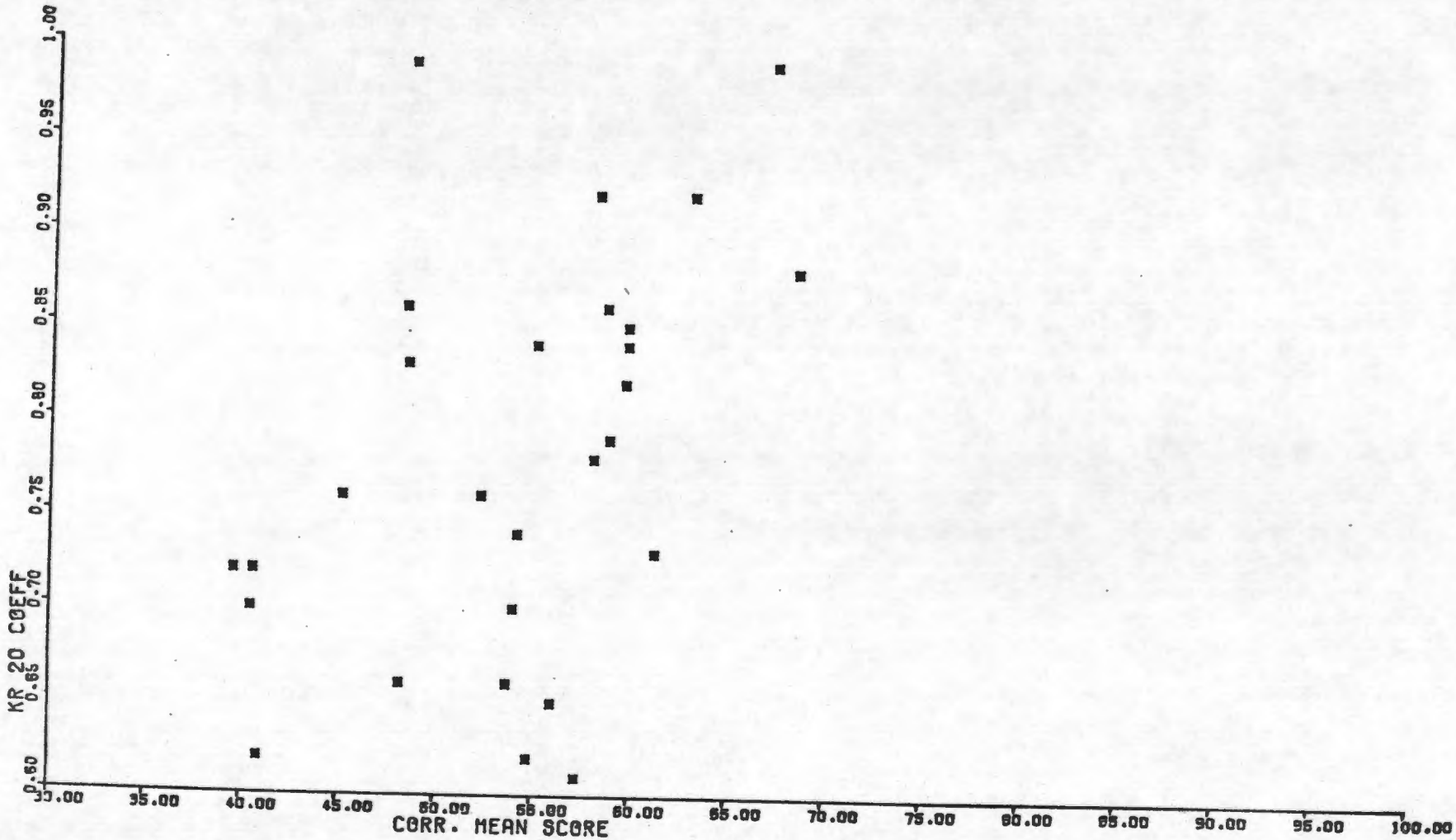


Fig. R5

RELATION OF VARIABLES TO TEST RELIABILITY

ALL TESTS - CORRECTED STANDARD DEVIATION

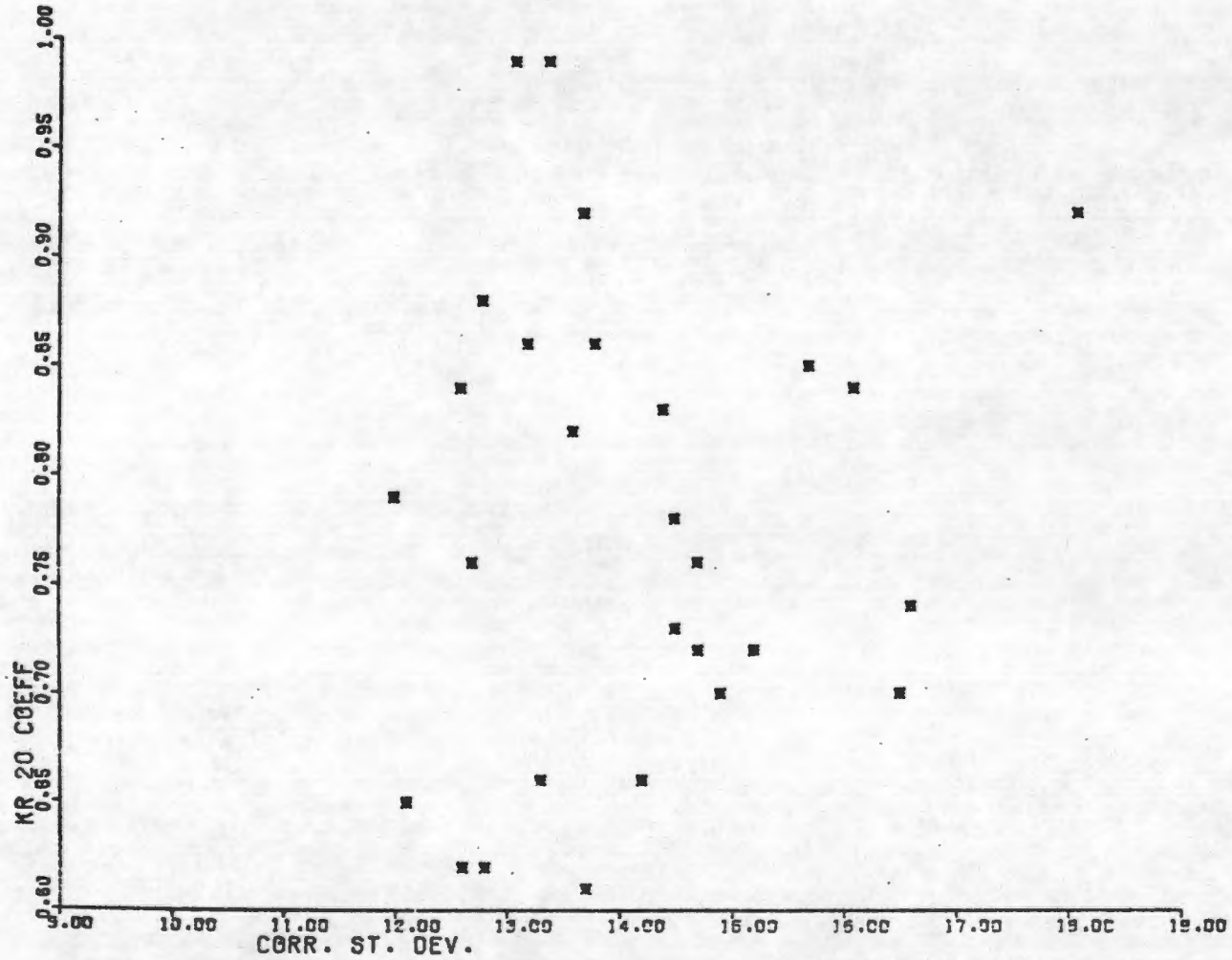


Fig. R6

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - NO OF QUESTIONS

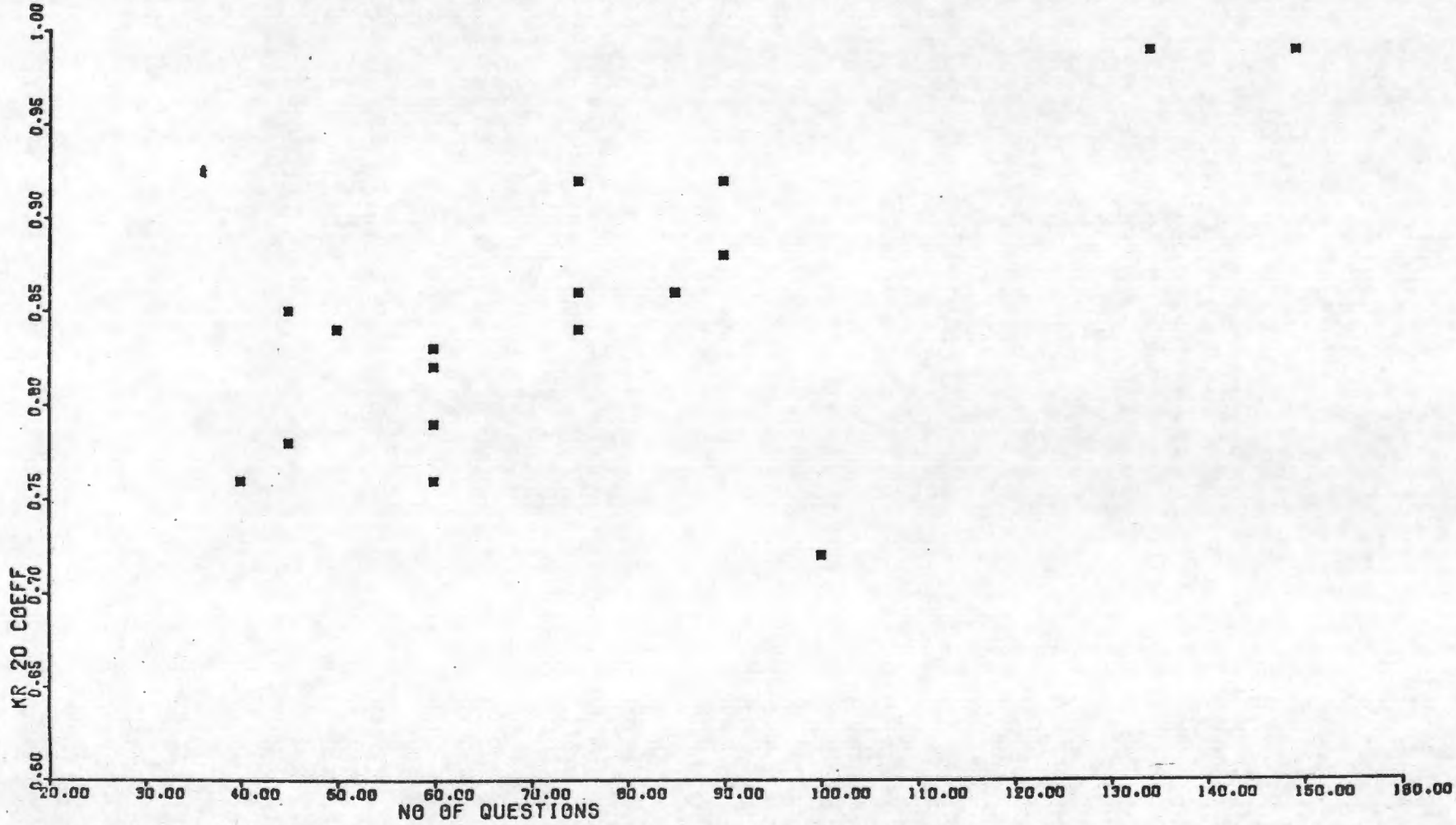


Fig. R7

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - NUMBER OF STUDENTS

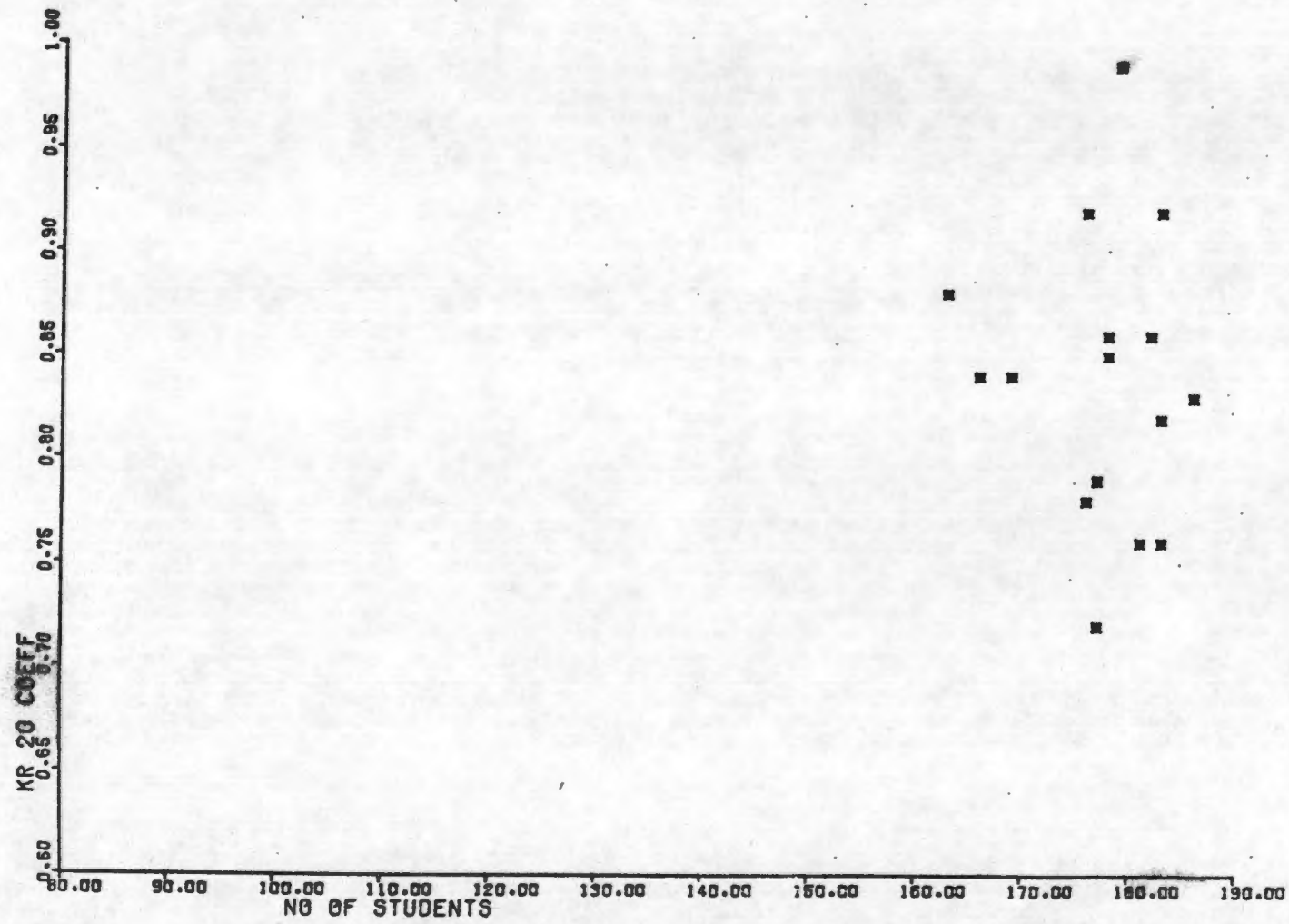


Fig. R8

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - RAW MEAN SCORE

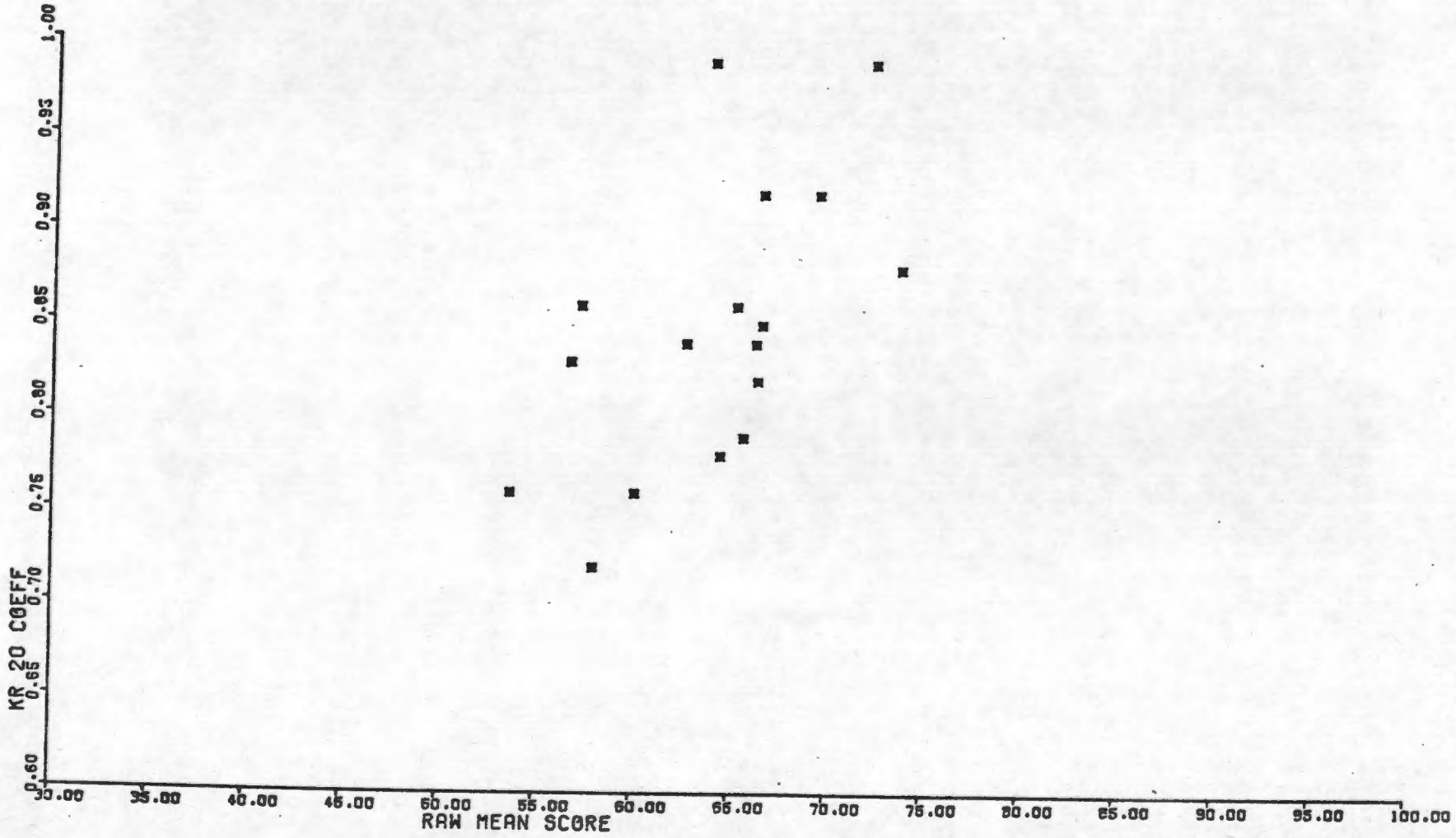


Fig. R9

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - RAW STANDARD DEVIATION

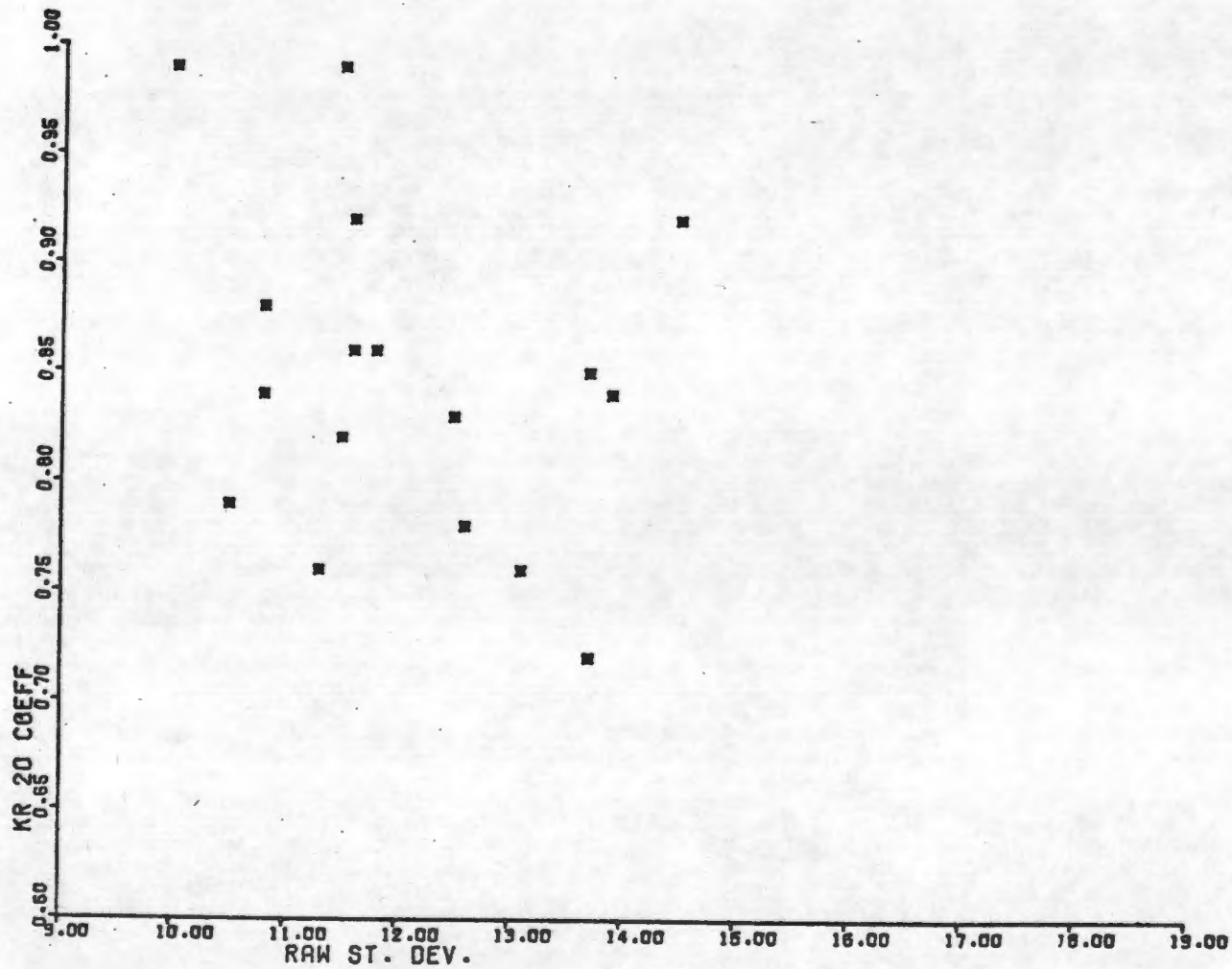


Fig. R10

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - CORRECTED MEAN SCORE

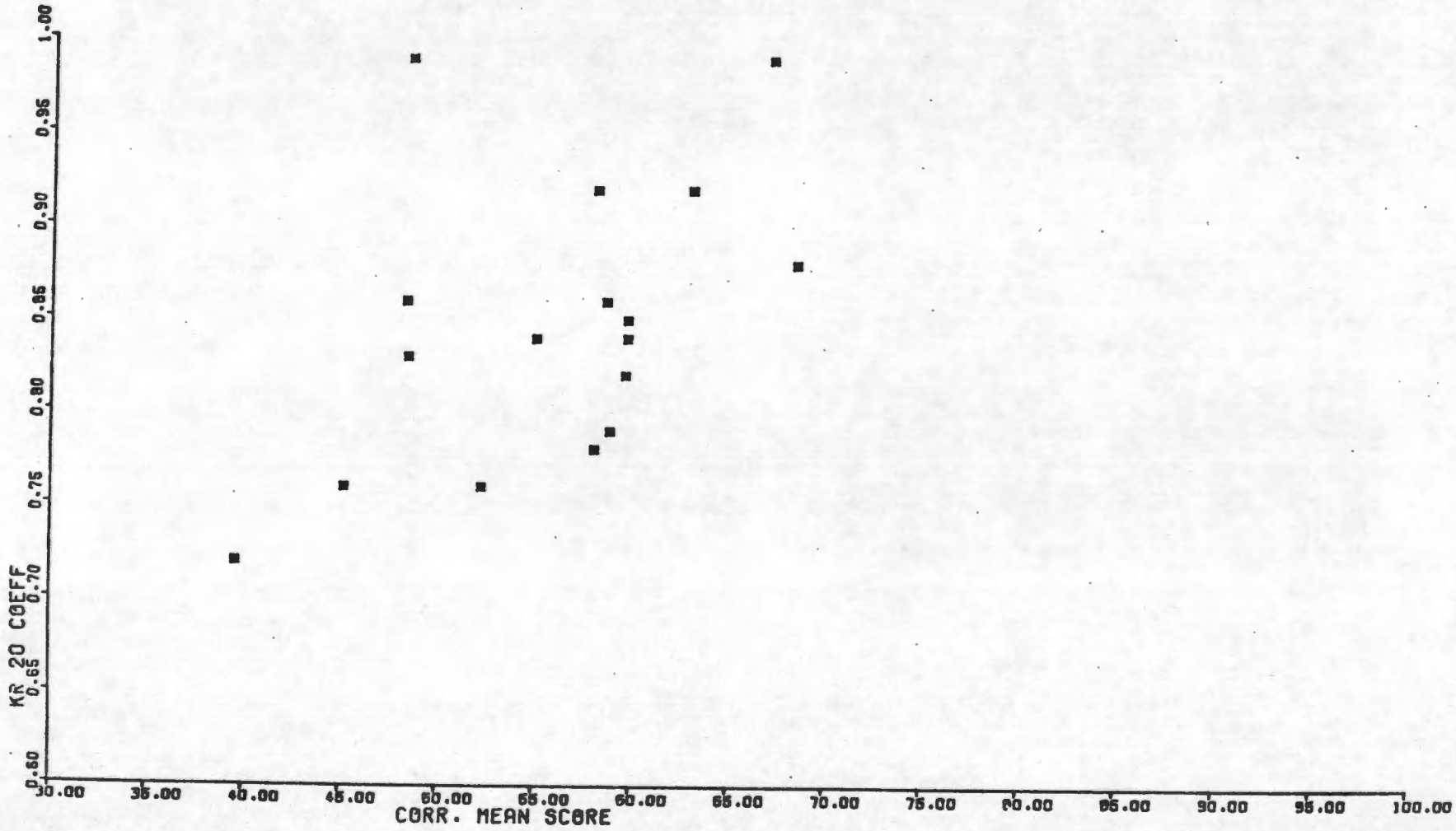


Fig. R11

RELATION OF VARIABLES TO TEST RELIABILITY

CLASS TESTS - CORRECTED STANDARD DEVIATION

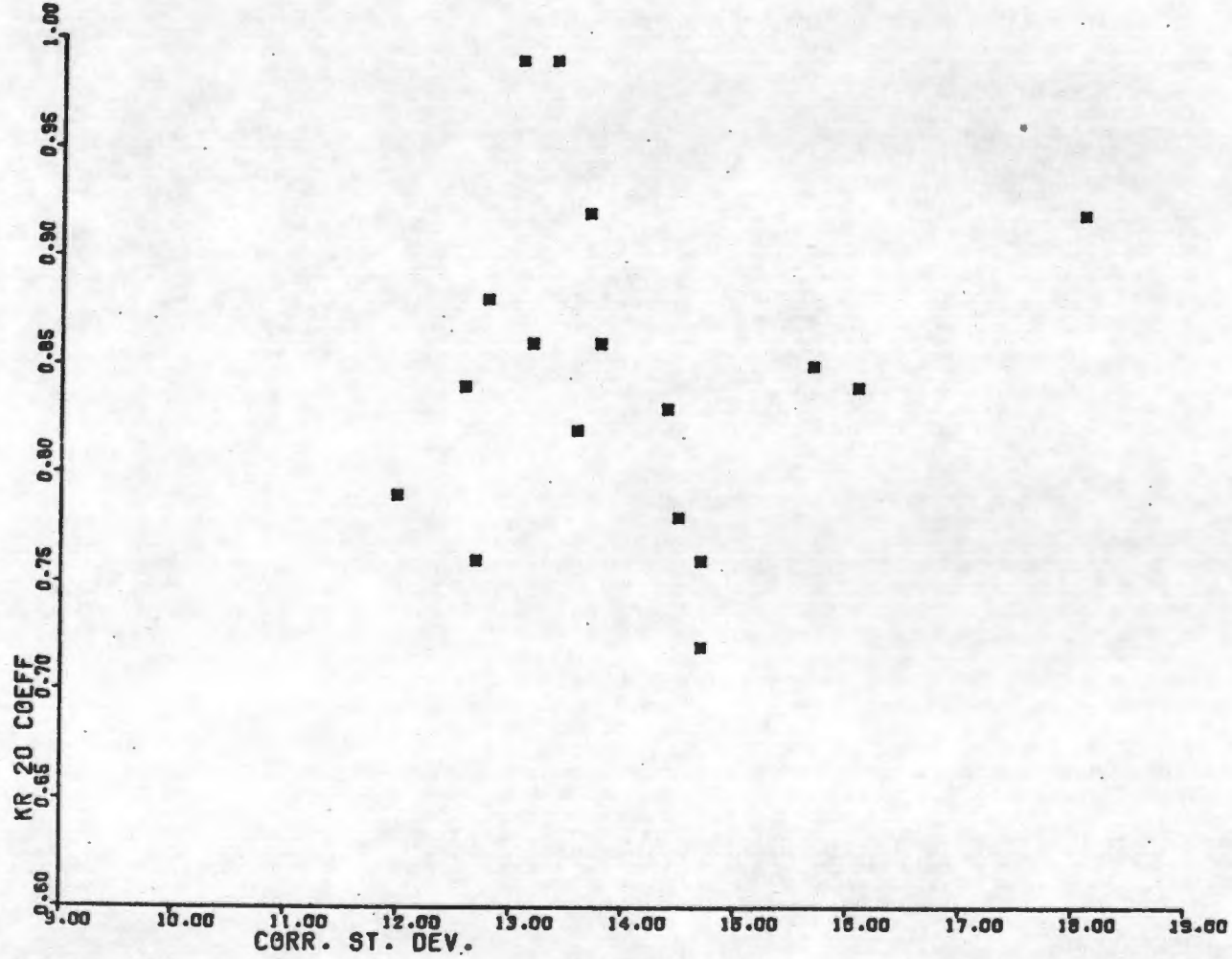


Fig. R12

TABLE R.4 RELIABILITY

Correlation of Variables - All Tests.

n = 28

No. of Questions	1,00						
No. of Students	,359	1,00					
Raw Score	,462	,250	1,00				
St. Dev. Raw	-,369	-,253	-,057	1,00			
Corrected Score	,269	,196	,941*	,064	1,00		
St. Dev. Corr.	-,209	,269	-,055	,638*	-,074	1,00	
KR 20	,824*	,479	,561*	-,220	,443	,100	1,00
	No.Q	No.S	RS	SDRS	CS	SDCC	KR 20

* = highly significant p = <.01

TABLE R.5 RELIABILITY

Correlation of Variables - Class Tests.

n = 17

No. of Questions	1,00						
No. of Students	-,083	1,00					
Raw Score	,354	-,442	1,00				
St. Dev. Raw	-,413	,226	-,215	1,00			
Corrected Score	,100	-,385	,921*	-,192	1,00		
St. Dev. Corr.	-,280	,242	-,038	,918*	-,054	1,00	
KR 20	,691*	-,057	,592*	-,281	,488	,031	1,00
	No.Q	No.S	RS	SDRS	CS	SDCS	KR 20

* = highly significant p = $<.01$

TABLE R.6 RELIABILITY

Changes In Variables When Items With Poor Phi Coefficient Are Removed.

Date. 26.04.72 n = 169 No. of Questions = 50

Cycle	ø	KR 20	% Q. Left	% Q. Del.	T.D.	Corr. M. Score	St.Dev.
1	0	,877	100,0	-	,663	59,8	16,1
2	,10	,884	82,0	18,0	,675	61,3	18,1
3	,12	,883	80,0	20,0	,671	60,8	18,3

TABLE R.7 RELIABILITY

Changes in Variables when Items with poor Phi Coefficient are Removed.

Date. 23.06.72 n = 166 No. of Questions = 75

Cycle	ø	KR 20	% Q. Left	% Q. Del.	T.D.	Corr. Mean Score	St. Dev.
1	-	,866	100	0	,627	55,1	12,6
2	,10	,822	82,7	17,3	,653	58,4	14,6
3	,12	,855	77,3	22,7	,654	58,6	15,3
4	,14	,886	69,3	30,7	,661	59,6	16,0
5	,16	,887	62,7	38,3	,649	58,1	17,0
6	,18	,887	50,7	49,3	,624	55,3	18,9

TABLE R.8 RELIABILITY

Change in Variables when items with poor Phi Coefficients are removed.

Date. 27.10.72 n = 163 No. of Question = 90

Cycle	ø	KR 20	% Q.left	% Q.Del.	T.D.	Corr.Mean	St. Dev.
1	-	,904	100	-	,738	68,5	12,8
2	,10	,905	72,2	27,8	,703	64,4	15,6
3	,12	,902	65,6	34,4	,693	63,2	16,2

TABLE R.9 RELIABILITY

Change in Variables when items with poor Phi Coefficients are Removed.

Date. 27.06.73 n = 178 No. of Question = 85

Cycle	ø	KR 20	% Q.Left	% Q. Del.	T.D.	Corr. Mean	St. Dev.
1	-	,876	100	-	,573	48,4	13,2
