The Reliability of Examinations

by A. E. H. BLEKSLEY

EXAMINATIONS are a type of measurement, and thus subject to the same principles and laws as any other measurement in science. It might be useful some time to lay down these principles in general terms, and then to apply them to the problem of examinations as a special case, but this is hardly the occasion for so ambitious an undertaking . I propose, therefore, to limit my discussion in the main to the problem of the reliability of examinations, as demonstrated by a number of investigations carried out, with the unstinting assistance of the Transvaal Education Department over a number of years. It is a pleasure to acknowledge my indebtedness to the authorities for their co-operation at all stages of these investigations.

The first questions to be asked of any examination are, briefly, "What?" and "Why?" What are we attempting to measure, and for what purpose? Quite often, one must admit, the answer to neither of these is very clear; in too many cases we are not by any means sure what it is that a particular test or examination actually measures, even if we are reasonably clear about what we would like to measure.

Applying these questions to most examinations with which we have to deal in school, we may say that in general the "Why?" falls into one of two main categories:

(a) To measure specific knowledge, as in a dictation test, a matriculation Latin paper, etc.
(b) To measure skill, as in the practical examination in manual training, the essay in a language examination, etc.

In many cases the objective is a combination of both, as in the geometry paper.

As to the "What?" the main reasons for examinations could be enumerated under the general headings of diagnosis and prognosis. Thus diagnosis of shortcomings in the taught (or in the teaching) by means of class tests carried out from time to time in order to determine whether the class as a whole and the pupils individually have mastered the work, and to decide whether the pupil meets the basic requirements for promotion to a higher standard.

The prognostic aim in examinations is that of predicting future success in another way of life. Thus the matriculation as a university entrance examination is an attempt to decide whether an individual is likely to benefit by higher study. In this latter connection, as everyone knows, the matriculation examination has not, in fact, proved to have as high a prognostic value as we would like, since roughly 40 percent of all First Year students at our universities fail, and perhaps 40 percent of all university entrants succeed in graduating.

Even in the restricted Faculties, such as Medicine and Engineering, where additional selection takes place, far too many students fail for our comfort. So far the position has not been given the attention it deserves. Its importance is such that one feels that every effort should be made to find whether any really adequate predictors of future university success can be found in the entrance examination.

So far it seems that the best predictor we have is the aggregate mark in the Matriculation examination, and even this seems to work best in a negative kind of way. Thus the matriculant with a third class aggregate has twice as high a probability of failing First Year Engineering as one with a second class, and four times as high as a student coming to the study of engineering with a first class aggregate in the Matric. But even first class matriculants fail far too often. We are, in fact, confronted here with a problem of major importance, on which far too little is being done.

It is not, however, my intention on this occasion to carry the discussion on this point any further. I am here concerned with still another question which one is entitled to ask of any examination, namely "How reliably does it measure what it is intended to measure?" This question can be extended somewhat when we are dealing with an examination such as the Matriculation in which there are often several papers in a given subject, and also different examiners even in the same paper. Two important questions to be asked of different parts of the same examination, or of two examiners in the same subject, are:

Do they measure the same thing? This means, do high scores in one paper tend to go with high scores in the other, for example. If so, we can determine this fact by means of the statistical

measure known as the coefficient of correlation. Thus two examinations which measure much the same thing will give scores which are highly correlated.

The second question is: do they measure it in the same way? In this case, not only will the coefficient of correlation between the two sets of marks for the same candidates be high, but also the mean and the scatter (as measured statistically by what is known as the standard deviation) will be nearly the same in both cases. Thus, if one states a set of what have come to be called "vital statistics" for a modern film star as 100-60-90, we need not be shocked, since these statistics would closely represent Miss Munroe in the c.g.s system. In other words, sets of vital statistics measurements carried out in inches and in centimetres as units would be strongly correlated: they measure the same thing. But the means and standard deivations of the two sets of measures would be very different: they measure it in different ways. It is clear that the requirement of high correlation is by far the more important; the other only arises when this requirement is satisfied.