2	,10	,895	80	20	,590	50,6	15,9
3	,12	,900	74,1	25,9	,596	51,2	16,8
4	,14	,892	56,5	43,5	,559	46,6	18,8

TABLE R.10 RELIABILITY

Change in Variables when items with poor Phi Coefficients are Removed.

Date. 23.04.74 n = 181 No. of Questions = 60

Cycle	ø	KR 20	% Q.Left	% Q. Del.	T.D.	Corr.Mean	St. Dev.
1	-	,850	100	0	,601	52,3	12,7
2	,10	,876	71,7	28,3	,659	59,6	17,0
3	,12	,879	68,3	31,7	,669	60,7	17,5
4	,14	,883	63,3	36,7	,686	62,8	18,0
5	,16	,882	60,0	40,0	,677	61,9	18,5

TABLE R.11 RELIABILITY

Change in Variables when items with poor Phi Coefficients are Removed.

Date = 180674 n = 176 No. items = 100

Cycle	ø	KR 20	% Q.Left	% Q.Del.	T.D.	Corr. Mean	St. Dev.
1	-	,984	100	-	,579	39,4	14,6
2	,10	,987	86	14	,574	38,9	16,3
3	,12	,988	81	19	,561	37,3	16,8
4	,14	,989	76	24	,567	38,4	17,4
5	,16	,991	65	35	,578	40,3	18,4
6	,18	,992	61	39	,574	40,0	19,0
7	,20	,994	53	47	,563	38,6	20,2
8	,22	,996	48	52	,562	38,7	20,6
9	,24	,999	36	64	,583	40,7	23,4

TABLE R.12 RELIABILITY

Change in Variables when items with poor Phi Coefficients are Removed.

Cycle	Ø	KR 20	% Q.Left	% W.Del.	T.D.	Corr. Mean	St. Dev.
1	-	,999	100	-	,723	67,2	13,4
2	,10	1,001	80,5	19,5	,703	64,6	16,0
3	,12	1, 002	78,5	21,5	,698	63,9	16,3
4	,14	1,002	71,1	28,9	,699	63,9	17,4
5	,16	1,003	65,1	34,9	,706	65,0	18,3
6	,18	1,005	54,4	46,6	,697	64,3	19,5
7	,20	1,006	50,3	49,7	,718	67,1	20,0
8	,22	1,011	36,2	63,8	,701	65,7	22,3
9	,24	1,011	35,6	64,4	,703	65,9	22,2

TABLE R.13 RELIABILITY

Changes in Variables when Items with poor Phi Correlations are Removed.

Date = 051174 n = 179 No. of Questions = 134

Cycle	ø	KR 20	% Q.Left	% Q.Del.	T.D.	Corr.Mean	St.Dev.
1	-	,990	100	-	,640	48,6	13,1
2	,10	,997	60,4	39,6	,627	49,8	17,3
3	,12	,998	56,0	44,0	,632	50,6	17,8
4	,14	1,000	47,8	52,2	,606	47,8	18,8
5	,16	1,000	46,3	53,7	,610	48,1	19,0
6	,18	1,002	41,8	58,2	,625	50,1	19,8
7	,20	1,005	34,3	65,7	,603	47,9	20,9
8	,22	1,007	30,6	69,4	,605	48,4	21,3
9	,24	1,013	24,6	75,4	,592	47,8	22,0

TABLE R.14 RELIABILITY

Changes in Variables when Items with poor Phi Coefficients are Removed.

Date = 190475 n = 184 No. of Questions = 84

Cycle	ø	KR 20	% Q.Left	% Q.Del.	T.D.	Corr.Mean	St. Dev.
1	-	,871	100	-	,689	62,7	11,6
2	,10	,885	68,3	31,7	,657	58,9	15,7
3	,12	,885	65,9	34,1	,651	58,2	16,0

TABLE R.15 RELIABILITY

Changes in Variables when Items with poor Phi Coefficients are Removed.

Date = 180675 n = 182 No. of Questions = 100

Cycle	g	KR 20	% Q.Left	% Q.Del.	T.D.	Corr.Mean	St. Dev.
1	_	,994	100	-	,630	54,4	12,9
2	,10	,999	78	22	,655	56,9	14,6
3	,12	1,001	66	34	,642	56,0	15,6
4	,14	1,001	65	35	,645	56,4	15,7
5	,16	1,003	58	42	,659	58,3	16,3
6	,18	1,004	55	45	,652	57,4	16,7
7	,20	1,006	49	51	,662	58,7	17,5
8	,22	1,009	41	59	,660	58,2	18,5
9	,24	1,012	35	65	,659	57,9	19,6

CHANGES OF VARIABLES IN SUCCESSIVE CYCLES

KR 20 RELIABILITY COEFFICIENT

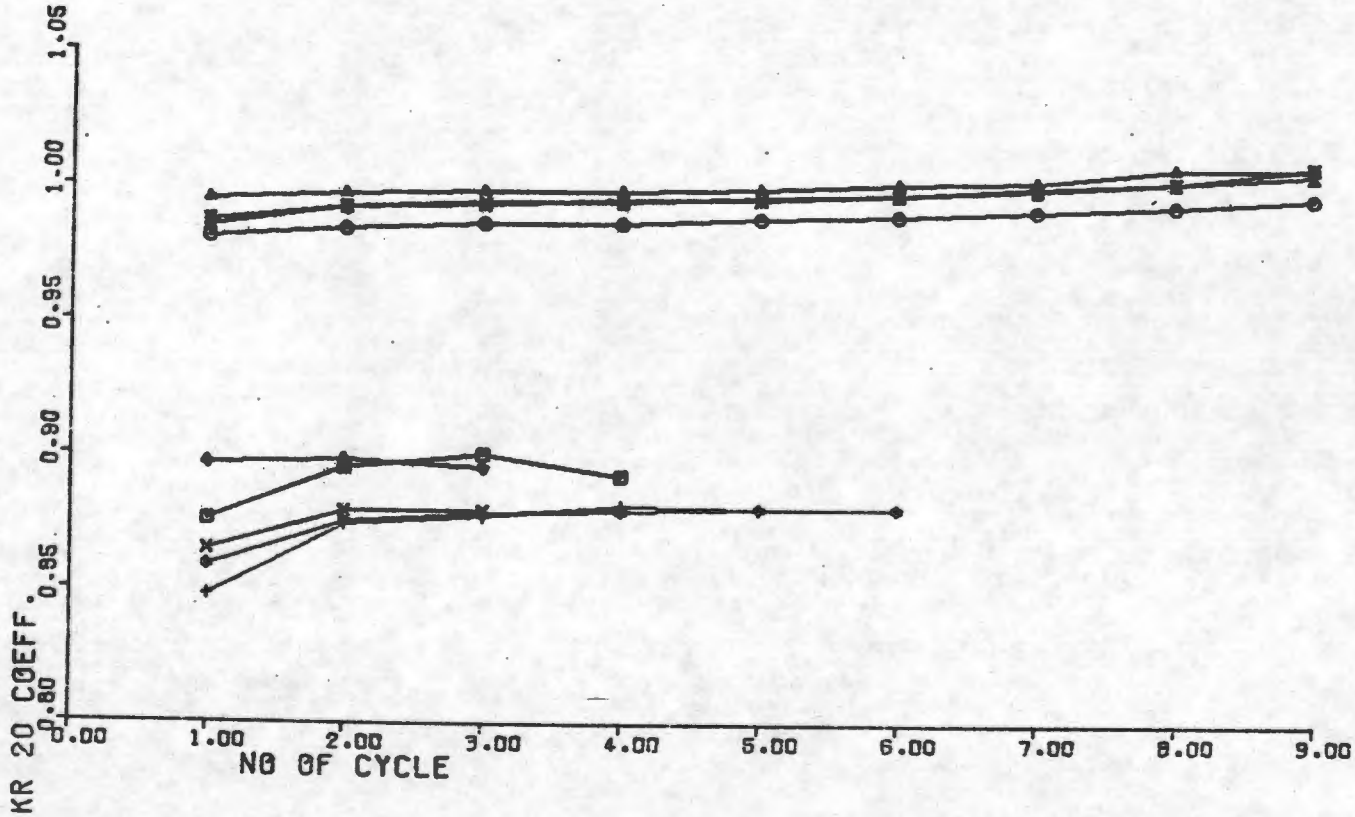


Fig. R13

CHANGES OF VARIABLES IN SUCCESSIVE CYCLES

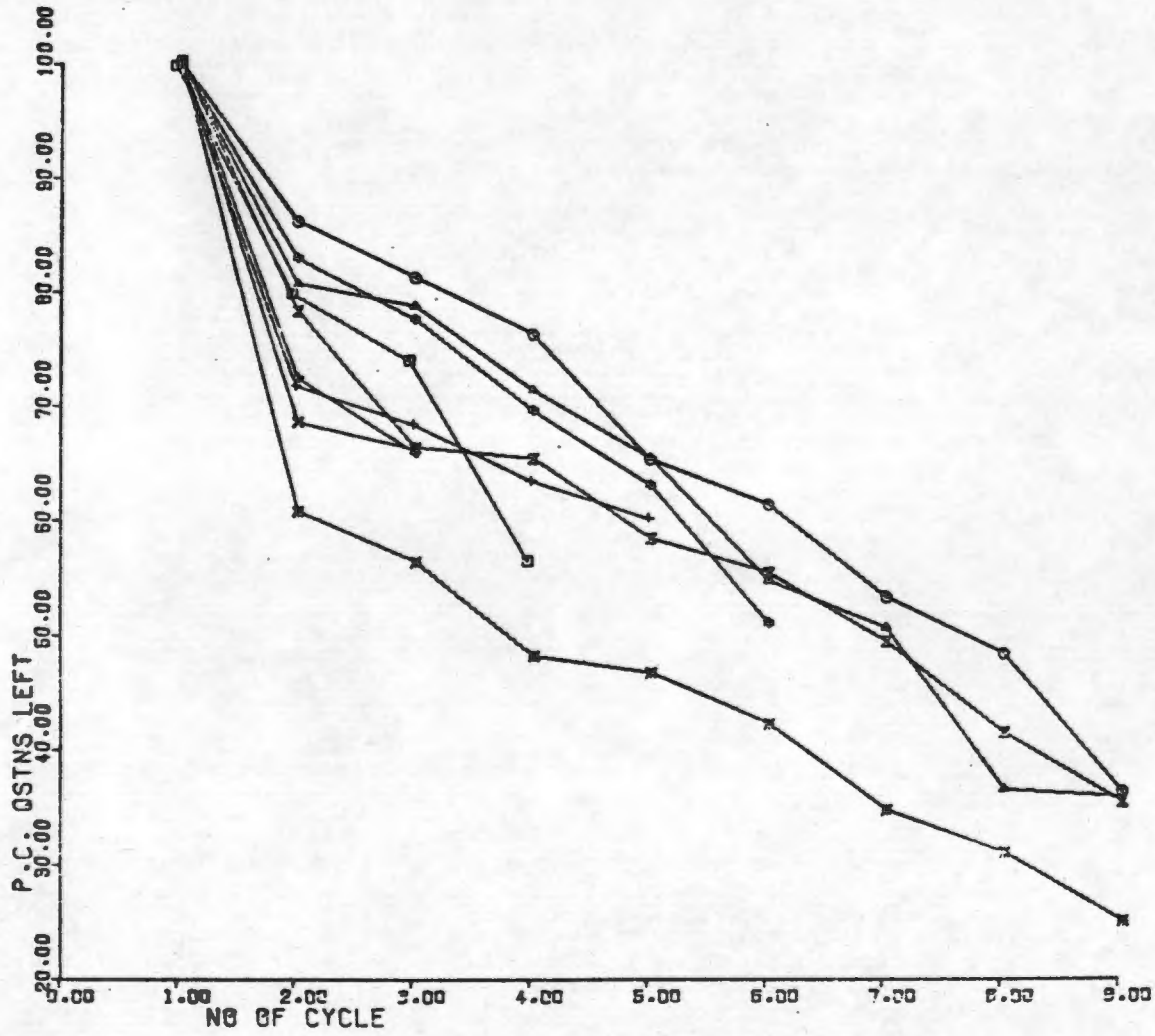
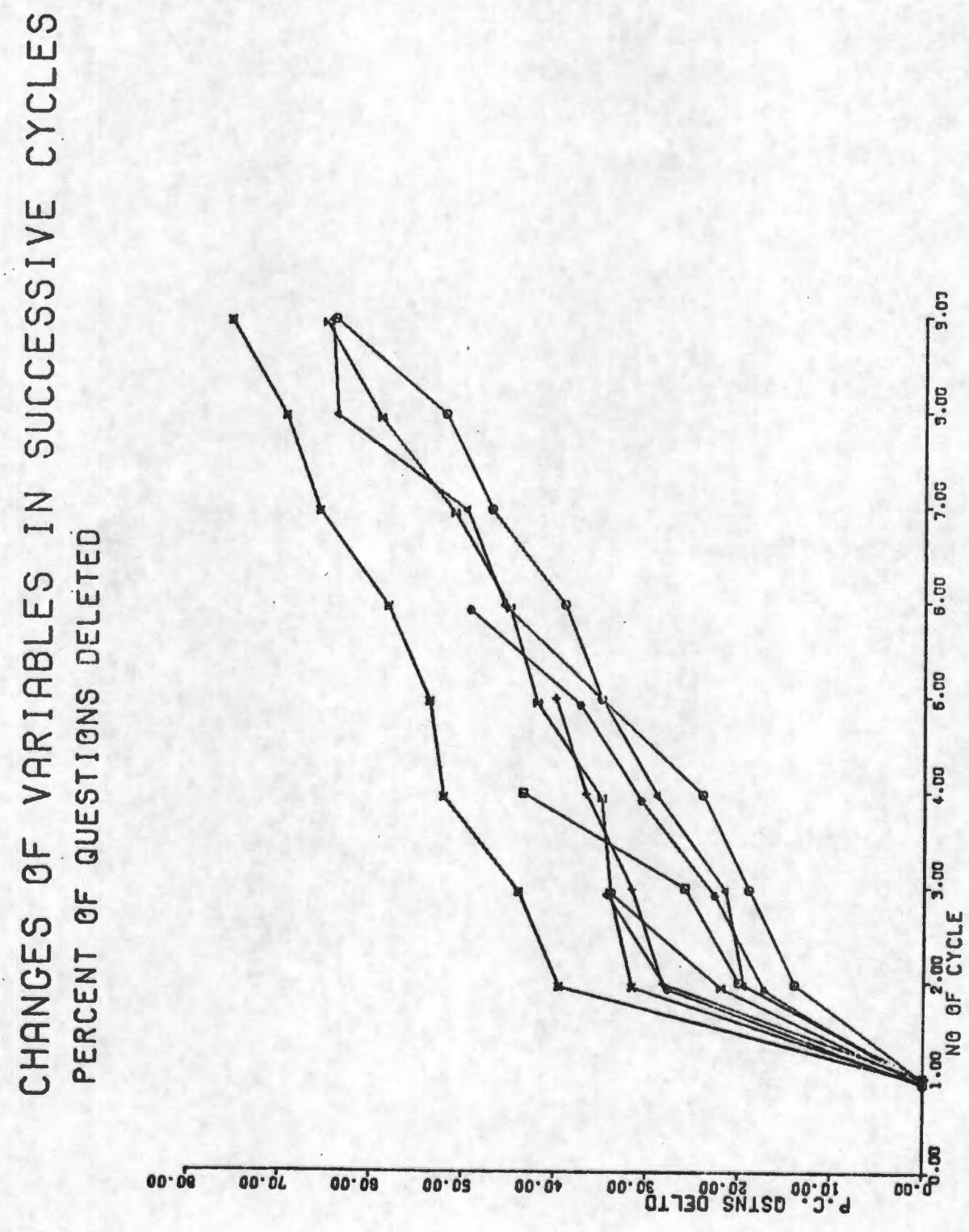


Fig. R14

Fig. R15


CHANGES OF VARIABLES IN SUCCESSIVE CYCLES
PERCENT OF QUESTIONS DELETED
P.C. QSTNS DELTD
0.00
10.00
20.00
30.00
40.00
50.00
60.00
70.00
80.00
NO OF CYCLE
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00

CHANGES OF VARIABLES IN SUCCESSIVE CYCLES

CORRECTED MEAN SCORE

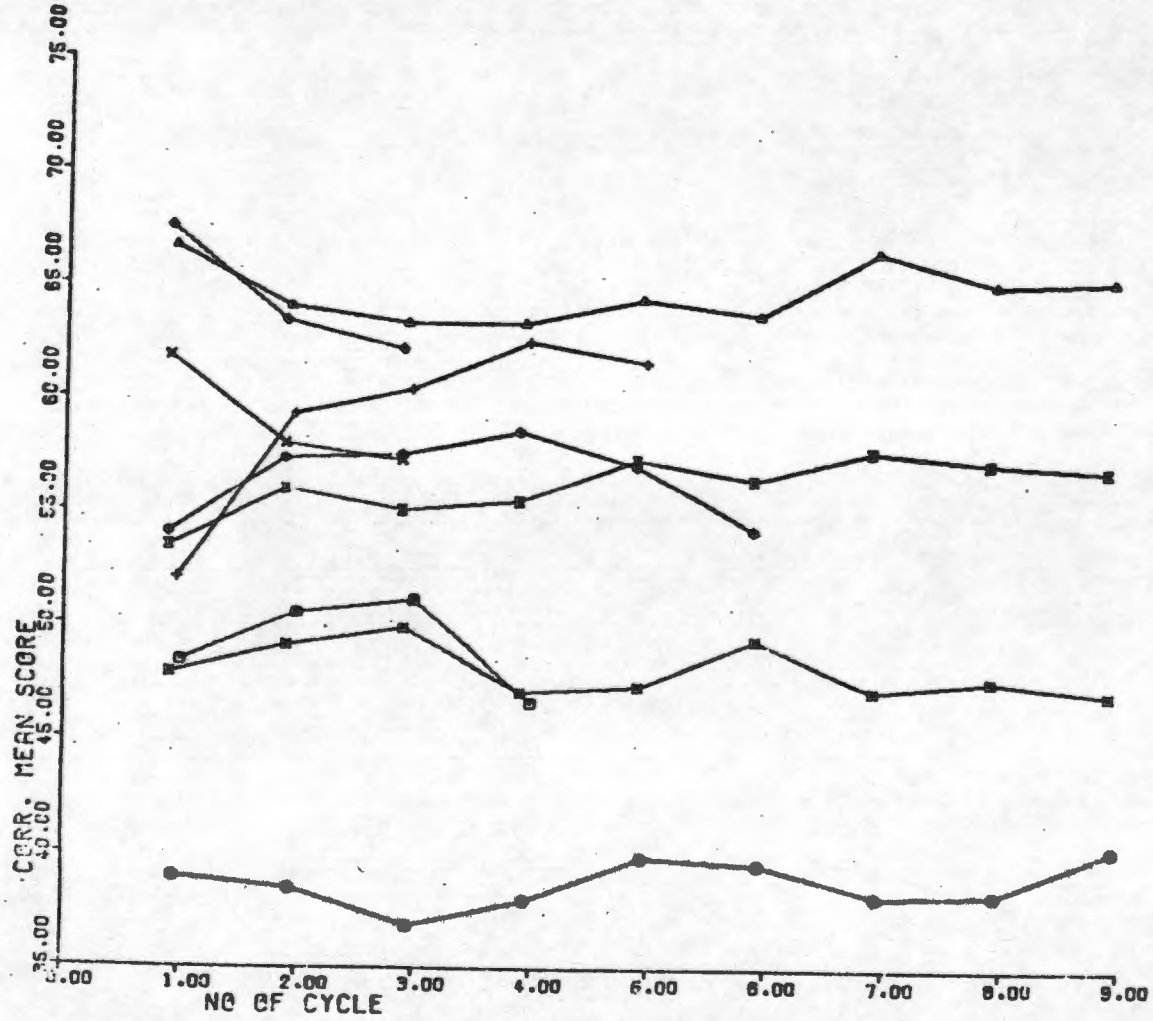


Fig. R16

Fig. R17

CHANGES OF VARIABLES IN SUCCESSIVE CYCLES
CORRECTED STANDARD DEVIATION
CORR. MEAN SCORE
10.00
12.50
15.00
17.50
20.00
22.50
25.00
27.50
30.00
NO OF CYCLE
0.00
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00

TABLE R.16 RELIABILITY

Variables Examined in results of Cycling Marking Program

Variable	Mean	St. Dev.	Max.	Min.
1. No. of students	177,12	5,79	184	163
2. Cycle	4,21	2,51	9	1
3. KR20 Rel. Coeffic.	0,957	0,057	1,013	,850
4. No. of questions in run	66,01	24,38	149	33
5. No. of questions deleted in Run	37,63	28,15	101	0
6. Test Difficulty	0,639	0,049	0,738	0,559
7. Mean Score	54,35	8,87	68,5	37,3
8. Mean St. Dev.	17,22	2,73	23,4	11,6
9. Chai Value for Normality of Dist.	12,74	7,31	40,83	3,96
10. Percentage of Qstns. remaining in Run.	65,62	20,75	100,0	24,6
11. Percentage of Qstns. deleted in Run.	34,38	20,75	75,4	0,0

n = 56

TABLE R.17 RELIABILITY

Correlation Coefficients of Variables examined in results of Cycling Marking Program

1. No. Sts.	1,00										
2. Cycle	,155	1,00									
3. KR20 Coeff.	,383*	,516*	1,00								
4. No. Q.	,019	-,587	,210	1,00							
5. No. Q. deleted	,229	,852*	,608*	-,444*	1,00						
6. Test diff.	-,057	-,066	-,068	,212	,004	1,00					
7. Corr. Mean Score	-,048	-,089	-,230	,097	-,026	-,957*	1,00				
8. Corr. Mean St. Dev.	,147	,898*	,480*	-,560*	,883*	-,164	-,198	1,00			
9. Chai dist.	,107	,404*	,202	-,209	,328	,202	,182	,420*	1,00		
10. % Q Sts. left	-,191	-,908*	-,459	,673*	-,930*	,056	,058	-,920*	-,303	1,00	
11. % Q Sts. del.	,191	,908*	,459*	-,673*	,930*	-,056	-,058	,920*	,303	-1,00	1,00

* = Highly significant. p = <,01 n = 56

TABLE R.18 RELIABILITY

Percentage of Question in Middle ID Range (40-60%)

Date	Type	No.Q	KR20 Coeff.	% Item Midd. ID Range
301069	FT	60	,82	28,3
220470	CT	40	,76	25,0
170670	CT	60	,83	33,3 *
051070	SE	34	,65	22,2
051170	FT	75	,86	26,7
290371	SE	30	,62	33,3 *
200471	CT	45	,78	35,5 *
140671	SE	30	,66	26,7
280671	CT	60	,79	21,7
021171	FT	75	,92	17,3
030372	SE	30	,62	40,0 *
170472	SE	30	,66	26,7
260472	CT	50	,84	30,0 *
120672	SE	30	,70	43,3 *
230672	CT	75	,84	28,0
271072	FT	90	,88	16,7
060373	SE	24	,70	20,8
100473	SE	30	,73	20,0
240473	CT	45	,85	13,3
110673	SE	30	,72	16,7
270673	CT	85	,86	32,9 *
071173	FT	90	,92	20,0
050374	SE	30	,61	16,7
090474	SE	30	,74	20,0
230474	CT	60	,76	15,0
180674	CT	100	,72	39,0 *
041174	NFT	149	,99	13,4
051174	GFT	179	,99	23,9

* = > 30%

TABLE R.19 RELIABILITY

Mean KR20 Values of Tests with more than 30% items of Middle Range I.D. compared to others.

		More than 30% Middle I.D. Range	Other Tests
Max. KR	=	,86	,99
Min. KR	=	,62	,61
Mean KR	=	,748	,793
St. Dev.	=	,094	,113
n	=	8	20

TABLE QT.1 QUESTION TYPES

Mean Item Difficulty.

Type	n	Mean I.D.	Stan. Deviation
All	577	64,61	20,17
1/5	130	71,32	16,40
W/5	107	57,54	20,81
C/4	110	62,75	20,44
CSL	82	56,59	22,15
RLTD	24	76,08	8,51
T/F	124	68,44	19,10

TABLE QT.2 QUESTION TYPES.

Rank Order Item Difficulty

		Mean I.D.
1	RLTD	76,08
2	1/5	71,32
3	T/F	68,44
4	C/4	62,75
5	W/5	57,54
6	CSL	56,59

TABLE QT.3 QUESTION TYPES.

T-values for the Differences of means of Item Difficulty and Probability Confidence Levels.

	RLTD	1/5	T/F	C/4	W/5	CSL
RLTD	-	2,11 (32)	3,13 (33)	5,11 (34)	2,66 (34)	6,50 (38)
1/5	,05	-	1,29 (251)	3,54 (231)	5,57 (226)	5,19 (172)
T/F	,01	N	-	2,19 (229)	2,64 (224)	3,97 (174)
C/4	,001	,001	,05	-	1,86 (215)	1,97 (175)
W/5	,05	,001	,01	N	-	0,30 (174)
CSL	,001	,001	,001	,05	N	-

Upper right figures = T-value and degrees of freedom in parenthesis.

Lower left figures = maximal p value for confidence of validity of difference.

N = Not significant.

TABLE QT.4 QUESTION TYPES

Mean Discrimination Index.

Q. Type	n	Phi. Coefficient	Standard Deviation
All	577	,1896	,1112
1/5	130	,2241	,1002
W/5	107	,1877	,1094
C/4	110	,2079	,1035
Causal	82	,1798	,1175
Related	24	,2588	,0940
T/F	124	,1325	,1051

TABLE QT.5 QUESTION TYPES

Value of Phi and Reliability for 180 Students.

Probability	ø
99,9%	,25
99,0%	,19
98,0%	,17
95,0%	,15

Source: U.C.T. M.C.Q. Analysis Program.

TABLE QT.6 QUESTION TYPES.

Rank Order of Question Formats by Discrimination Index.

	Type	phi	Confidence	Mean I.D.
1	RELATED	,2558	99,9%	76,08
2	1/5	,2241	99%	71,32
3	C/4	,2079	99%	62,75
4	W/5	,1877	98%	57,54
5	CAUSAL	,1798	98%	56,69
6	T/F	,1325	< 95%	68,44
	All	,1896	98%	64,61

TABLE QT.7 QUESTION TYPES.

T value for Difference of Means of Phi Coefficients and Probability Confidence Levels.

	RLTD	1/5	C/4	W/5	CSL	T/F
RLTD	-	1,64 (32)	2,36 (34)	3,25 (34)	3,41 (38)	5,90 (33)
1/5	N	-	1,23 (231)	2,65 (226)	2,83 (172)	7,10 (251)
C/4	,05	N	-	1,40 (215)	1,72 (175)	5,52 (229)
W/5	,01	,01	N	-	,47 (174)	3,89 (224)
CSL	,01	,01	N	N	-	2,95 (174)
T/F	,001	,001	,001	,001	,01	-

Upper Rt. figures = T value + (Degrees of Freedom).

Lower Lt. figures = maximal p value for confidence of validity of difference.

N = Not significant.

TABLE QT.8 QUESTION TYPES

Mean Item Difficulties of Question Types

Q. Type	n	Mean Item Diff.	Mean Stand. Dev.
All	577	64,61	20,17
1/5	130	71,32	16,40
W/5	107	57,54	20,81
C/4	110	62,75	20,44
Causal	82	56,59	22,15
Related	24	76,08	8,51
R/F	124	68,44	19,10