* * * * * *

Having laid down, then, the criteria which we shall apply to examinations to decide on their reliability, let us consider a particular case. We wish. say, to test the capacity of a pupil to use his home language. To do so, we can use any one of a considerable variety of possible methods. We might, for example, set him an oral examination. in the course of which we assess his capacity to express his ideas through the spoken word. Or we might set him the task of writing an essay, in which he is assessed on his capacity to express his ideas through the written word. Or we might set him a comprehension test, or a series of questions on the machinery of the language. Or — and this is what we do in general practice — we set a comprehensive examination which contains all of these different kinds of test as component parts; we then add the various marks together and so obtain a final mark on which to base our final judgement. Having taken by this time a great deal of trouble in reaching this final assessment, we probably feel that we now have a reasonably high correlation between the various parts of the language examination. Experiment, however, shows that this is by no means the case. Investigations undertaken on the coefficient of correlation between the various parts of the Secondary School Examination papers in the two Higher languages are summarised in *Table 1*.

TABLE 1.

Coefficient of Correlation between the various parts of the Higher Language papers, T.S.S.C.

ENGLISH

	Language	*Comprehension*
Essay	0.69	0.25
Language		0.85

AFRIKAANS

	Language	*Comprehension*
Essay	0.34	0.28
Language		0.57

To give an idea of the significance of the correlation coefficients quoted, the correlation obtained between the marks obtained in the examination by the same candidates in certain other pairs of subjects are given in *Table 2*.

TABLE 2.

Coefficients of Correlation between marks in various other subjects.

	Maths Paper 2 (Geometry)	History	Science	Latin
Maths Paper 1 (Algebra)	0.64	0.40	0.59	0.53
Maths Paper 2		0.31	0.63	0.36
History			0.51	0.53
Science				0.58

A comparison of Tables 1 and 2 shows that, on the whole, the component parts of the Language examinations are not more closely related than the marks in entirely different subjects, as, for example, Science and Latin. Indeed, the observed correlation of about 0.6 between these two subjects shows that these two examinations must have a considerable common content; since this is clearly not the subject matter, it seems that the common element is general intelligence. This conclusion has been verified by further studies.

In particular, then, the low correlation found in both languages between Essay and Comprehension suggests that these two parts of the examination have very little in common, and what we have just said about the common element between Science and Latin suggests that it might be worth assessing the role of I.Q. in this case as well. The correlations between Essay and Comprehension on the one hand and the Intelligence Quotient on the other are given in *Table 3*.

TABLE 3.

Coefficient of Correlation between Language Marks and the I.Q.

	English Essay	*English Comprehension*
Intelligence Quotient	0.35	0.52

This Table shows that both Essay and Comprehension marks are correlated with general intelligence as measured by I.Q.. They therefore have this factor in common. By well-known statistical techniques it is possible to eliminate the common element due to I.Q. If these are applied, we find that for children with the same I.Q., the correlation between Essay and Comprehension drops to 0.08, which is quite insignificant. In other words, we are forced to conclude that the Essay and Comprehension tests have no common element that we can ascribe to the capacity to use the mother tongue! If either of these tests does in fact assess the pupil's capacity to use the mother tongue (which we have not yet established anyway), one thing is abundantly clear, namely that the other then certainly does not.

I have here referred specifically to the case of the parts of the Higher language examination, but I am by no means suggesting that this situation is unique to these subjects. If there should be anyone who is convinced that the two Mathematics papers measure some indefinable but real quality which one might call "mathematical ability", we need only refer once again to Table 2, which shows that the correlation between the Algebra and Geometry papers is 0.64, which is not significantly higher than that between Science and Latin, namely 0.58. If we were to eliminate the effect of the known correlation between the marks in the two Mathematics papers and the I.Q., we would again be left with a relatively low correlation, which once again suggests that in these two papers we are in fact measuring different things.

We must therefore face the fact that our examinations cannot be assumed without careful scrutiny to measure what we intend and believe them to measure. In other words, our examinations do not necessarily possess high *validity*. It seems to me to be of the highest importance to determine, by careful statistical analysis, just what our examinations do in fact measure. A study of this kind might very well have revolutionary results.

Having concluded that we do not always know what our examinations measure, let us next try to answer the question of their reliability: how accurately do we manage to measure whatever it is that we are measuring. Here we find that there are, in general, two major sources of error.