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE ALL QUESTIONS

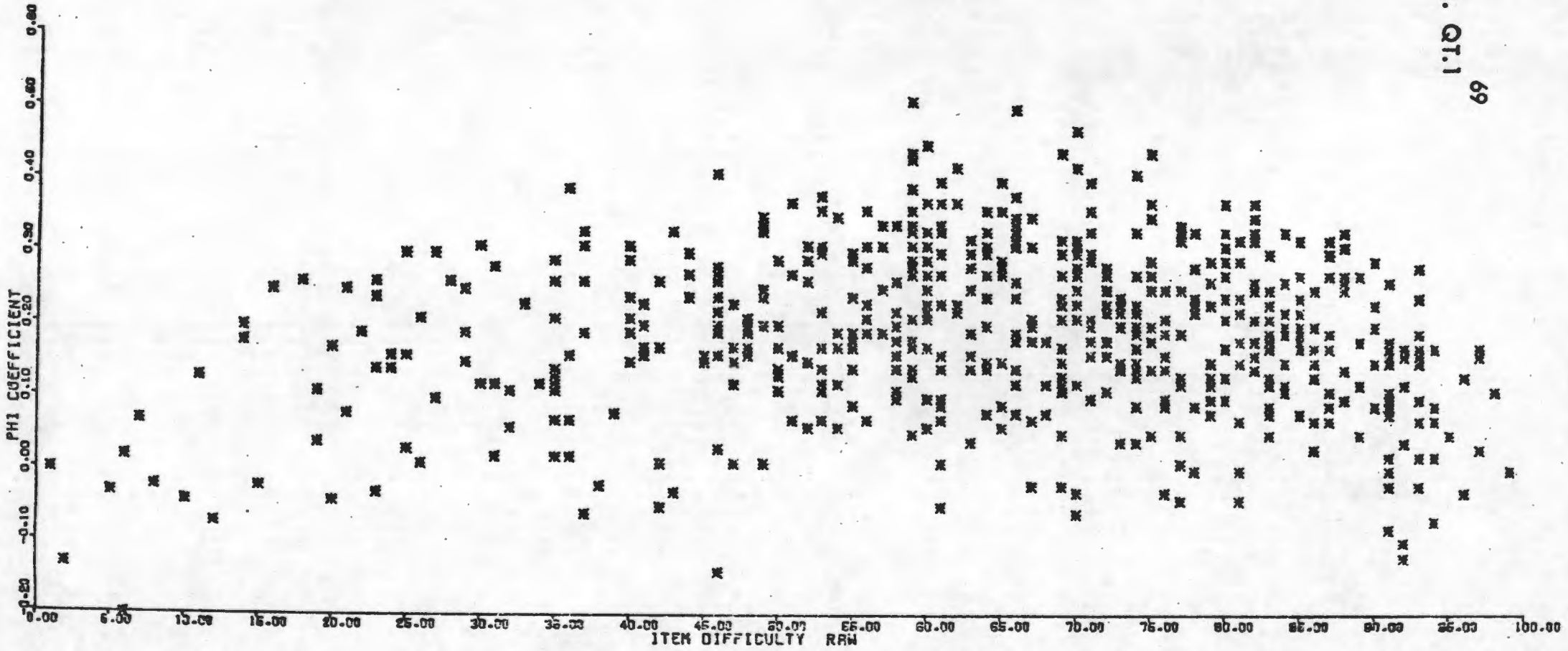


Fig. QT.1

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE 1/5

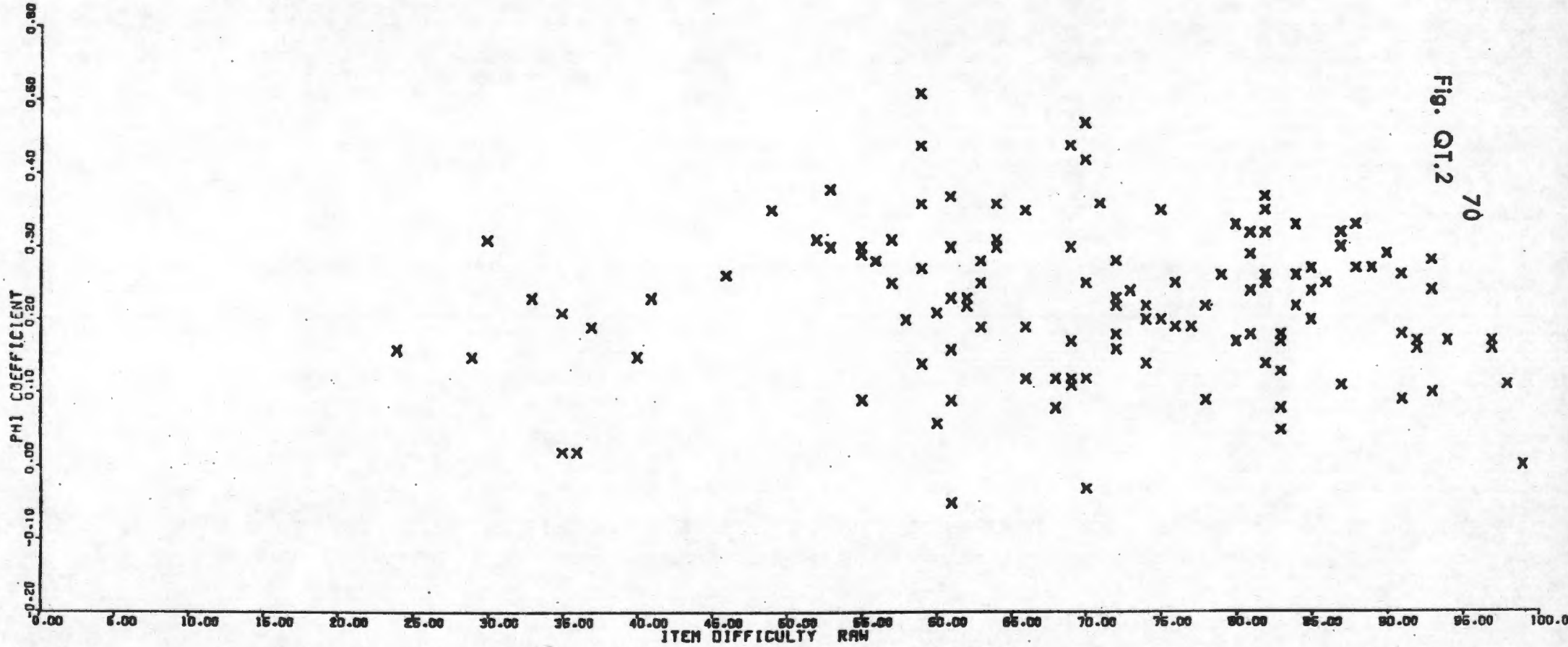


Fig. QT.2

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE W/5

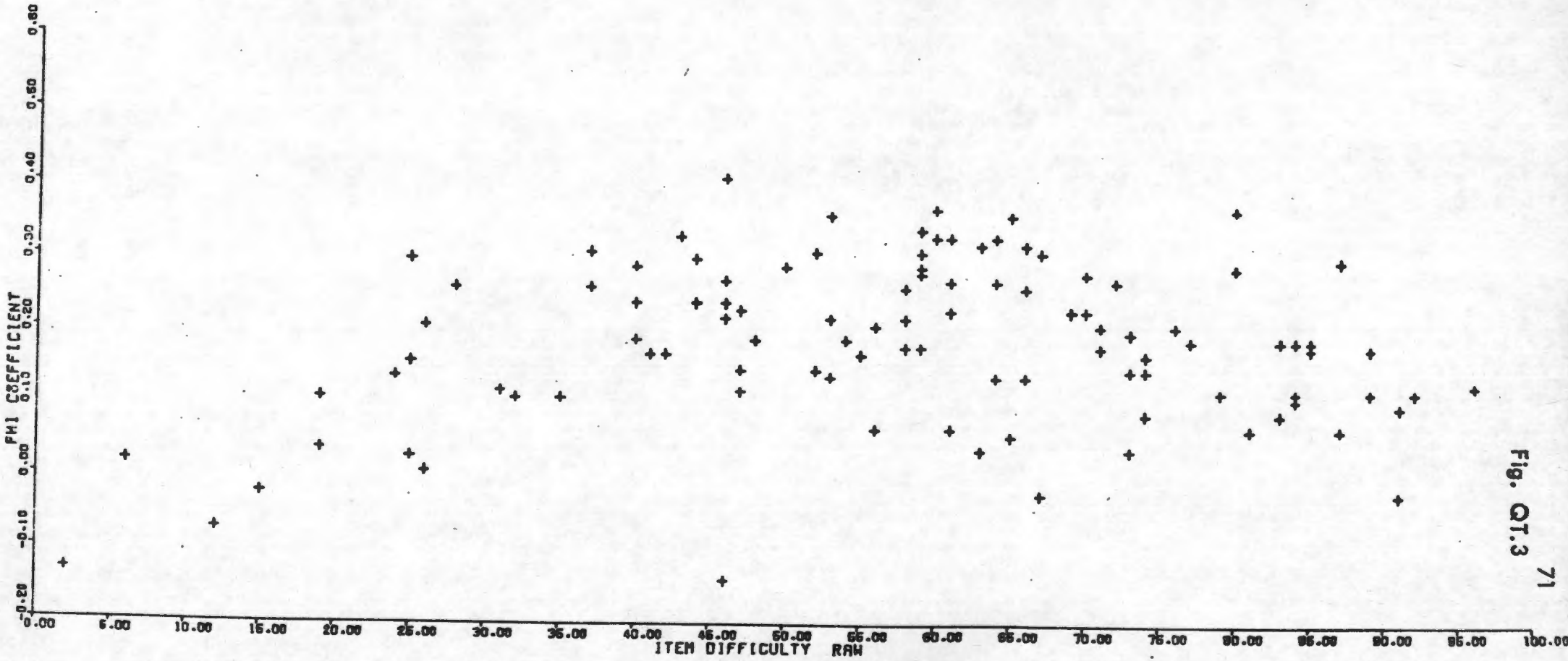


Fig. QT.3

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE C/4

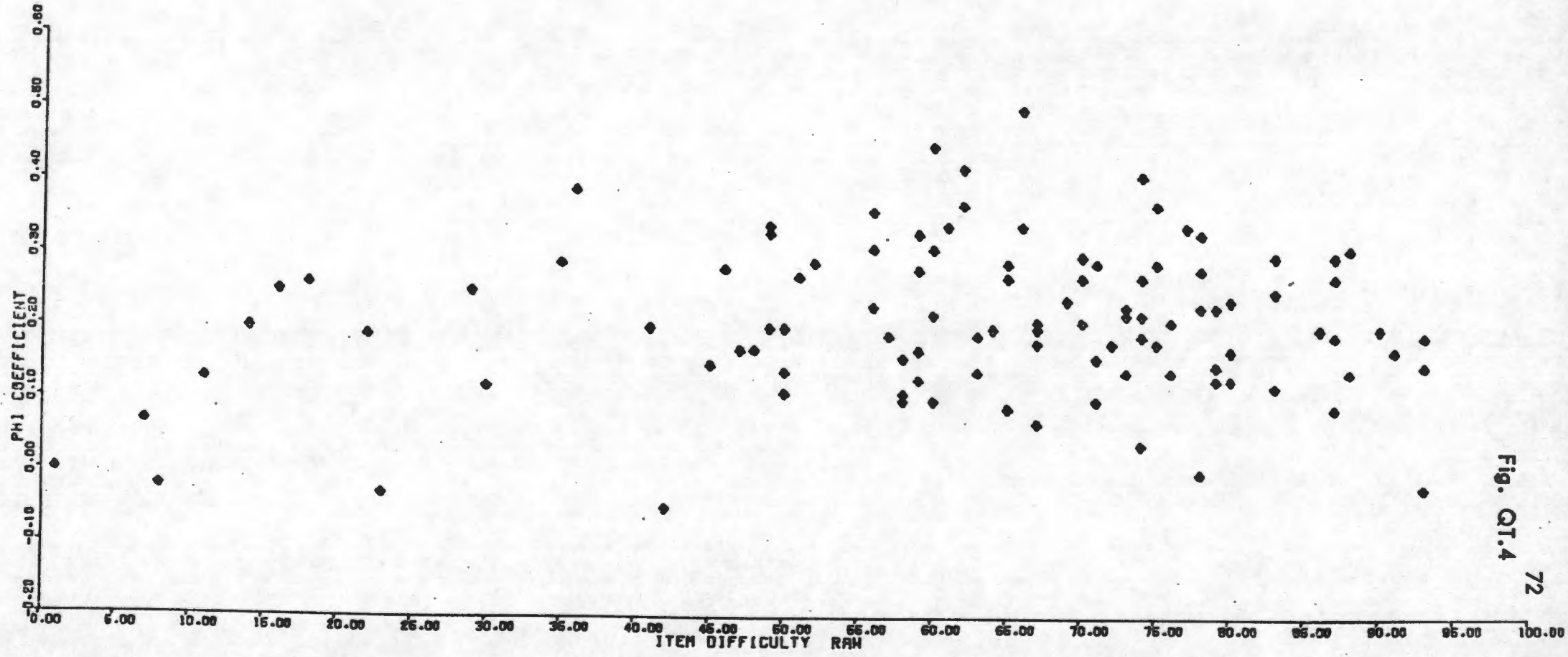


Fig. QT.4

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE CAUSAL

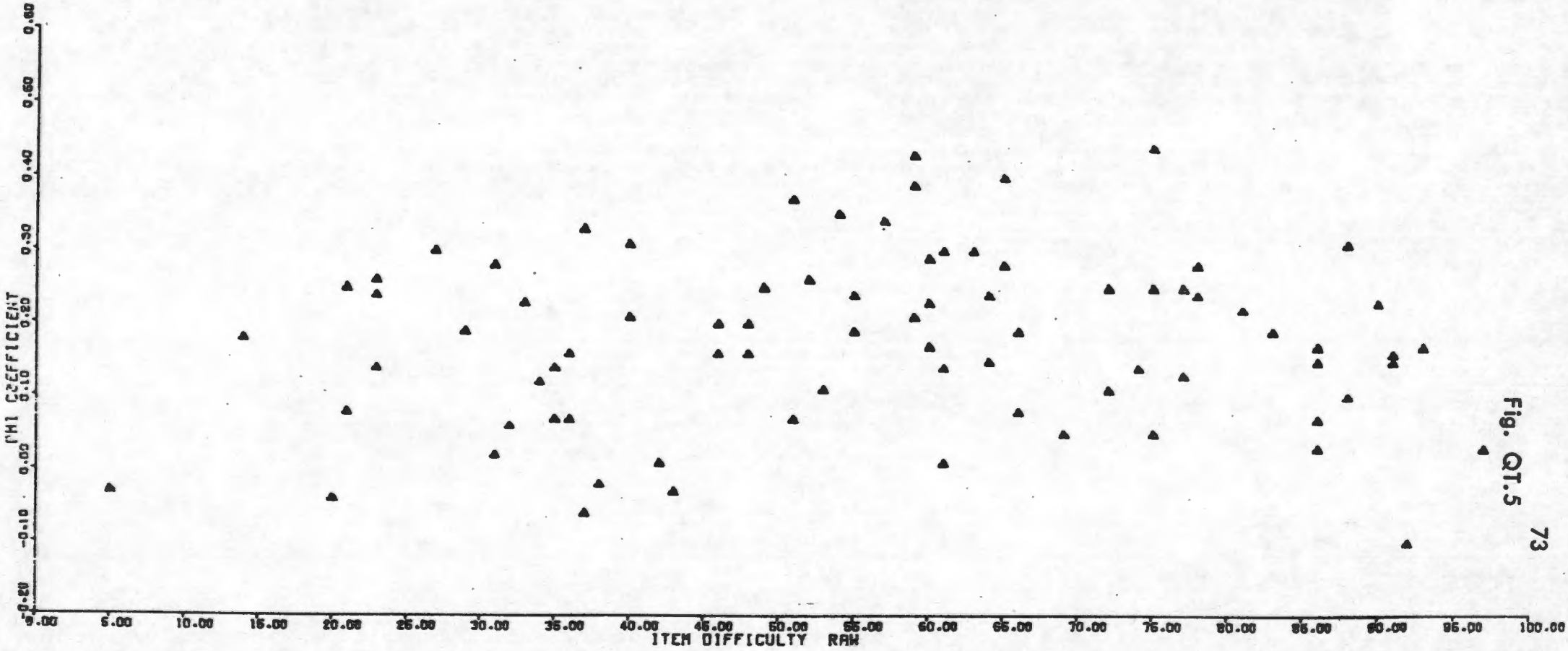


Fig. QT.5

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE RELATED/EXCLUSI0N

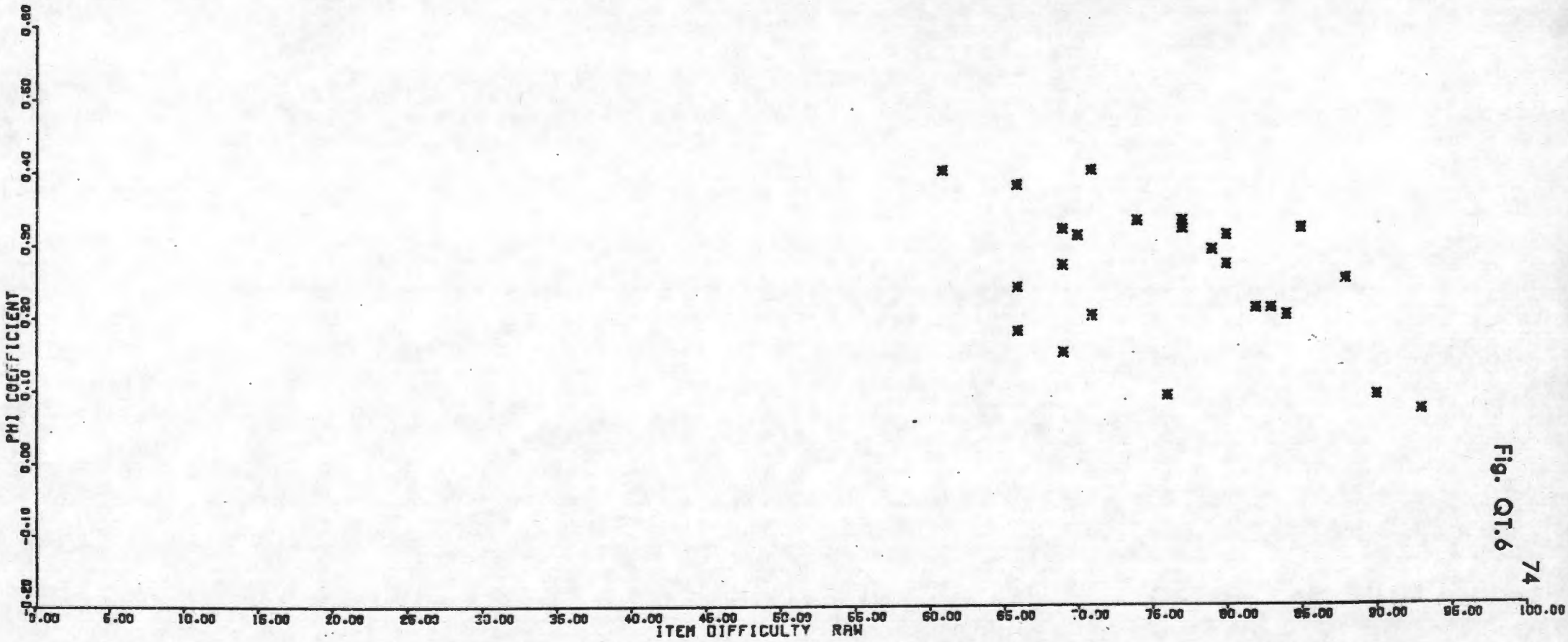


Fig. QT.6

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE TRUE/FALSE

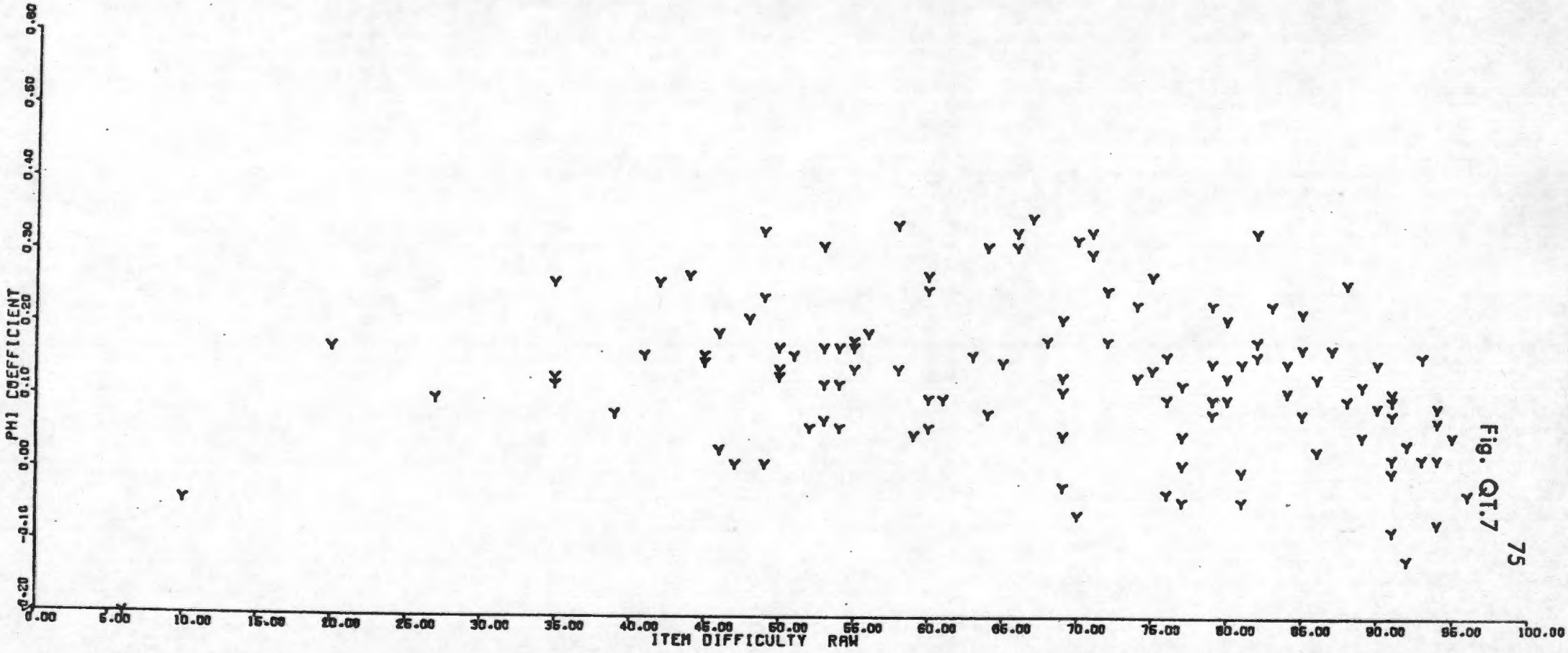


Fig. QT.7

TABLE QT. 9 QUESTION TYPES

Correlation Coefficient between Item Difficulty and PHI

Q. Type	n	r =
All	577	,0475
1/5	130	-,0557
W/5	107	,1536
C/4	110	,1392
CSL	82	,0872
RELTD	24	-,4717 *
T/F	124	-,1257

* p = $<.02$

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE ALL QUESTI0NS

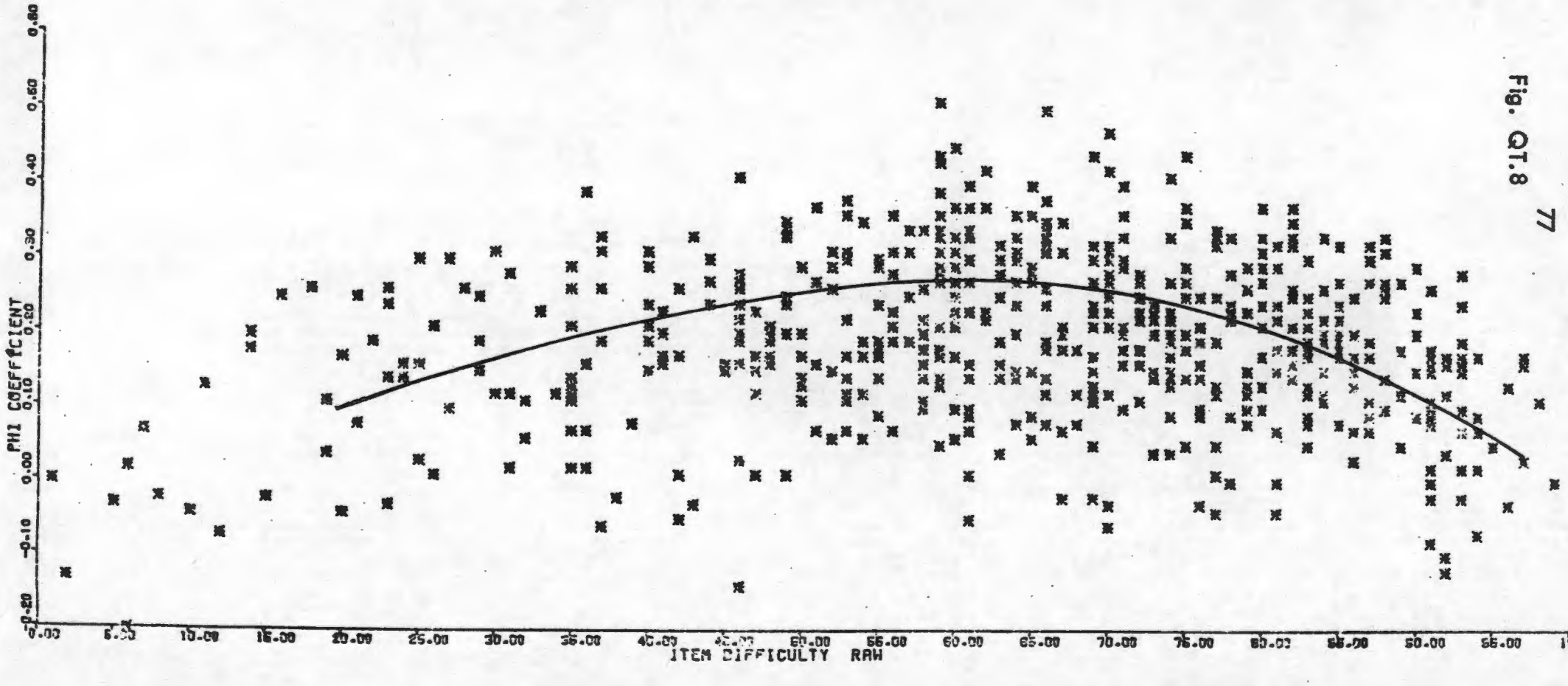


Fig. QT.8

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE ALL QUESTI0NS

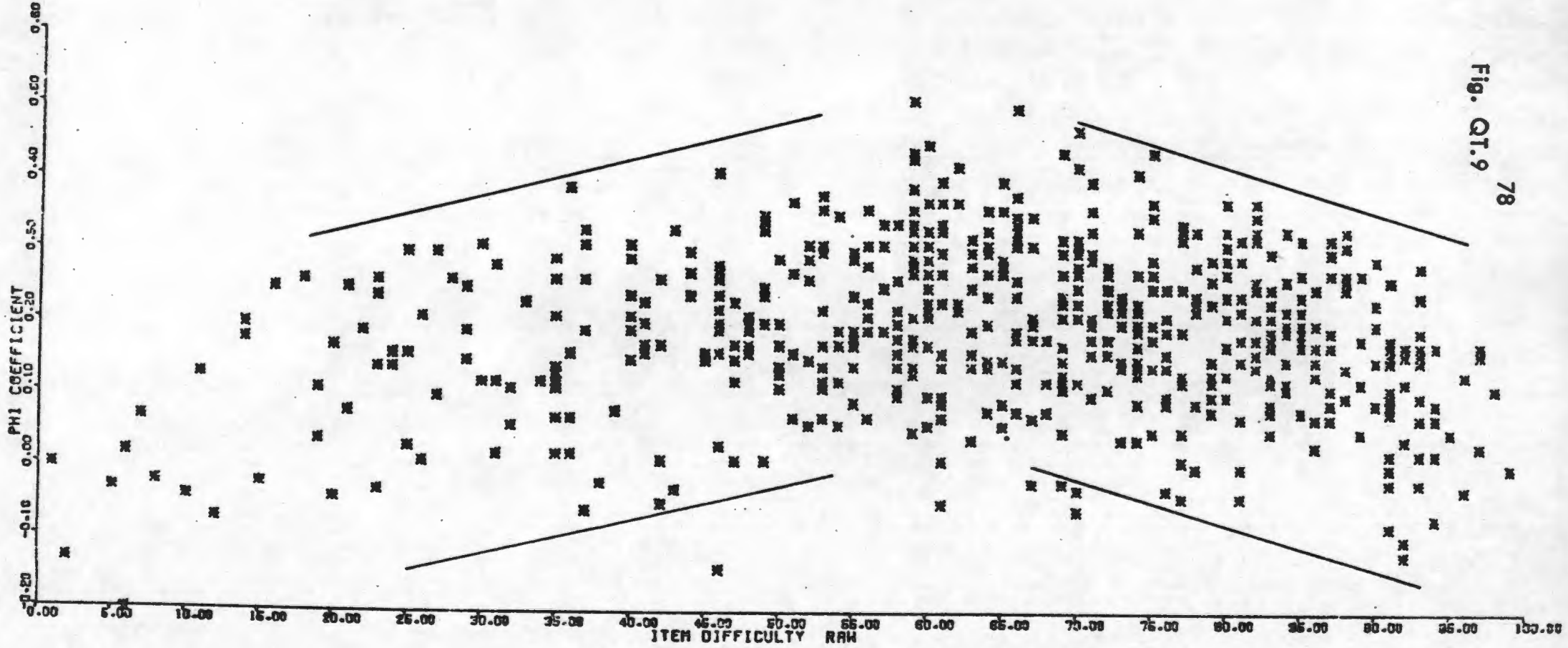


Fig. QT.9

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE ALL QUESTI0NS

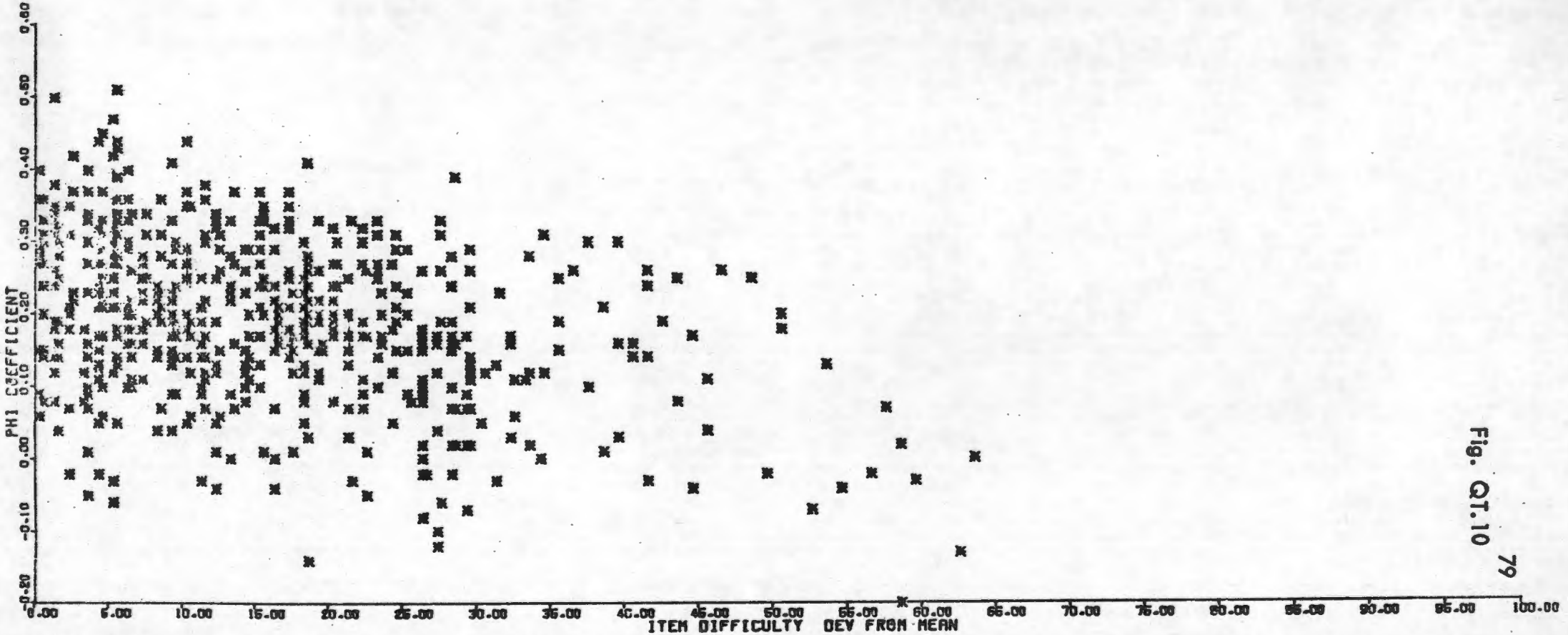


Fig. QT.10

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE ALL QUESTI0NS

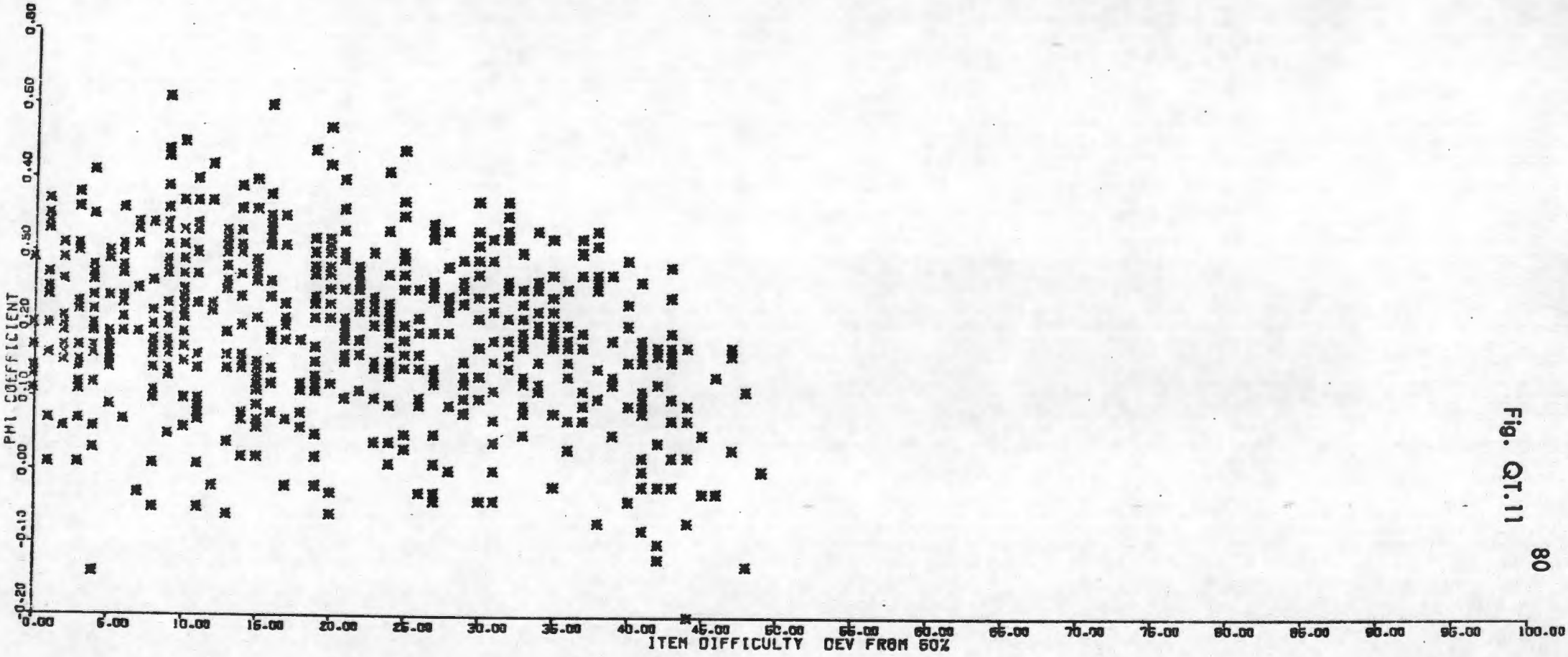


Fig. QT.11

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE 1/5

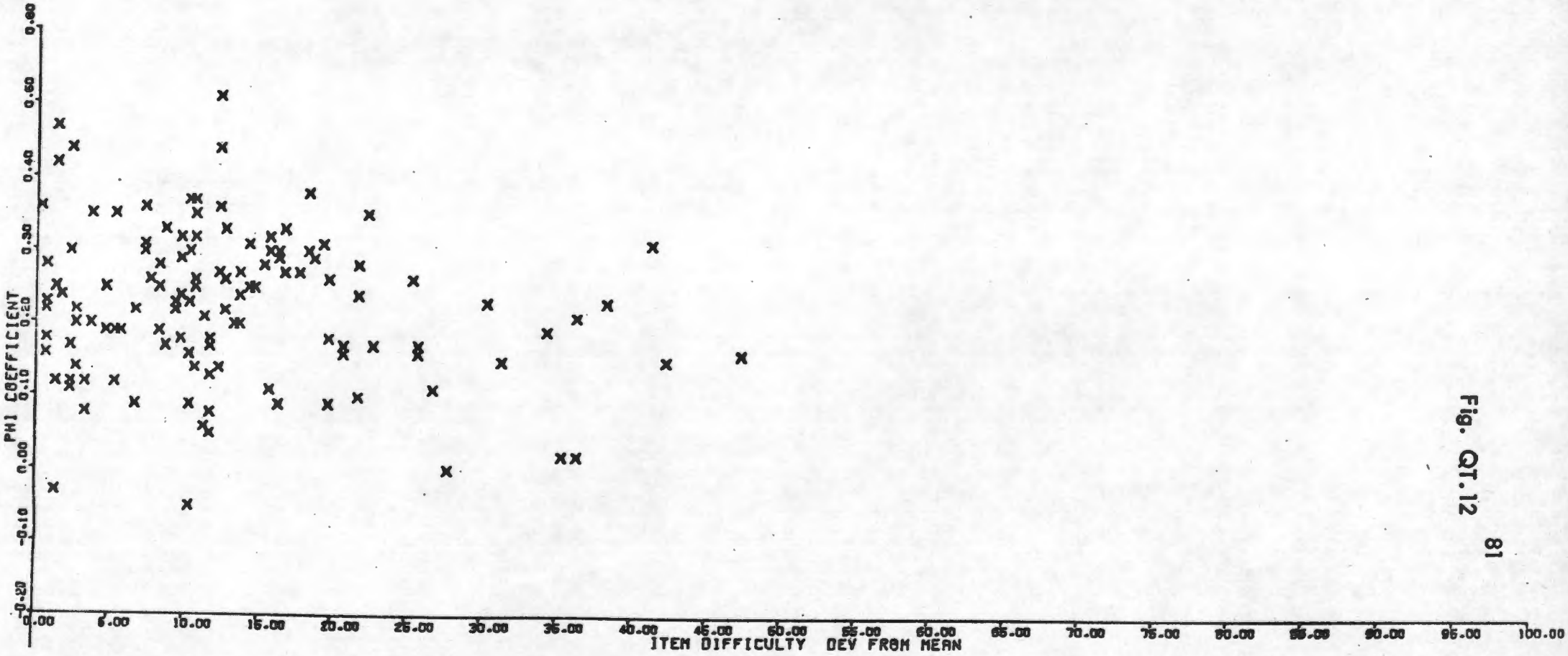


Fig. QT.12

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE 1/5

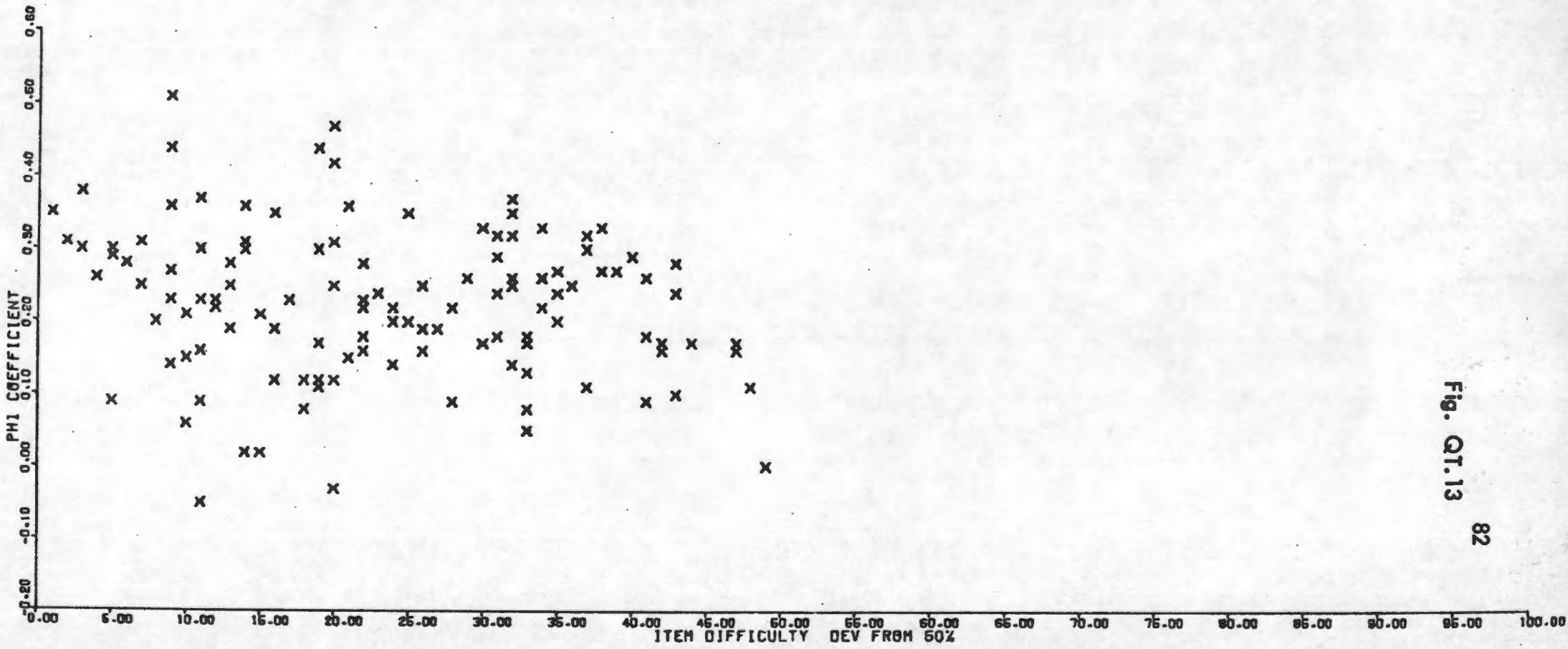


Fig. QT.13

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE W/5

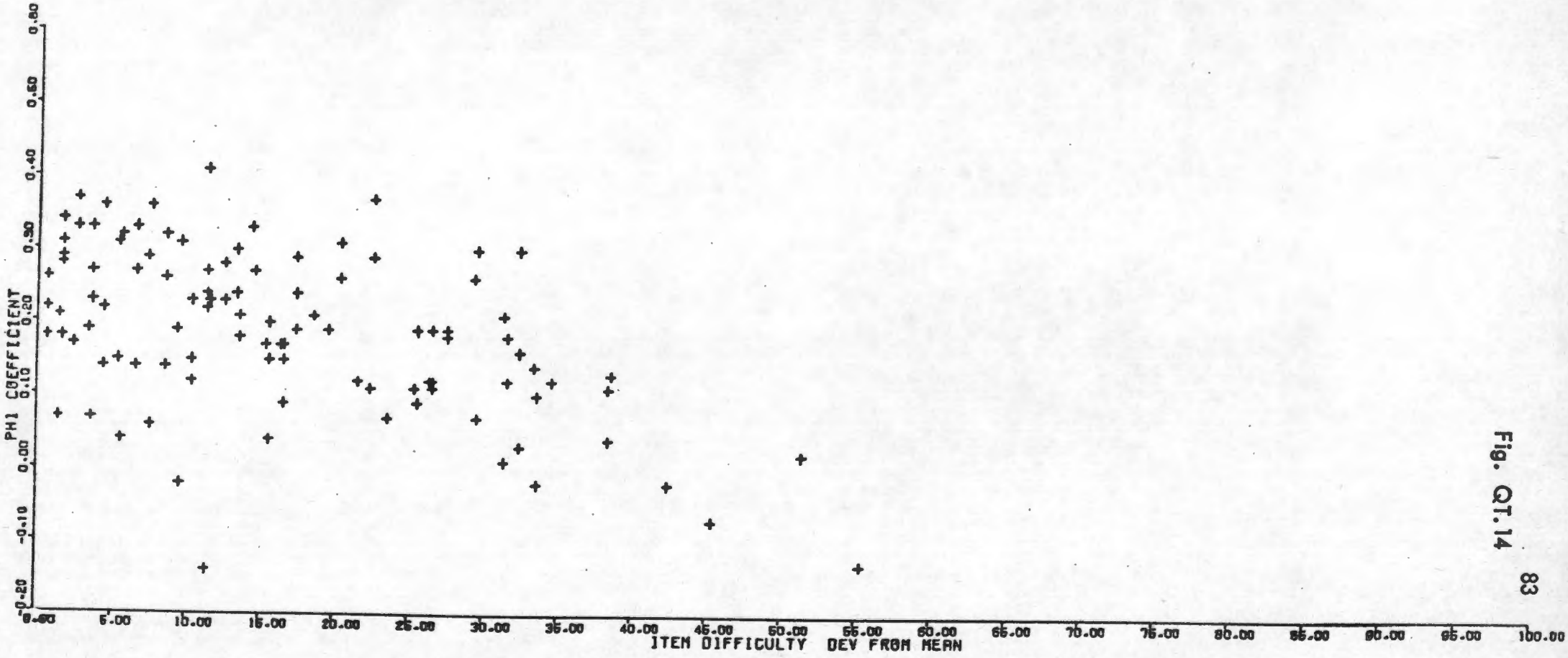


Fig. QT.14

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE W/5

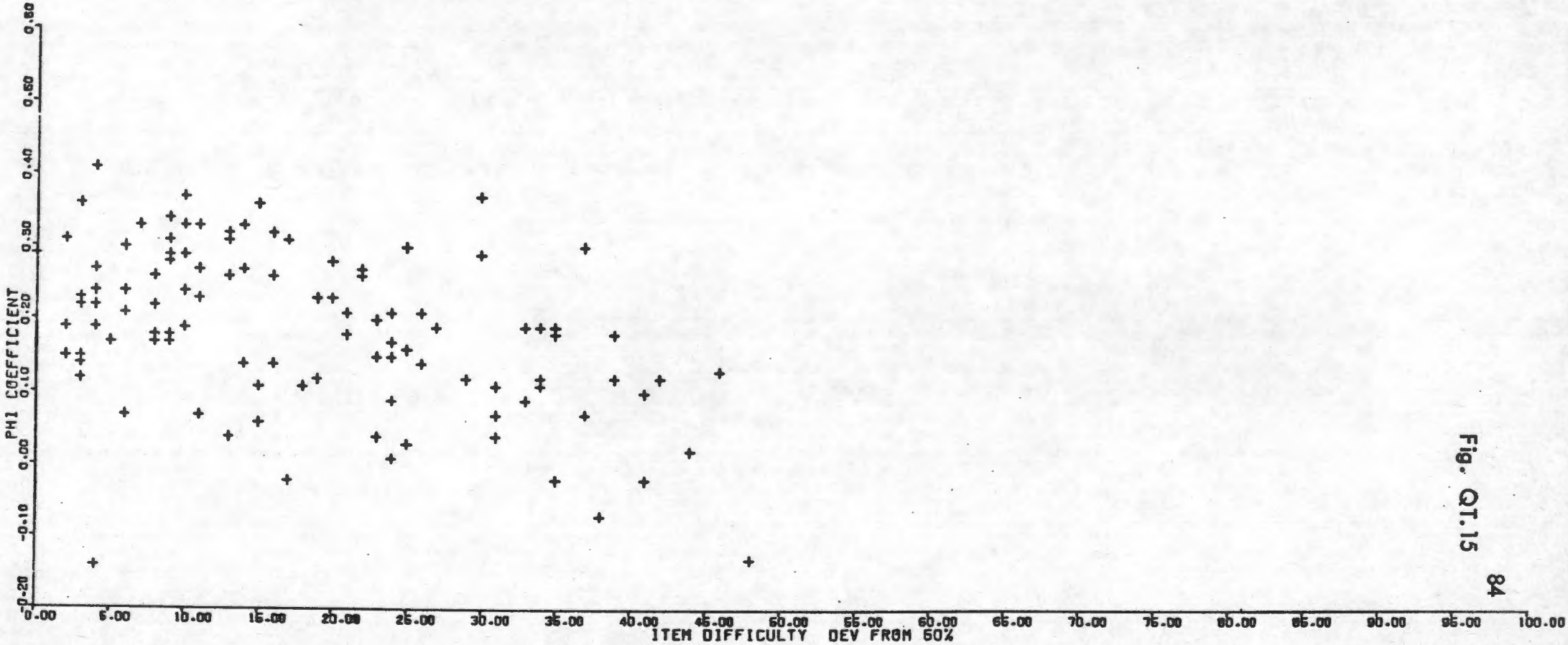


Fig. QT.15

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE C/4

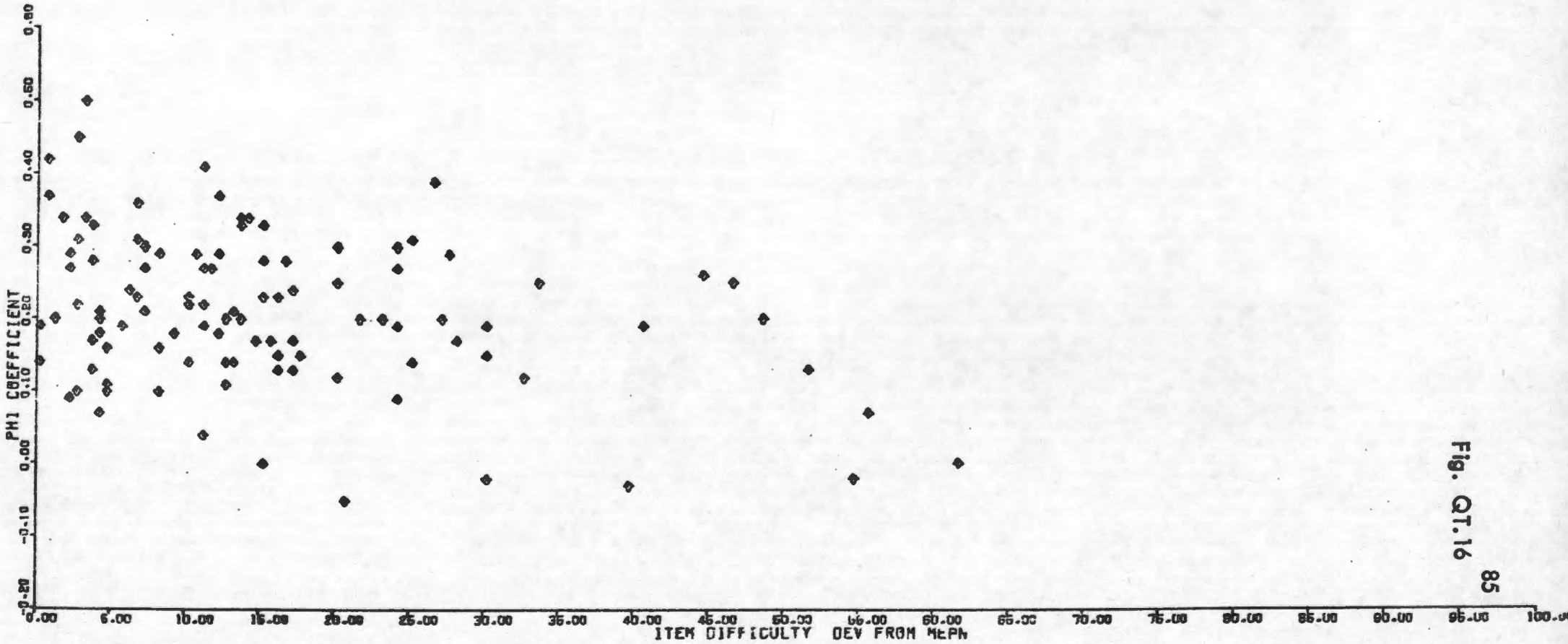


Fig. QT.16

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE C/4

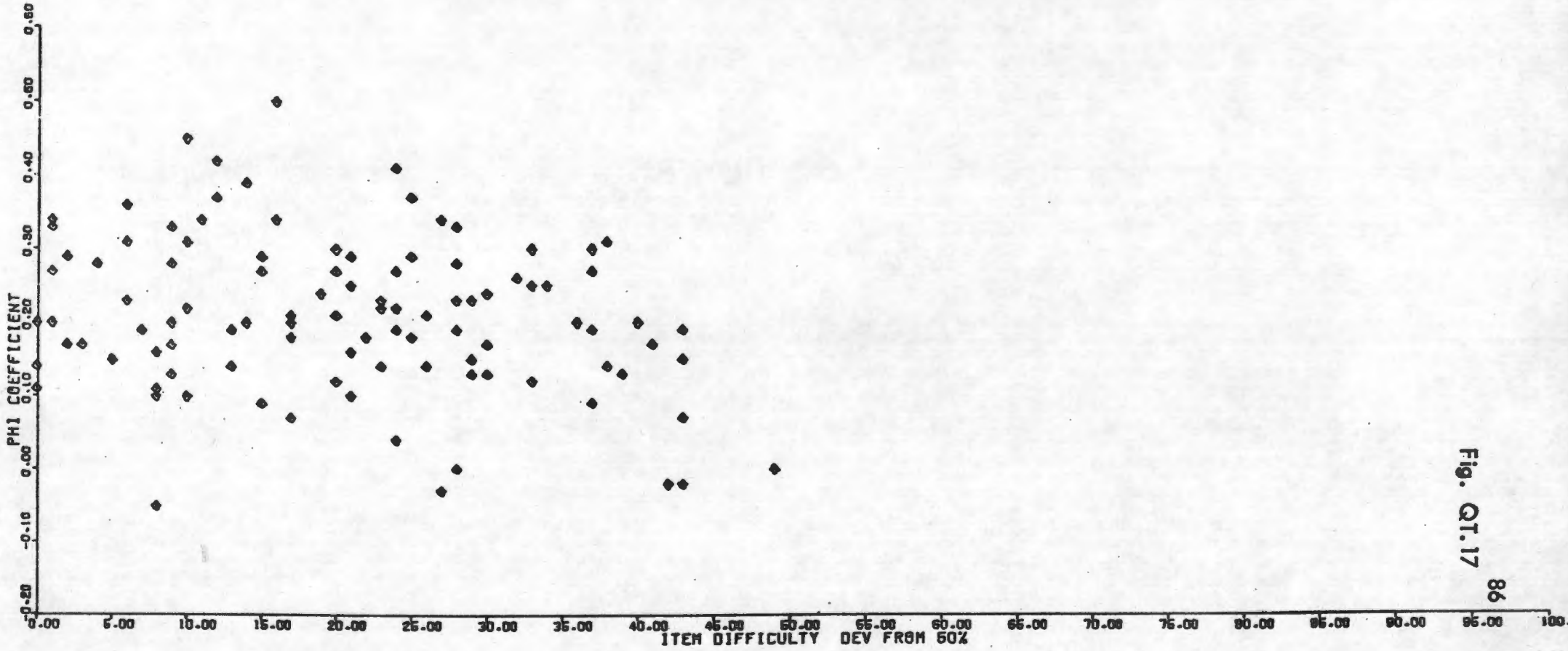


Fig. QT.17

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE CAUSAL

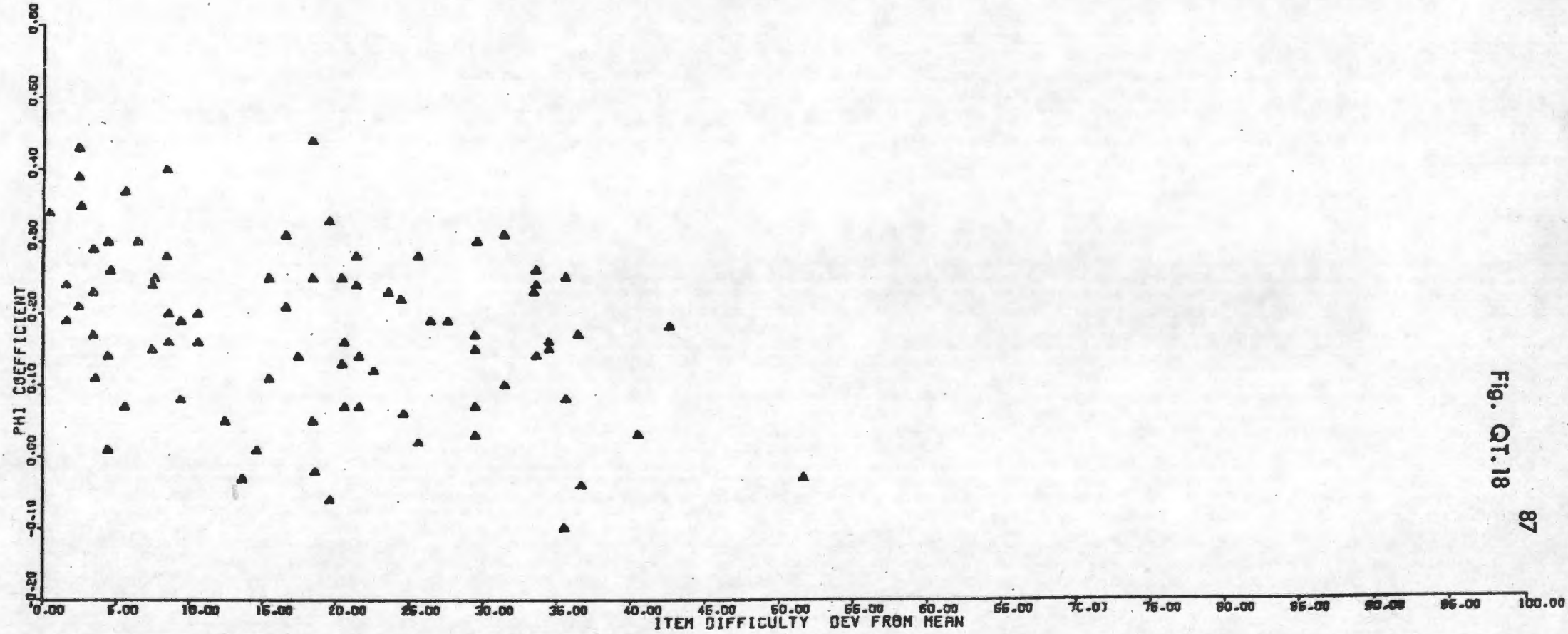


Fig. QT.18

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE CAUSAL

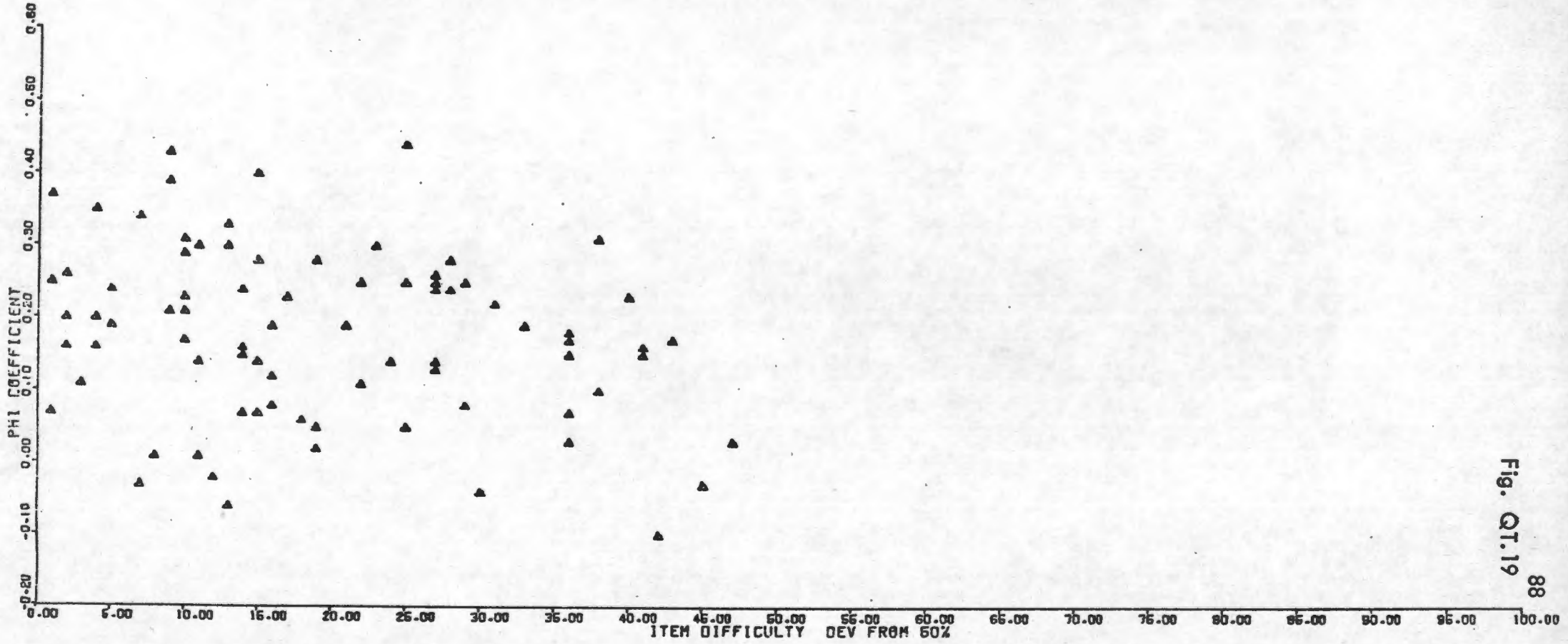


Fig. QT.19

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE RELATED/EXCLUSION

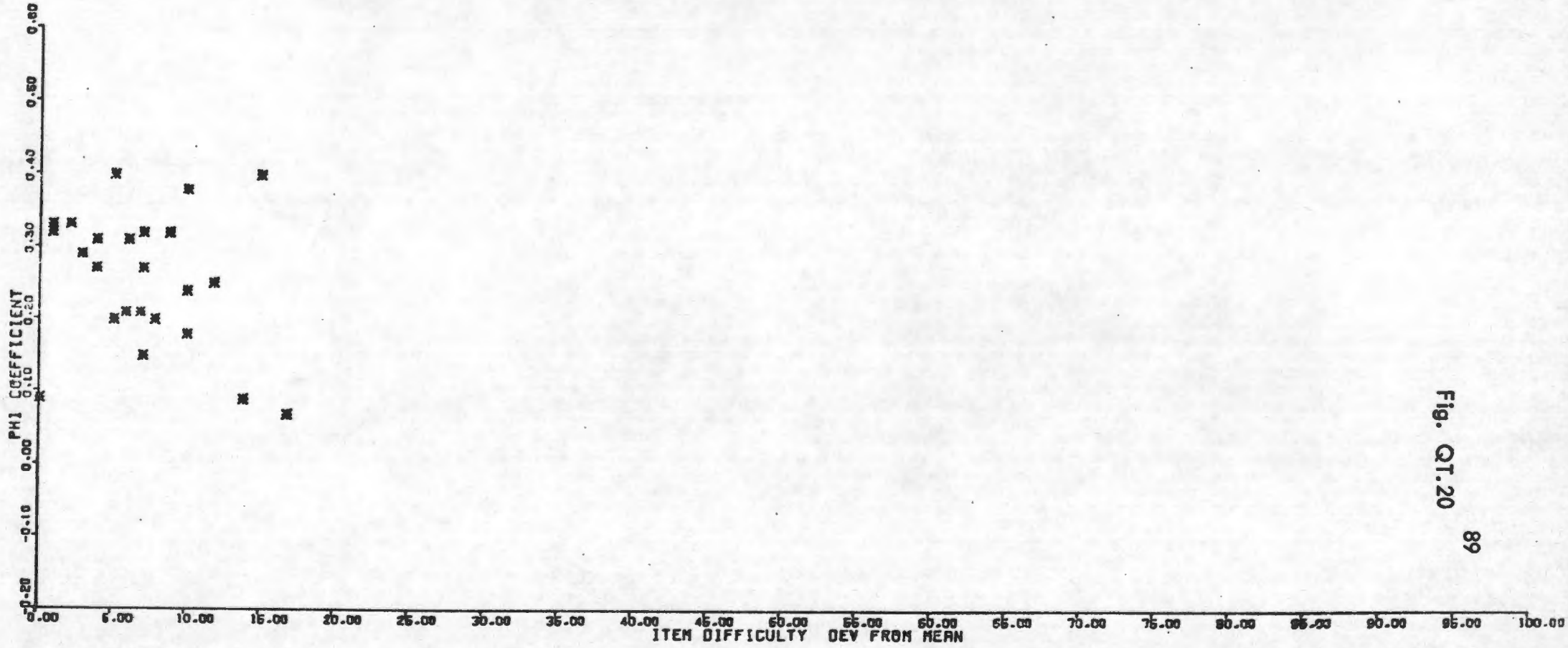


Fig. QT.20

ITEM DIFFICULTY AND PHI C0EFFICIENT

QUESTI0N TYPE RELATED/EXCLUSI0N

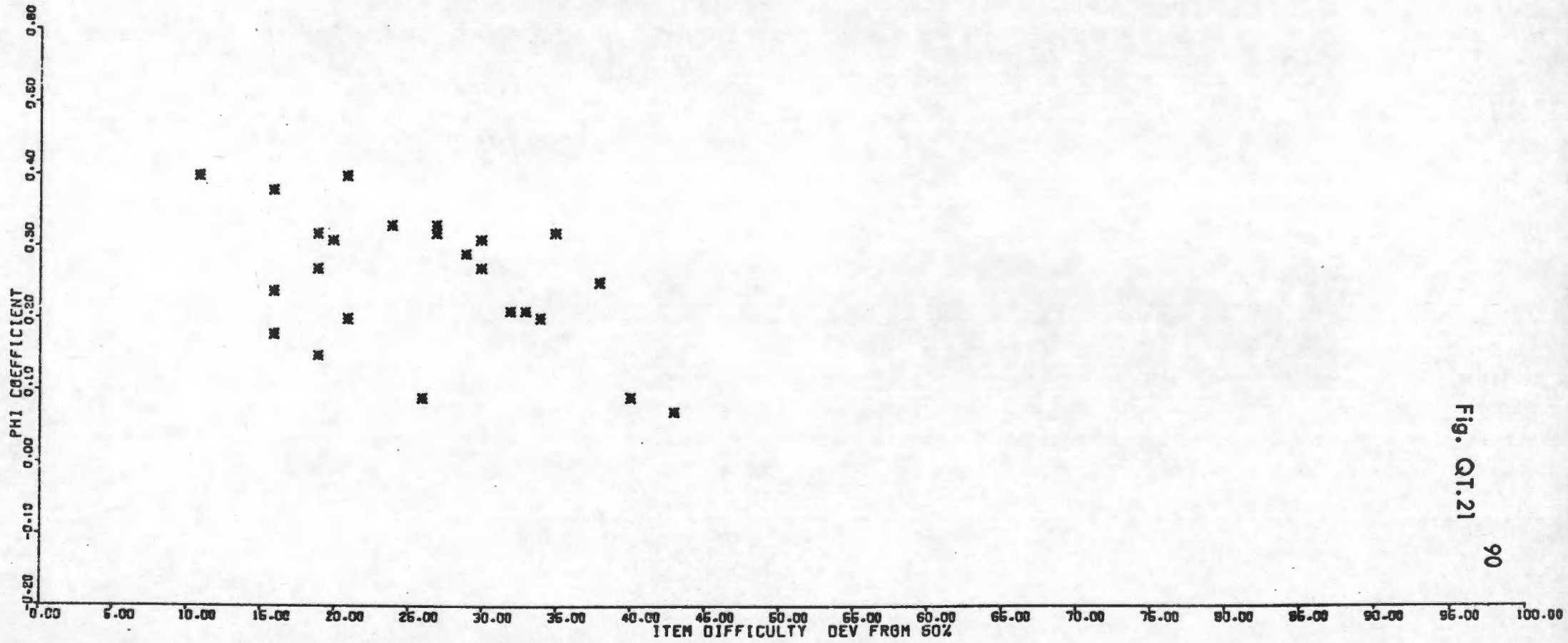


Fig. QT.21

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE TRUE/FALSE

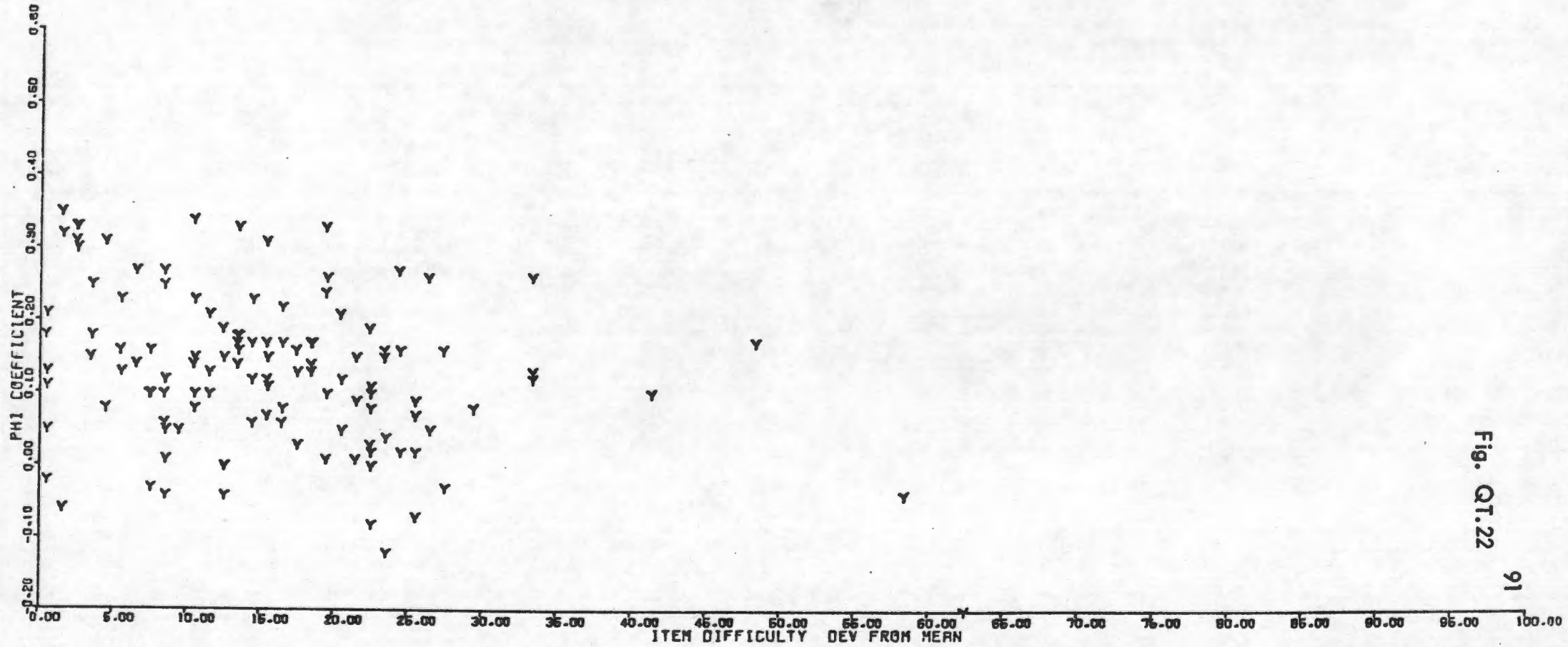


Fig. QT.22

ITEM DIFFICULTY AND PHI COEFFICIENT

QUESTION TYPE TRUE/FALSE

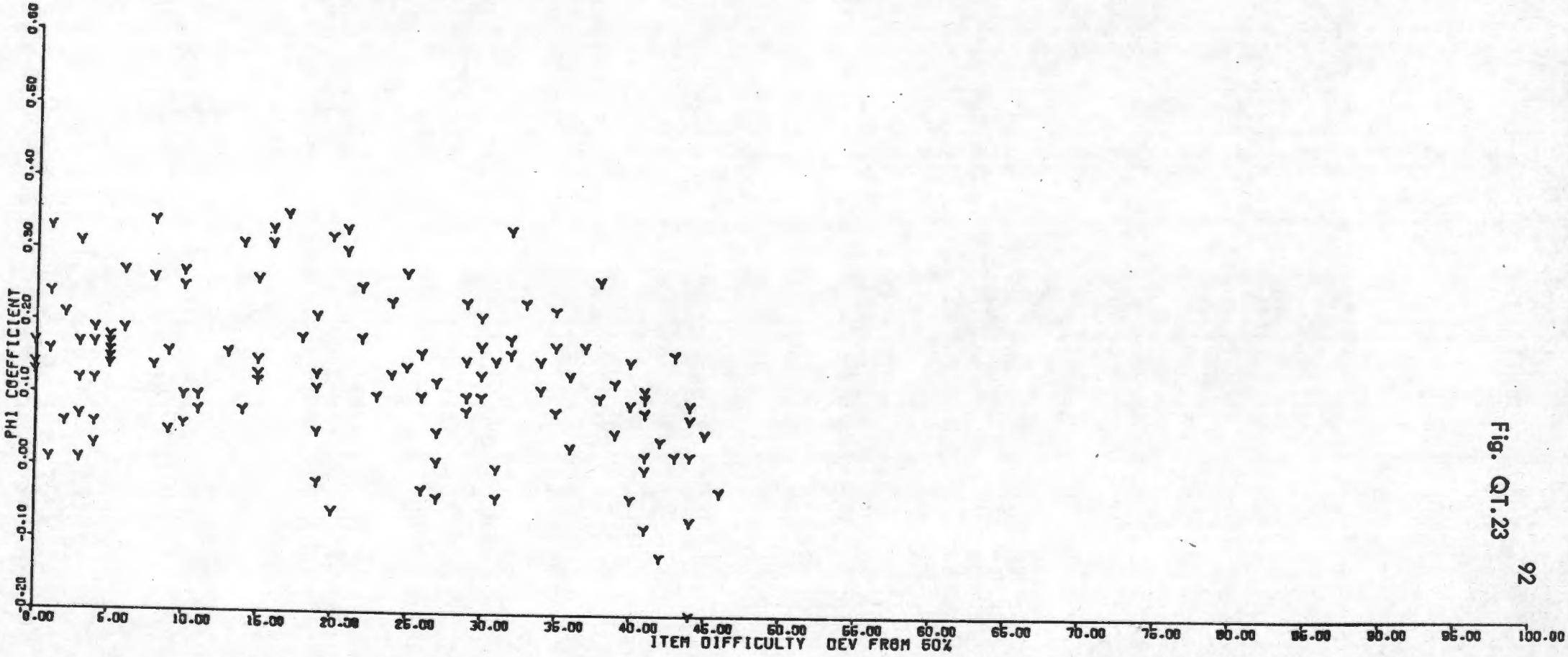


Fig. QT.23

TABLE QT.10 QUESTION TYPES

Correlation Coefficients between Deviations of Item Difficulty and PHI

Q. Type	n	Dev. from Mean	Dev. from 50%
		r =	r =
All	577	-,3556 **	-,2614 **
1/5	130	-,1555	-,1781
W/5	107	-,5057 **	-,4367 **
C/4	110	-,3308 **	-,2297 *
Causal	82	-,3562 **	-,2412 *
Related	24	-,2328	-,4717 *
T/F	124	-,3455 **	-,3618 **

* = p = <,05

** = _ = <,01

TABLE QT.11 QUESTION TYPES

Ascending Difficulty of Question Types and Correlation between PHI and Deviation of Item Difficulty. *

Q. Type	Mean I.D.	Correlation between Ø deviation from Mean I.D. r =		Ø deviation from 50% I.D. r =	
Related	76,08	-,2328	NS	-,4717	
1/5	71,32	-,1555	NS	-,1781	NS
T/F	68,44	-,3455		-,3618	
C/4	62,75	-,3308		-,2297	
W/5	57,54	-,5057		-,4367	
Causal	56,59	-,3652		-,2412	

NS = Not significant at p = <,05

* Note: as the Item Difficulty is in fact conversely related to the difficulty of a question the easier questions have higher Item Difficulty values.

TABLE QT.12 QUESTION TYPES

Attributes of 6 Types of Question Format.

Type	Item Difficulty	Discriminative Ability	Relation of Difficulty to discrimination.
1/5	Easy	Very Good	No relation
W/5	Moderate	Good	Related
C/4	Moderate	Very Good	Related
CSL	Moderate	Good	Related
RLTD	Very Easy	Excellent	Anomalous
T/F	Easy	Poor	Related

Item Difficulty		Discrimination	
Very easy	75% +	Excellent	99,9% +
Easy	65-75%	Very Good	99% +
Moderate	55-65%	Good	98% +
		Fair	95% +
Difficult	< 55%	Poor	< 95% -

Choice of Formats

1. Difficulty
   - Very Easy : RLTD
   - Easy : 1/5, T/F
   - Moderate : W/5, C/4, CSL

2. Discrimination
   - Excellent : RLTD
   - Very good : 1/5, C/4
   - Good : W/5, CSL
   - Poor : T/F

3. Discrimination and Item Difficulty
   - Not related : 1/5
   - Related : W/5, C/4, CSL, T/F
   - Anomalous : RLTD

TABLE CR.1 CONFIDENCE AND RELIABILITY

Consequences of different Strategies when using the Scoring System of Rothman

Students Knowledge	Conventional Score	Confidence Score		
	% correct by chance	% score obtained when confidence answer marked:-		
		very sure	sure	guess
Nil	20	0	20	40
Knows 1 choice incorrect	25	8,3	25	41,7