The first arises from the fact that most examinations involve an element of subjective judgement on the part of the examiner. This element is probably most pronounced in the case of the marking of essays, but it certainly exists even in the case of mathematics. In this case, once the examiner had drawn up his marking memorandum, the application of the memorandum is relatively objective. But the original allocation of marks in the memorandum involves considerable subjectivity of judgement, as the well-known investigation of Hartog and Rhodes has shown.

In the Transvaal Board of Moderators our investigations were, however, largely confined to the marking of essays. The investigations were lengthy, and I wish to do no more here than to refer to the results in broad outline. The important facts that emerged were the following:—

(i) The marks of an individual examiner are not highly reliable over long periods. Thus in one investigation, a team of experienced examiners were set the task of marking for a second time, some months after the first occasion, 300 scripts in English (and also 300 in Afrikaans) essays which had been written and marked during the T.S.S.C. examination. The coefficient of correlation between the marks awarded by the same examiners to the same scripts on two occasions was found to be about 0.6 — thus no greater than that existing between the marks for Science and Latin.

(ii) The marks of two individual examiners, working together in order to establish a common standard, and obtained independently for a large number of Matriculation essays were found to correlate to the extent of a coefficient 0.6. In other words, two examiners working together vary by about as much as one examiner does over a period of time.

(iii) The candidates themselves are not able to reproduce their own form with much precision over a period of time. This investigation appeared to lead to a rather new conclusion, and may be worth reporting in some detail. A group of about 300 Standard X pupils in High Schools in Johannesburg were set an essay on a general topic, and these essays were marked independently by two examiners A and B. A month later an essay on a similar topic was again set to the

same pupils, and these were marked by the same examiners. The two sets of marks were then correlated in various ways; in the following results I have taken the mean of the results for English and Afrikaans, and where applicable the mean of the marks awarded to the same essay by the two examiners.

1. Essay I: Average correlation between marks of two independent examiners: 0.6.
2. Average correlation between marks of either examiner on essays 1 and 2: 0.4.
3. Mark of Examiner A for Essay 1 with mark of Examiner B for Essay 2: 0.4.
4. Average mark of both examiners for Essay 1 with average mark of both examiners for Essay 2: 0.55.

The really striking results are those given under 2 and 3 which show the effect of the combined variability of pupil and examiner. If we eliminate statistically the effect of the lack of consistency of the Examiner, we find the correlation of the candidate with himself to be about 0.6, which means that the examiner and the candidate are about equally inconsistent. It appears, then, that in the marking of essays, at any rate, we are attempting the educational equivalent of measuring the distance between the tip of the nose and the tip of the tail of a live and wriggling eel by using a piece of elastic string graduated by eye.

We have shown, then, that for various reasons there is an inherent lack of reliability in the writing and marking of examinations of the standard type. Since we cannot hope to eliminate examinations, however, it seems clear that we should try and improve their reliability.

In the first place there can be little doubt that we should make more use of the so-called *objective type* tests which have been shown to possess much greater internal reliability than the subjective type test in common use at present. These tests, when well drawn up, are not only internally reliable, but also consistent, i.e. they measure the same thing.

The other point that I would like to stress is that no single test, however well designed, can eliminate the factor of the candidate's own variability. Hence we need to use, not one test, nor one single final examination, but a whole series of tests, to obtain by an average over the entire series, a more reliable estimate of the candidate's true capacity. And since this cannot, obviously, be done in any reasonable final examination, it follows, or so it appears to me, that we must make far more use of the teacher's assessment — of the class record which the pupil can establish over a period of two years or so. This suggestion is based on the now well-established fact that the competent teacher is better fitted than anyone else to rank the pupils in order of merit. What the teacher may well find difficult is to assess the correct numerical mark to assign to each pupil. What is neededd, therefore, is to provide the teacher with material by means of which he can establish absolute standards, such as a battery of standardised objective-type tests. These, taken in conjunction with the means and standard deviations of the entire class, would enable the teacher's raw scores to be interpreted in terms of reliable marks, thus providing, one hopes, a more reliable assessment of the true potential and knowledge of the pupil as an individual than any we possess at present.