Knows 2 choices incorrect	33,3	22,2	33,3	44,4
Knows 3 choices incorrect	50,0	50,0	50,0	50,0
Knows 4 choices incorrect	100,0*	133,3*	100,0*	66,7*

* no chance involved

After Paton (1971)

TABLE CR.2 CONFIDENCE AND RELIABILITY

Consequences of different strategies when using the scoring system of Paton.

Students Knowledge	Conventional Score	Confidence Score		
	% correct	% score obtained when confidence answer marked: very sure	sure	guess
Nil	20	-46,7	-13,3	-6,7
Knows 1 choice incorrect	25	-33,3	- 4,2	0
Knows 2 choices incorrect	33,3	-11,2	+11,1	+11,1
Knows 3 choices incorrect	50,0	+33,3	+42,7	+33,3
Knows 4 choices incorrect	100,0*	+166,7*	+133,0*	+100,0*

* = no chance involved

After Paton (1971)

TABLE CR.3 CONFIDENCE AND RELIABILITY

Consequences of different Strategies when using the Scoring System proposed - Fredman

Students Knowledge	Conventional Score	Confidence Score		
	% correct by chance	% score obtained when confidence answer marked:		
		very sure	sure	guess
Nil	20	-60	-48	-12
Knows 1 choice incorrect	25	-50	-25	-10
Knows 2 choices incorrect	33,3	-33	-13,3	- 6,7
Knows 3 choices incorrect	50,0	0	+10	0
Knows 4 choices incorrect	100,0*	100*	+80*	+20*

* = no chance involved

TABLE CR.4 CONFIDENCE AND RELIABILITY

Reliability Coefficients of Conventional and Confidence Marking of the Same Examination.

Date	No.Q.	KR 20 by Conventional Mark	KR 20 by Confidence Mark
1. 271072	90	,904	1,009
2. 060373	24	,711	1,035
3. 270673	85	,876	1,008
4. 031173	42	,877	1,009
5. 090474	30	,848	1,021
6. 230474	60	,893	1,016
7. 180674	100	,984	1,009
8. 041174	149	,999	1,007
9. 051174	134	,990	1,007
10. 190475	84	,875	1,009

n = 10

RELIABILTY AND CONFIDENCE

CONVENTIONAL TO CONFIDENCE MARKING

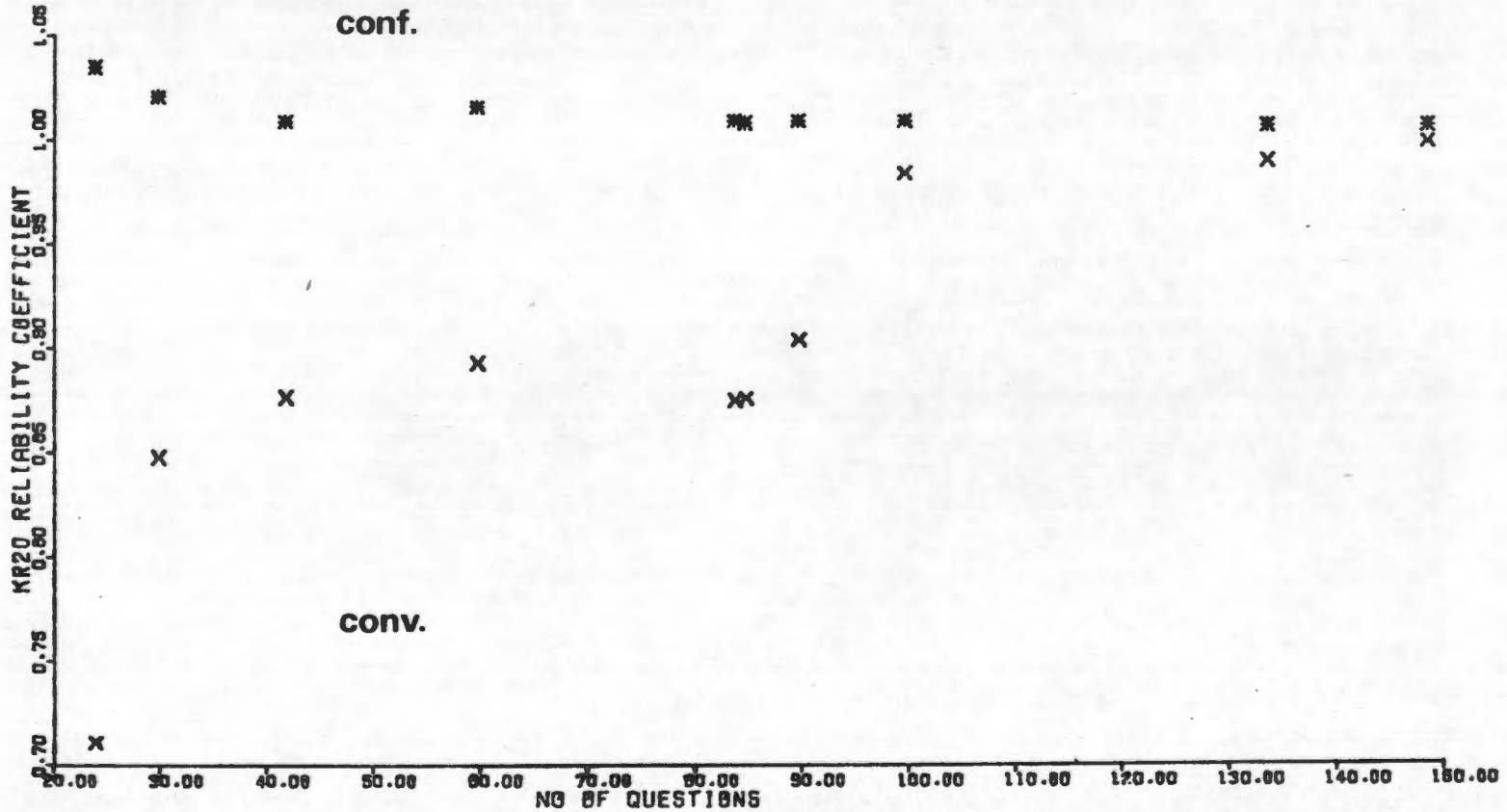


Fig. CR 1.

TABLE CR.5 CONFIDENCE AND RELIABILITY

Actual and Hypothetical Test Length and Test Efficiency .

Date.	No. Q	Actual Time Time for Conf. Examination - in minutes	Effective Test Length - Conv. Marking	Time of Equivalent Conventional Exam. in minutes	Test Efficiency
271072	90	90	155	116	1,29
060373	24	30	179	134	4,47
270673	85	90	154	115	1,71
031173	42	40	155	116	2,90
090474	30	30	166	124	4,13
230474	60	60	161	121	2,02
180674	100	90	155	116	1,72
041174	149	90	153	114	1,70
051174	134	90	153	114	1,70
190475	84	60	155	116	1,93

TABLE CR.6 CONFIDENCE AND RELIABILITY

Test Efficiency Means

No. of Questions	No. of Test	Mean Confidence Test Efficiency
20 - 59	3	3,84
60 - 90	4	1,74
More than 90	3	1,71

TABLE CAR.1 ANALYSIS OF CONFIDENCE.

Mean Values for Correct Answers in each Confidence Category.

		Mean % Correct Answers for Test.		
Test	No.Q	Guess	Sure	Very Sure
230672	75	48,54	59,15	74,09
271072	90	55,83	66,19	84,68
270673	85	42,13	52,71	67,81
230474	60	43,90	56,27	71,23
180674	100	50,51	58,10	72,19
041174	148	52,97	60,05	83,19
051174	134	50,03	58,08	75,64
190475	84	42,12	60,53	78,85
180675	99	45,99	59,87	76,89
Mean of series				
n = 9		48,00	52,42	76,01

TABLE CAR.2 ANALYSIS OF CONFIDENCE.

Responses per Quintile - % of Questions answered Very Sure (Right or Wrong)

Date	Q.1	Q.2	Q.3	Q.4	Q.5
230672	74,07	60,88	59,39	52,02	46,82
271072	81,48	72,67	66,09	59,51	51,39
270673	73,43	62,10	53,54	58,86 *	52,98
230474	76,35	72,93	61,19	56,63	52,78
180674	70,83	57,42	53,65	52,64	55,58 *
041174	88,13	81,06	74,37	61,06	50,39
051174	76,02	65,85	63,10	57,57	51,21
190475	80,74	70,38	64,51	68,36 *	56,63
180675	80,17	72,08	67,33	64,95	56,41
Mean =	77,91	68,37	62,57	59,07	57,13

TABLE CAR.3 ANALYSIS OF CONFIDENCE.

Responses per Quintile - % of questions answered as Guess (right or wrong).

Date	Q.1	Q.2	Q.3	Q.4	Q.5
230672	7,60	13,48	13,08 *	18,04	23,63
271072	6,55	8,38	12,00	12,89	23,19
270673	4,03	5,80	9,63	14,35	12,70 *
230474	7,98	10,09	15,04	18,09	17,59 *
180674	11,09	17,37	19,75	20,91	16,60 *
041174	4,21	6,92	10,32	16,15	21,54
051174	8,44	11,97	16,16	17,91	20,21
190475	3,85	6,71	10,86	8,99 *	13,62
180675	4,55	7,96	9,10	10,08	14,51
Mean =	6,48	9,85	12,88	15,26	18,17

TABLE CAR.4 ANALYSIS OF CONFIDENCE.

% of Questions correct in each Confidence Category - Quintile 1.

Date	Very Sure	Sure	Guess
230672	83,97	62,66	47,77
271072	73,93	64,03	41,07
270673	83,33	53,32	36,31
230474	82,10	61,79	43,45
180674	86,04	56,80	48,88
041174	94,54	64,99	43,12
051174	86,33	64,43	46,95
190475	89,47	55,67	34,07
180675	89,08	57,39	45,28
Mean =	85,42	60,12	42,98

TABLE CAR.5 ANALYSIS OF CONFIDENCE.

% of Questions correct in each Confidence Category - Quintile 2.

Date	Very Sure	Sure	Guess
230672	81,42	59,28	45,37
271072	86,94	63,76	46,20
270673	77,03	58,31	36,62
230474	76,62	52,97	40,88
180674	80,90	58,11	50,11
041174	89,57	63,75	48,95
051174	82,55	62,49	43,35
190475	86,70	56,04	37,43
180675	84,71	58,91	30,81
Mean =	82,93	59,29	42,19

TABLE CAR.6 CONFIDENCE ANALYSIS.

% of Questions correct in each Confidence Category - Quintile 3.

Date	Very Sure	Sure	Guess
230672	77,25	56,73	46,52
271072	88,19	66,14	44,40
270673	73,95	46,99	36,76
230474	75,11	54,63	40,50
180674	76,55	56,00	43,34
041174	85,53	61,73	37,69
051174	79,81	54,97	47,09
190475	86,62	56,42	34,81
180675	81,46	53,00	45,83
Mean =	79,72	56,29	41,88

TABLE CAR.7 CONFIDENCE ANALYSIS.

% of Questions correct in each Confidence Category - Quintile 4.

Date	Very Sure	Sure	Guess
230672	75,64	54,75	40,50
271072	85,38	59,41	38,31
270673	68,54	42,74	25,84
230474	73,04	53,22	37,79
180674	71,09	50,75	45,53
041174	81,87	57,18	41,16
051174	74,88	51,75	47,37
190475	79,98	42,48	33,25
180675	76,27	48,40	34,49
Mean =	76,28	51,19	38,25

TABLE CAR.8 CONFIDENCE ANALYSIS.

% of Questions correct in each Confidence Category - Quintile 5.

Date	Very Sure	Sure	Guess
230672	66,46	45,85	32,47
271072	75,30	53,18	39,86
270673	52,81	35,29	26,90
230474	63,26	44,09	28,05
180674	63,22	44,72	33,61
041174	66,03	48,90	41,27
051174	69,78	49,74	37,82
190475	74,21	44,75	26,55
180675	71,24	42,63	28,53
Mean =	59,59	45,46	32,78

TABLE CAR . 9 CONFIDENCE ANALYSIS.

Confidence Score - Raw and as a Percentage per Quintile.

Date	Q1		Q2		Q3		Q4		Q5		Class Mean	
	Raw	%	Raw	%	Raw	%	Raw	%	Raw	%	Raw	%
230672	192,36	85,49	176,39	78,40	171,64	76,28	158,45	70,42	152,47	67,76	170,16	75,63
271072	243,28	90,10	229,58	85,03	215,09	79,66	206,97	76,66	190,27	70,47	217,04	80,38
270673	217,49	85,29	201,68	79,09	188,97	74,11	186,17	73,32	180,81	70,91	195,09	76,51
230474	154,27	85,70	151,81	84,34	133,95	74,42	124,59	69,22	122,92	68,92	137,75	76,53
180674	244,34	81,45	215,91	71,97	209,67	69,89	212,97	70,99 *	179,78	59,93	212,33	70,78
041174	259,46	94,01	243,86	88,36	233,17	84,48	207,14	75,05	181,80	65,87	225,03	81,53
051174	345,57	85,96	313,33	77,94	302,46	75,23	295,29	73,46	277,25	68,97	306,60	76,27
190475	226,44	89,86	210,62	83,58	198,11	78,62	204,95	81,33 *	183,86	72,96	204,76	81,25
180675	239,67	87,79	224,25	82,14	217,59	79,70	210,67	77,17	193,05	70,71	216,92	79,46
Mean =		87,29		81,21		76,93		74,18		68,50		77,59

TABLE CAR. 10 CONFIDENCE ANALYSIS.

Mean Confidence Index per Quintile.

Date	Q1	Q2	Q3	Q4	Q5	Class Mean.
230672	2,66	2,48	2,46	2,34	2,24	2,44
271072	2,75	2,64	2,54	2,47	2,29	2,54
270673	2,65	2,50	2,42	2,44	2,35	2,47
230474	2,68	2,63	2,46	2,38	2,36	2,50
180674	2,60	2,41	2,34	2,32	2,39	2,41
041174	2,84	2,74	2,64	2,45	2,31	2,60
051174	2,68	2,54	2,47	2,40	2,36	2,49
190475	2,77	2,64	2,54	2,60*	2,44	2,60
180675	2,76	2,64	2,58	2,55	2,42	2,59
Mean =	2,71	2,58	2,49	2,44	2,35	2,51

TABLE CAR.11 CONFIDENCE ANALYSIS.

Correlation between Knowledge & Confidence.

Knowledge = Number of questions correct.

Date	n	Confidence Score	Confidence Index	Class Confidence Ratio
230672	166	,597	,446	,467
271072	163	,497	,355	,356
270673	178	,579	,351	,352
230474	181	,733	,503	,504
180674	176	,798	,305	,305
041174	179	,810	,673	,672
051174	179	,671	,485	,483
190475	184	,618	,442	,447
180675	182	,654	,411	,408
Total	1588	,674	,412	,389

All significant at p = <,01

TABLE CAR.12 CONFIDENCE ANALYSIS.

Correlation Coefficients between Knowledge and Confidence.

Knowledge = No. wrong.

Date	n	Confidence Score		Confidence Index	Class Confidence Ratio
230672	166	-,171	N	-,308	-,309
271072	163	-,180	N	-,246	-,245
270673	178	-,061	N	-,280	-,281
230474	181	-,155	N	-,349	-,351
180674	176	-,120	N	-,282	-,280
041174	179	-,439		-,544	-,543
051174	179	-,105	N	-,315	-,314
190475	184	-,219		-,316	-,318
180675	182	-,226		-,373	-,369
Total	1588	-,221		-,356	-,285

N = Not significant at p = <,01

TABLE CAR.13 CONFIDENCE ANALYSIS

Correlation Coefficients between Knowledge and Confidence.

Knowledge = Number of questions correct less No. Wrong.

Date	n	Confidence Score	Confidence Index	Class Confidence Ratio
230672	166	,431	,410	,411
271072	163	,366	,318	,318
270673	178	,372	,351	,352
230474	181	,540	,484	,485
180674	176	,551	,368	,368
041174	179	,691	,652	,651
051174	179	,460	,451	,449
190475	184	,465	,423	,416
180675	182	,500	,408	,419
Total	1588	,507	,445	,370

All significant at p = <,01

TABLE CAR.14 CONFIDENCE ANALYSIS

Correlation between Knowledge and Confidence.

Knowledge = No. Correct + ½ of No Attempt.

Date	n	Confidence Score	Confidence Index	Class Confidence Ratio
230672	166	,431	,410	,411
271072	163	,366	,318	,318
270673	178	,372	,351	,352
230474	181	,540	,484	,485
180674	176	,551	,368	,368
041174	179	,690	,652	,651
051174	179	,460	,451	,449
190475	184	,500	,408	,416
180675	182	,465	,423	,419
	1588	,507	,446	,369

All significant at p = < 01,

TABLE CAR.15 CONFIDENCE ANALYSIS

Correlation between Knowledge and Confidence

Knowledge = No. Correct + ½ of No Guess - No. Wrong.

Date	n	Confidence Score	Confidence Index	Class Confidence Ratio
230672	166	,309	,367	,368
271072	163	,278	,286	,286
270673	178	,220	,324	,325
230474	181	,367	,433	,435
180674	176	,242	,356	,354
041174	179	,584	,613	,611
051174	179	,292	,395	,394
190475	184	,351	,370	,372
180675	182	,378	,407	,403
Total	1588	,372	,412	,336

All significant at p = < 01,

TABLE CAR.16 CONFIDENCE ANALYSIS

Correlation of Confidence to Change in Results.

Confidence = Confidence Score.

Date	n	Difference in % Score	Difference in Z-Score	Difference in Rank
230672	166	-,270	+,012 N	+,030 N
271072	163	-,573	+,249	+,176 N
270673	178	+,120 N	-,492	-,449
230474	181	+,070 N	-,664	-,359
180674	176	+,180 N	-,335	-,395
041174	179	-,469	-,133 N	+,005 N
051174	179	+,002	-,273	-,214
190475	184	-,142 N	-,396	-,282
180675	182	-,092 N	-,418	-,352
Total	1588	-,191	-,347	-,250

N = Not significant at p = <,01

TABLE CAR.17 CONFIDENCE ANALYSIS

Confidence to Change in Results.

Confidence = Confidence Index

Date	n	Difference in % Score	Difference in Z-Scores	Difference in Rank
230672	166	-,458	+,226	+,280
271072	163	-,682	+,443	+,374
270673	178	-,070 N	-,174 N	-,125 N
230474	181	-,054 N	-,329	-,054 N
180674	176	+,114 N	-,216	-,253
041174	179	-,603	+,140 N	+,239
051174	179	-,167 N	+,065 N	-,036 N
190475	184	-,210	-,165 N	-,047 N
180675	182	-,246	-,079 N	-,029 N
Total	1588	-,328	-,252	-,140

N = Not significant at p = <,01

TABLE CAR.18 CONFIDENCE ANALYSIS

Confidence to Change in Results.

Confidence = Class Confidence Ratio

Date	n	Difference in % Scores	Difference in Z-Scores	Difference in Rank
230672	166	-,456	+,264	+,279
271072	163	-,682	+,444	+,373
270673	178	-,070 N	-,176 N	-,126 N
230474	181	-,143 N	-,329	-,055 N
180674	176	+,113 N	-,215	-,252
041174	179	-,601	+,138	+,239
051174	179	-,167 N	-,063 N	+,037 N
190475	184	-,213	-,165 N	-,050 N
180675	182	-,242	-,082 N	-,032 N
Total	1588	-,271	-,220	-,115

N = Not significant at p = <,01

TABLE CAR.19 CONFIDENCE ANALYSIS

Confidence to Change in Results.

Confidence = Quintile Confidence Ratio

Date	n	Difference in % Scores	Difference in Z-Scores	Difference in Rank
230674	166	-,213	+,201 N	+,230
270673	178	+,135 N	-,227	-,169 N
230474	181	+,142 N	-,196 N	-,132 N
180674	176	+,256	-,291	-,286
041174	179	-,104 N	+,053 N	+,226
051174	179	+,115 N	-,163 N	-,025 N
190675	184	+,105 N	-,211	-,083 N
180675	182	+,069 N	-,128 N	-,068 N

N = Not significant at p = <,01

TABLE CAR.20 CONFIDENCE ANALYSIS

Knowledge to Change in Results.

Knowledge = Number of Questions Correct.

Date	n	Difference in % Scores	Difference in Z-Scores	Difference in Rank
230672	166	-,315	-,169 N	-,165 N
271072	163	-,393	-,348	-,130 N
270673	178	-,406	-,241	-,228
230474	181	-,263	-,664	-,208
180674	176	-,046 N	-,193 N	-,225
041174	179	-,621	-,156 N	-,119 N
051174	179	-,240	-,230	-,223
190475	184	-,627	-,184 N	-,156 N
180675	182	-,537	-,191 N	-,207
Total	1588	-,442	-,306	-,144

N = Not significant at p = <,01

TABLE CAR.21 CONFIDENCE ANALYSIS

Knowledge to Change in Results.

Knowledge = Number of questions Wrong.

Date	n	Difference in % Score	Difference in Z-Score	Difference in Rank
230672	166	+,529	+,144 N	-,106 N
271072	163	+,433	+,233	-,036 N
270673	178	+,739	-,310	-,317
230474	181	+,671	+,076 N	-,310
180674	176	+,350	-,178	-,200
041174	179	+,768	-,170	-,058 N
051174	179	+,507	-,131 N	-,148 N
190475	184	+,796	-,196	-,177
180675	182	+,810	-,280	-,245
Total	1588	+,649	+,125	+,245

N = Not significant at p = <,01

TABLE CAR.22 CONFIDENCE ANALYSIS

Knowledge to Change in Results

Knowledge = Questions Right - Questions Wrong.

Date	n	Difference in % Scores	Difference in Z-Scores	Difference in Rank
230674	166	-,439	-,029 N	-,044 N
271072	163	-,430	-,304	-,056 N
270673	178	-,620	+,018 N	+,029 N
230474	181	-,481	-,459	+,011 N
180674	176	-,204	-,061 N	-,075 N
041174	179	-,621	-,156 N	-,119 N
051174	179	-,398	-,075 N	-,061 N
190475	184	-,743	-,075 N	-,007 N
180675	182	-,704	+,007 N	-,000 N
Total	1588	-,573	-,244	-,069 N

N = Not significant at p = <,01

TABLE CAR.23 CONFIDENCE ANALYSIS

Correlation between Knowledge and Confidence - All Examinations.

Knowledge as	1 % Confidence Score	2 Confidence Index	3 Class Confidence Ratio
1. % of No. of questions Right	+,674	+,459	+,389
2. % of No. of questions Wrong	-,211	-,356	-,285
3. % of No. of questions Right - number wrong	+,507	+,455	+,370
4. % of No. of questions Right + ½ No Guess	+,507	+,446	+,369
5. % of No. of questions Right + ½ No Guess - No. Wrong	+,372	+,412	+,336

n = 1588

All Significant at p = <,01

TABLE CAR.24 CONFIDENCE ANALYSIS

Correlation between Confidence and Change in Results - All Examinations

	1	2	3
Confidence as	Difference in Score	Difference in Z-Score	Difference in Rank
1. % Confidence Score	-,191	-,347	-,250
2. Confidence Index	-,328	-,252	-,140
3. Class Confidence Ratio	-,271	-,220	-,115

n = 1588

All Significant at p = <,01

TABLE CAR.25 CONFIDENCE ANALYSIS

Correlation between Knowledge and Change in Results - All Examinations.

Knowledge as	1 Difference in Score	2 Difference In Z-Score	3 Difference in Rank	
1. Percentage No. of questions correct	-,442	-,306	-,144	
2. Percentage No. of questions wrong	,649	,125	,245	
3. Percentage No. of questions right - wrong	-,573	-,244	-,069	N
4. Percentage No. of questions right + ½ NG	-,625	-,192	-,019	N
5. Percentage No. of questions right + ½ NG - wrong	-,575	-,246	-,071	N

n = 1588

N = Not significant at p = <,01

APPENDIX A.

MULTIPLE-CHOICE QUESTION FORMATS IN USE

IN THE DEPARTMENT OF ANATOMY

UNIVERSITY OF CAPE TOWN

1. One From Five:

In this format a stem or statement is set out followed by 5 alternatives, of which one is the correct answer. The student is required to select the correct answer:-

e.g. Complete the following sentences (questions 1 - 10) by choosing the best of the alternatives 1-5 below. If you do NOT know the answer please mark 0.

Example:

The venous sinus in the inferior border of the falx cerebri is the:

1. inferior sagittal
2. transverse
3. sigmoid
4. straight
5. cavernous

2. Wrong from Five:

In this format a stem or statement is followed by 4 alternatives, one of which may be wrong, or alternatively all 4 may be correct. The student is required to select the incorrect answer, or to choose 5, (all statements are correct), if none of the alternatives is wrong:-

e.g. Select the INCORRECT statement from statements 1 to 4 in the following questions (11-15) or 5 if all statements are correct.

Example:

The ophthalmic nerve:

1. is a division of the trigeminal nerve
2. passes through foramen rotundum
3. is a sensory nerve
4. gives off the frontal nerve
5. all 4 statements are correct

3./.....

3. Correct from Four:

In this format the stem is followed by 4 alternatives any number of which may be correct. The student is required to identify the correct statements by using the key below:-

e.g. Answer the following questions (16-20) according to the following key:

1. If A, B, C are correct
2. If A, C are correct
3. If B, D are correct
4. If any other combination (including all 4 statements), or only one of the statements, is correct
5. If none of the 4 statements is correct

Example:

Structures passing through the jugular foramen include:

A. vagus nerve
B. hypoglossal nerve
C. glossopharyngeal nerve
D. superior petrosal sinus

4. Causal:

In this format two statements are linked with the word "because". Both statements may be correct and there may or may not be a causal linkage, or only one, or neither statement may be correct. The student is required to identify which of these situations exists by the use of the following key:-

e.g. Answer the following questions (21-24) according to the following key:

1. If statements A & B are both true and their relation IS causal
2. If statements A & B are both true but their relation is NOT causal
3. If statement A is true and B is false

4./.....

4. If statement A is false and B is true
5. If statement A is false and B is also false

Example:

A. In carpal tunnel compression loss of sensation and muscle power is evident in both ulnar and median nerve distribution

BECAUSE

B. both these nerves pass deep to the flexor retinaculum at the wrist.

5. Related - Exclusion:

In this format the student is presented with 2 columns. In column 1 on the left hand side the student is required to identify which one statement is not in the same category as the other 4 (i.e. the odd man out) and to choose the category from the right hand column (column 2) that establishes the exclusion applicable in column 1:-

e.g. In column 1 there are 5 structures mentioned - one of these is NOT in the same group as the other four.

In the first question state which structure this is.

In the second question choose the group (from Column 2) to which the four related structures refer.

Example:

Question 25		Question 26	
1	coeliac	1	parietal branches of aorta
2	middle suprarenals	2	visceral branches of aorta
3	testicular	3	tributaries of inferior vena cava
4	superior mesenteric	4	tributaries of portal vein
5	common iliacs	5	structures found only in the foetus

(In the above two questions in September 1970 86% of students correctly identified the common iliacs as being different from the other 4 arteries named, and 77% of students identified it as not being a visceral branch of the aorta.

6./.....

6. True or False:

In this format a stem is presented followed by a number of statements each of which might be true or false. The student is required to state in each case whether the statements are true or false:-

e.g. In this section there is a heading followed by subsidiary statements referring to the heading. Each statement is numbered and constitutes a question. If the statement is TRUE mark 1 as the answer to the question; if the statement is FALSE, mark 2 as the answer.

Example:

The Ureter:

27. Crosses above the uterine artery lateral to the cervix
28. Crosses the origin of external iliac artery
29. Runs forwards in base of broad ligament
30. Descends on the lateral pelvic wall
31. Changes direction at the level of the ischial spine
32. Is crossed anteriorly by the ductus deferens.

APPENDIX B

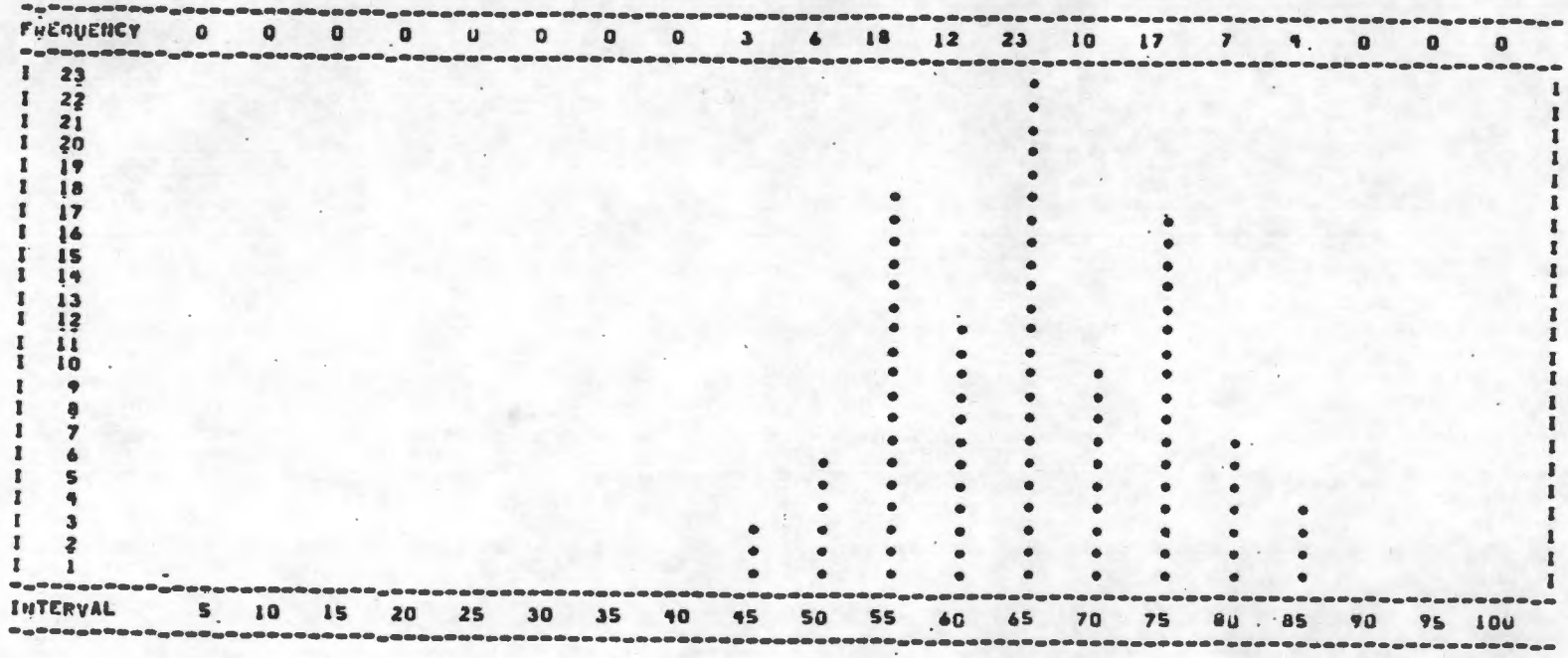
SAMPLE COMPUTER LISTINGS

OF COMPUTER PROGRAM TO

CORRELATE COMPONENTS OF AN EXAMINATION.

47.00	24.00	41.00	12.00	17.00	32.00	20.00	8.00	1.00	29.00	65.00
72.00	29.00	51.00	13.00	18.00	37.00	19.00	10.00	6.00	37.00	67.00
62.00	26.00	51.00	14.00	21.00	35.00	22.00	11.00	10.00	34.00	68.00
62.00	29.00	44.00	13.00	13.00	28.00	21.00	9.00	8.00	31.00	64.00
55.00	26.00	46.00	12.00	19.00	34.00	21.00	10.00	4.00	32.00	70.00
59.00	28.00	45.00	15.00	21.00	28.00	16.00	9.00	5.00	36.00	64.00
54.00	26.00	42.00	13.00	16.00	31.00	22.00	8.00	4.00	29.00	64.00
75.00	33.00	52.00	12.00	21.00	39.00	25.00	11.00	5.00	41.00	68.00
69.00	30.00	49.00	12.00	17.00	37.00	17.00	9.00	7.00	37.00	60.00
47.00	27.00	32.00	10.00	9.00	28.00	19.00	8.00	1.00	24.00	50.00
60.00	29.00	44.00	10.00	15.00	37.00	22.00	8.00	3.00	36.00	60.00
71.00	31.00	50.00	13.00	21.00	37.00	20.00	10.00	8.00	37.00	65.00
35.00	22.00	30.00	10.00	10.00	30.00	19.00	3.00	4.00	21.00	48.00
62.00	30.00	48.00	15.00	20.00	31.00	15.00	8.00	8.00	31.00	73.00
63.00	26.00	43.00	12.00	20.00	37.00	21.00	10.00	5.00	34.00	69.00
55.00	28.00	39.00	11.00	16.00	33.00	22.00	9.00	5.00	32.00	63.00
64.00	28.00	41.00	14.00	20.00	34.00	20.00	9.00	7.00	33.00	74.00
66.00	26.00	45.00	13.00	21.00	36.00	20.00	12.00	6.00	39.00	70.00
70.00	29.00	49.00	12.00	16.00	33.00	22.00	9.00	6.00	41.00	63.00
43.00	27.00	33.00	12.00	10.00	26.00	17.00	10.00	4.00	29.00	54.00
51.00	21.00	34.00	12.00	12.00	20.00	20.00	8.00	4.00	30.00	43.00
43.00	23.00	22.00	10.00	10.00	23.00	22.00	11.00	4.00	21.00	55.00
69.00	30.00	46.00	17.00	24.00	37.00	23.00	14.00	4.00	38.00	70.00
52.00	28.00	44.00	10.00	16.00	33.00	17.00	9.00	4.00	31.00	62.00
56.00	27.00	42.00	12.00	14.00	30.00	20.00	8.00	6.00	30.00	68.00
36.00	18.00	18.00	12.00	11.00	25.00	14.00	8.00	5.00	22.00	55.00
68.00	30.00	51.00	11.00	22.00	34.00	23.00	9.00	4.00	35.00	73.00
60.00	26.00	44.00	11.00	16.00	32.00	22.00	10.00	6.00	28.00	68.00
46.00	28.00	34.00	11.00	11.00	29.00	20.00	8.00	8.00	29.00	54.00
67.00	29.00	46.00	13.00	19.00	34.00	18.00	10.00	4.00	35.00	71.00
54.00	27.00	37.00	13.00	19.00	31.00	18.00	10.00	5.00	26.00	61.00
66.00	29.00	43.00	11.00	15.00	33.00	20.00	9.00	4.00	32.00	60.00
54.00	27.00	38.00	11.00	18.00	25.00	21.00	11.00	5.00	31.00	70.00
63.00	28.00	54.00	15.00	25.00	40.00	23.00	12.00	10.00	39.00	85.00
73.00	31.00	55.00	13.00	23.00	38.00	22.00	10.00	6.00	39.00	75.00
74.00	32.00	53.00	13.00	22.00	37.00	19.00	10.00	6.00	36.00	73.00
46.00	25.00	30.00	12.00	8.00	18.00	15.00	8.00	2.00	23.00	55.00
55.00	24.00	42.00	12.00	15.00	32.00	22.00	9.00	5.00	39.00	64.00
45.00	29.00	36.00	14.00	18.00	32.00	16.00	8.00	8.00	33.00	62.00
57.00	29.00	47.00	11.00	16.00	39.00	19.00	9.00	4.00	36.00	67.00
70.00	29.00	47.00	14.00	23.00	37.00	23.00	10.00	8.00	35.00	58.00
53.00	28.00	35.00	12.00	15.00	24.00	16.00	9.00	2.00	35.00	63.00
69.00	30.00	53.00	13.00	22.00	38.00	19.00	9.00	4.00	35.00	63.00
35.00	24.00	27.00	11.00	7.00	26.00	19.00	6.00	2.00	23.00	48.00
73.00	30.00	49.00	12.00	24.00	32.00	21.00	13.00	5.00	39.00	68.00
70.00	29.00	53.00	13.00	25.00	38.00	21.00	10.00	7.00	36.00	70.00
35.00	17.00	28.00	12.00	13.00	30.00	20.00	11.00	3.00	27.00	58.00
62.00	29.00	47.00	12.00	16.00	34.00	19.00	11.00	7.00	31.00	67.00
54.00	29.00	47.00	12.00	19.00	34.00	20.00	11.00	5.00	31.00	70.00
62.00	27.00	30.00	13.00	13.00	25.00	14.00	9.00	3.00	31.00	52.00
53.00	25.00	34.00	13.00	14.00	26.00	16.00	10.00	2.00	26.00	55.00
53.00	28.00	44.00	13.00	22.00	31.00	24.00	10.00	5.00	33.00	72.00

STUDENTS MARK PER COMPONENT



HISTOGRAM OF QUESTION 10 WITH MARKS EXPRESSED AS PERCENTAGES.

NUMBER OF OBSERVATIONS 177

NUMBER OF VARIABLES 11

VARIABLE	MEAN	STD DEV	TOP	MEDIAN
1	57.4	9.9	77.0	57.0
2	68.9	6.8	82.5	70.0
3	69.5	13.0	96.7	71.7
4	60.5	8.3	85.0	60.0
5	54.2	14.7	83.3	53.3
6	62.4	10.0	80.0	64.0
7	78.0	10.3	100.0	80.0
8	61.4	10.7	93.3	60.0
9	50.3	19.6	100.0	50.0
10	62.3	9.9	84.0	62.0
11	63.7	7.6	85.0	64.0

MATRIX OF CORRELATION COEFFICIENTS

1.000										
.677	1.000									
.803	.631	1.000								
.363	.255	.397	1.000							
.634	.389	.688	.378	1.000						
.671	.480	.726	.215	.606	1.000					
.268	.234	.344	.211	.174	.282	1.000				
.464	.291	.419	.255	.544	.383	.189	1.000			
.352	.225	.404	.228	.367	.334	.172	.253	1.000		
.735	.516	.738	.343	.657	.660	.229	.458	.417	1.000	
.598	.521	.635	.383	.529	.554	.205	.422	.327	.614	1.000

VARIABLE 10 COMPARED WITH THE SUM OF THE OTHERS

MATRIX OF CORRELATION COEFFICIENTS

1.000

.803 1.000

APPENDIX C

SAMPLE COMPUTER LISTINGS OF COMPUTER PROGRAMS TO ANALYZE RESULTS OF CONFIDENCE MARKING

ANALYSIS OF MCQ CLASS TEST 180675 MARKED BY CONFIDENCE.

	RG	WG	RS	WS	RVS	WVS
UNWEIGHTED MARKS :-	1.00	1.00	4.00	3.00	5.00	5.00

a

NUMBER OF BLOCKS :- 2

b

BLK	WT	WTS	BLK	WT	WTS
1	3	78	2	1	22

	QUINTILE 1	QUINTILE 2	QUINTILE 3	QUINTILE 4	QUINTILE 5
QUINTILE RANGES (OLD) :-	69.14	61.51	53.14	44.46	24.14

d

	QUINTILE 1	QUINTILE 2	QUINTILE 3	QUINTILE 4	QUINTILE 5
QUINTILE RANGES (NEW) :-	52.19	42.03	33.44	22.50	.39

ANALYSIS OF CONFIDENCE MCQ 180675

STUDENT NUMBERS	V. SURE R (%R)	V. SURE W (%W)	SURE R (%R)	SURE W (%W)	GUESS R (%R)	GUESS W (%W)	TOTAL R (%R)	TOTAL W (%W)	TOTAL V. SURE (%T) (%A)	TOTAL SURE (%T) (%A)	TOTAL GUESS (%T) (%A)	NO GUESS (%T) (%A)	NEW SCORE (%)	OLD SCORE (%)	QUIN TILE (N) (U)	RANK (N) (U)
75001	59 77.6	17 22.4	4 44.4	5 55.6	5 71.4	2 28.6	68 73.9	24 26.1	76 76.0 82.6	9 9.0 9.8	7 7.0 7.6	8 8.0 8.7	540.00 42.19	63.08	2 2	71 69
75002	54 81.8	12 18.2	15 53.6	13 46.4	2 33.3	4 66.7	71 71.0	29 29.0	66 66.0 66.0	28 28.0 28.0	6 6.0 6.0	0 .0 .0	543.00 42.42	61.51	2 2	70 74
75003	52 83.9	10 16.1	6 40.0	9 60.0	4 50.0	4 50.0	62 72.9	23 27.1	62 62.0 72.9	15 15.0 17.6	8 8.0 9.4	15 15.0 17.6	501.00 39.14	56.17	3 3	88 100
75004	34 87.2	5 12.8	11 42.3	15 57.7	10 52.6	9 47.4	55 65.5	29 34.5	39 39.0 46.4	26 26.0 31.0	19 19.0 22.6	16 16.0 19.0	351.00 27.42	49.48	4 4	132 130
75005	35 76.1	11 23.9	20 57.1	15 42.9	7 41.2	10 58.8	62 63.3	36 36.7	46 46.0 46.9	35 35.0 35.7	17 17.0 17.3	2 2.0 2.0	358.00 27.97	50.31	4 4	130 122
75006	40 72.7	15 27.3	12 60.0	8 40.0	5 50.0	5 50.0	57 67.1	28 32.9	55 55.0 64.7	20 20.0 23.5	10 10.0 11.8	15 15.0 17.6	337.00 26.33	46.64	4 4	141 134
75007	52 64.2	29 35.8	7 58.3	5 41.7	0 .0	0 .0	59 63.4	34 36.6	81 81.0 67.1	12 12.0 12.9	0 .0 .0	7 7.0 7.5	400.00 31.25	55.44	4 3	118 102
75008	21 75.0	7 25.0	14 46.7	16 53.3	8 27.6	21 72.4	43 49.4	44 50.6	28 28.0 32.2	30 30.0 34.5	29 29.0 33.3	13 13.0 14.9	167.00 13.05	30.86	5 5	169 175
75009	73 82.0	16 18.0	5 45.5	6 54.5	0 .0	0 .0	78 78.0	22 22.0	89 89.0 89.0	11 11.0 11.0	0 .0 .0	0 .0 .0	705.00 55.08	73.33	1 1	25 19
75010	50 90.9	5 9.1	18 72.0	7 28.0	3 100.0	0 .0	71 85.5	12 14.5	55 55.0 66.3	25 25.0 30.1	3 3.0 3.6	17 17.0 20.5	729.00 56.95	71.23	1 1	19 30
75011	55 67.1	27 32.9	3 50.0	3 50.0	2 25.0	6 75.0	60 62.5	36 37.5	82 82.0 85.4	6 6.0 6.3	8 8.0 8.3	4 4.0 4.2	341.00 26.64	50.21	4 4	138 149
75012	41 70.7	17 29.3	10 35.7	18 64.3	2 25.0	6 75.0	53 56.4	41 43.6	58 58.0 61.7	28 28.0 29.8	8 8.0 8.5	6 6.0 6.4	220.00 17.19	40.17	5 5	162 162
75013	58 95.1	3 4.9	10 90.9	1 9.1	12 52.2	11 47.8	80 84.2	15 15.8	61 61.0 64.2	11 11.0 11.6	23 23.0 24.2	5 5.0 5.3	843.00 65.86	80.23	1 1	4 4
75015	77 93.9	5 6.1	4 100.0	0 .0	3 33.3	6 66.7	84 88.4	11 11.6	82 82.0 86.3	4 4.0 4.2	9 9.0 9.5	5 5.0 5.3	979.00 76.48	84.52	1 1	2 3
75016	45 90.0	5 10.0	20 66.7	10 33.3	7 53.8	6 46.2	72 77.4	21 22.6	50 50.0 53.8	30 30.0 32.3	13 13.0 14.0	7 7.0 7.5	669.00 52.27	67.36	1 2	35 44
	a	b	e	f	j	k	n	o	r	u	x	aa	dd		gg	hh	ii	kk
75017	38 84.4	7 15.6	19 79.2	5 20.8	10 45.5	12 54.5	67 73.6	24 26.4	45 45.0 49.5	24 24.0 26.4	22 22.0 24.2	9 9.0 9.9	580.00 45.31	63.39	2 2	55 62
	c	d	g	h	l	m	p	q	s t	v w	y z	bb cc	ee	ff				
75018	40 90.9	4 9.1	14 66.7	7 33.3	8 42.1	11 57.9	62 73.8	22 26.2	44 44.0 52.4	21 21.0 25.0	19 19.0 22.6	16 16.0 19.0	526.00 41.09	59.21	3 3	77 83

CPO 2.

ANALYSIS OF COOFIDENCE MCQ 180675

	QUINTILE 1		QUINTILE 2		QUINTILE 3		QUINTILE 4			QUINTILE 5	
	MEAN (S.D.)	NO. QUES (%)	MEAN (S.D.)	NO. QUES (%)	MEAN (S.D.)	NO. QUES (%)	MEAN (S.D.)	NO. QUES (%)		MEAN (S.D.)	NO. QUES (%)
V. SURE (RIGHT)	87.99	2290	83.13	1935	81.57	1672	79.75	1547	a	69.33	1207 c
(% OF TOTAL (RIGHT))	4.28	83.45	5.46	78.75	6.34	71.27	5.86	73.49	b	9.53	66.72 d
SURE (RIGHT)	55.37	351	53.86	398	52.94	503	50.98	423	e	41.78	423 g
(% OF TOTAL (RIGHT))	22.79	12.79	16.30	16.20	14.55	21.44	12.44	20.10	f	16.84	23.38 h
GUESS (RIGHT)	41.34	103	32.22	124	40.65	171	39.34	135	i	25.74	179 k
(% OF TOTAL (RIGHT))	32.92	3.75	27.32	5.05	29.77	7.29	25.52	6.41	j	20.39	9.89 l
TOTAL (RIGHT)	81.00	2744	74.48	2457	70.12	2346	65.40	2105	m	56.00	1809 o
(% OF TOTAL ANSWERED)	3.81	80.90	2.96	74.36	3.09	69.99	3.24	65.43	n	4.66	55.85 p
V. SURE (WRONG)	12.01	327	16.87	426	18.43	420	25.25	555	q	30.67	598 s
(% OF TOTAL (WRONG))	4.28	50.46	5.46	50.30	6.34	41.75	5.86	49.91	r	9.53	41.82 t
SURE (WRONG)	39.07	221	46.14	267	47.06	373	46.25	382	u	50.11	510 w
(% OF TOTAL (WRONG))	20.65	34.10	16.30	31.52	16.55	37.08	11.89	34.35	v	18.79	35.66 x
GUESS (WRONG)	41.99	100	37.22	154	43.14	213	49.54	175	y	63.45	322 aa
(% OF TOTAL (WRONG))	33.09	15.43	30.06	18.18	30.46	21.17	27.71	15.74	z	28.89	22.52 bb
TOTAL (WRONG)	19.00	648	25.52	847	29.88	1006	34.40	1112	cc	44.00	1430 ee
(% OF TOTAL ANSWERED)	3.81	19.10	2.96	25.64	3.09	30.01	3.24	34.57	dd	4.66	44.15 ff
V. SURE (TOTAL)	77.04	2617	71.79	2361	62.47	2092	65.25	2102	gg	55.75	1805 ii
(% OF TOTAL ANSWERED)	13.07	77.15	18.21	71.46	17.62	62.41	16.22	65.34	hh	21.85	55.73 jj
SURE (TOTAL)	16.93	572	19.86	665	26.01	876	25.15	805	kk	28.58	933 mm
(% OF TOTAL ANSWERED)	11.55	16.86	12.95	20.13	13.07	26.13	11.66	25.02	ll	15.67	28.81 nn
GUESS (TOTAL)	6.02	203	8.35	278	11.51	384	9.61	310	oo	15.67	501 qq
(% OF TOTAL ANSWERED)	5.49	5.98	9.13	8.41	10.01	11.46	7.72	9.64	pp	13.95	15.47 rr
NO GUESS (TOTAL)	6.62	208	9.66	296	11.13	346	12.87	383	ss	16.11	461 uu
(% OF TOTAL ANSWERED)	7.76	6.13	9.17	8.96	9.56	10.38	11.11	11.91	tt	16.38	19.23 vv
TOTAL ANSWERED	94.22	3392	91.78	3304	90.59	3352	89.36	3217	ww	67.54	3239 yy
(% OF TOTAL QUESTIONS)	6.15	94.22	7.22	91.78	7.36	90.59	8.02	89.36	xx	10.48	87.54 zz
SCORE	59.57		46.53		37.90		28.70		ab	14.38	
(AS A %)	6.79		3.07		2.89		2.89		ac	5.97	
NUMBER OF STUDENTS	36		36		37		36		ad	37	

NUMBER	GUESS	RIGHT	SURE	RIGHT	V. SURE	RIGHT	NO GUESS
a	b	c	d	e	f	g	h
76	9.89	55.56	28.57	88.46	52.75	98.98	8.79
77	18.68	35.29	26.37	47.92	28.02	45.10	26.92
78	2.20	25.00	8.79	75.00	84.07	80.39	4.95
79	7.69	64.29	18.68	67.65	70.33	67.19	3.30
80	2.20	100.00	7.14	100.00	89.56	94.48	1.10
81	2.20	50.00	8.79	75.00	87.91	93.75	1.10
82	8.24	60.00	13.74	76.00	71.43	90.77	6.59
83	15.38	67.86	24.73	66.67	47.25	73.26	12.64
84	16.48	63.33	13.19	66.67	58.24	92.45	12.09
85	8.24	73.33	8.79	75.00	77.47	100.00	5.49
86	22.53	41.46	23.63	58.14	28.02	70.59	25.82
87	1.65	66.67	8.24	80.00	88.46	97.52	1.65
88	3.30	66.67	8.24	66.67	86.26	93.63	2.20
89	1.10	50.00	7.14	92.31	90.66	96.97	1.10
90	3.85	85.71	17.03	83.87	74.73	96.32	4.40
91	3.85	57.14	14.29	88.46	79.12	95.14	2.75
92	19.78	22.22	23.63	18.60	22.53	21.95	34.07
93	20.33	21.62	30.22	12.73	31.32	12.28	18.13
94	17.03	93.55	25.27	97.83	43.41	100.00	14.29
95	22.53	70.73	23.08	85.71	17.58	90.63	36.81
96	18.68	44.12	19.78	63.89	32.42	79.66	29.12
97	17.58	46.88	18.68	73.53	10.44	63.16	53.30
98	19.23	40.00	24.73	35.56	28.57	53.85	27.47
99	14.29	50.00	28.57	80.77	29.12	84.91	28.02
100	12.09	18.18	21.98	17.50	51.65	7.45	14.29
k → MEAN	9.21	45.00	21.16	58.51	60.31	74.79	9.32

ANALYSIS OF CONFIDENCE MCQ 180675

	V. SURE	SURE	GUESS
MARKS AWARDED	3.00	2.00	1.00

NEW MEAN	NEW S.D.	OLD MEAN	OLD S.D.
39.30	17.00	56.80	13.60

MEAN CLASS CONFIDENCE SCORE 232.47

MEAN CLASS CONFIDENCE INDEX 2.56

	QUINTILE 1	QUINTILE 2	QUINTILE 3	QUINTILE 4	QUINTILE 5
MEAN PERCENTAGE SCORE	59.57	46.53	37.90	28.70	19.38
MEAN CONFIDENCE SCORE	255.50	241.42	227.35	220.50	210.32
MEAN CONFIDENCE INDEX	2.71	2.63	2.51	2.56	2.90

ANALYSIS OF CONFIDENCE MCQ 180675

STUDENT NUMBERS	TTL R	TTL W	TTL VS	TTL S	TTL G.	TTL NG	CNF SCORE	CNF INDEX	CLS CNF RATIO	QNT CNF RATIO	NEW % SCORE	OLD % SCORE	DIF % SCORE	NEW QNT	NEW RNK	OLD RNK	DIF RNK	NEW Z SCR	OLD Z SCR	DIF Z SCR
75001	68	24	76	9	7	8	253.00	2.75	1.07	1.05	42.19	63.08	20.89	2	71	64	-7	.17	.46	-.29
75002	71	29	66	28	6	0	260.00	2.60	1.01	.99	42.42	61.51	19.09	2	70	72	2	.18	.35	-.16
75003	62	23	62	15	8	15	224.00	2.64	1.03	1.05	39.14	56.17	17.03	3	88	100	12	-.01	-.05	.04
75004	55	29	39	26	19	16	188.00	2.24	.87	.88	27.42	49.48	22.06	4	132	130	-2	-.70	-.54	-.16
75005	62	36	46	35	17	2	225.00	2.30	.90	.90	27.97	50.31	22.34	4	130	122	-8	-.67	-.48	-.19
75006	57	28	55	20	10	15	215.00	2.53	.99	.99	26.33	48.64	22.31	4	141	134	-7	-.76	-.60	-.16
75007	59	34	81	12	0	7	267.00	2.87	1.12	1.12	31.25	55.44	24.19	4	118	102	-16	-.47	-.10	-.37
75008	43	44	28	30	29	13	173.00	1.99	.78	.83	13.05	30.86	17.81	5	169	175	6	-1.54	-1.91	.36
75009	78	22	89	11	0	0	289.00	2.89	1.13	1.07	55.08	73.33	18.25	1	25	19	-6	.93	1.24	-.29
75010	71	12	55	25	3	17	218.00	2.63	1.02	.97	56.95	71.23	14.28	1	19	30	11	1.09	1.06	-.02
75011	60	36	82	6	8	4	266.00	2.77	1.08	1.08	26.64	50.21	23.57	4	138	129	-14	-.74	-.48	-.26
75012	53	41	58	28	8	6	238.00	2.53	.99	1.05	17.19	40.17	22.98	5	162	162	0	-1.30	-1.22	-.08
75014	80	15	61	11	23	5	228.00	2.40	.94	.89	65.86	80.23	14.37	1	4	4	0	1.56	1.72	-.16
75015	84	11	82	4	9	5	263.00	2.77	1.08	1.02	76.48	84.52	8.04	1	2	3	1	2.19	2.04	.15
75016	72	21	50	30	13	7	223.00	2.40	.94	.88	52.27	67.36	15.09	1	35	44	9	.76	.78	-.01
75017	67	29	45	24	22	9	205.00	2.25	.88	.86	45.31	63.39	18.08	2	55	62	7	.35	.48	-.13
75018	62	22	44	21	19	16	193.00	2.30	.90	.92	41.09	59.21	18.12	3	77	83	6	.11	.18	-.07
75019	63	29	73	18	1	8	256.00	2.78	1.09	1.09	31.33	57.64	26.31	4	117	93	-24	-.47	.06	-.53
75020	49	44	67	18	8	7	245.00	2.63	1.03	1.10	8.28	37.03	28.75	5	177	168	-9	-1.82	-1.45	-.37
75021	66	30	79	12	5	4	266.00	2.77	1.08	1.08	32.50	54.92	22.42	4	111	103	-8	-.40	-.14	-.26
75022	85	15	98	2	0	0	298.00	2.98	1.16	1.10	72.03	85.04	13.01	1	3	2	-1	1.93	2.08	-.15
75023	68	32	43	46	11	0	232.00	2.32	.90	.92	40.94	62.13	21.19	3	79	69	-10	.10	.39	-.30
75024	70	18	76	9	3	12	249.00	2.83	1.10	1.08	51.95	66.53	14.58	2	37	47	10	.74	.72	.03
75025	66	17	69	9	5	17	230.00	2.77	1.08	1.05	51.72	66.00	14.28	2	38	49	11	.73	.68	.05
75026	66	34	32	36	32	0	200.00	2.00	.78	.76	42.03	60.56	18.53	2	72	76	4	.16	.28	-.12
a	b	c	d	e	f	g	h	i	j	l	m	n	o	p	q	r	s	t	u	v