A NOTE ON THE VALIDITY OF A MEDICAL COLLEGE SELECTION PROGRAMME*

By NAOMI H. CARMAN and JOY C. PAUL
Christian Medical College, Vellore
and
A. EDWIN HARPER, Jr., B. DAS GUPTA and S. P. SANGAL
Indian Statistical Institute

SUMMARY. In 1946 the Christian Medical College of Vellore (South India), introduced new and more scientific methods for selection of students. The new methods have resulted in a substantial reduction in the number of papers failed per student and years repeated. This paper describes the selection programme very briefly giving various types of evidence on the value of these scientific selection methods and concludes with a technical note on the correction of validity coefficients for the differential effects of restriction of range. This is a first report on full-scale validity studies which are in progress.

1. Brief description

In 1946 a special testing programme for the selection of medical students was introduced in the Christian Medical College, Vellore. It was organized by Dr. Frank Lake, on the basis of his experience in the officer selection methods used in the British Army. The selection method used is a two-stage process. The first stage consists of a series of objective ("new-type") tests of knowledge and ability. The second stage includes various types of tests which try to get at the student's personality, level of motivation and drive, interests, emotional stability, etc.

At the first stage, examinations are administered all over India, and even in Pakistan, Burma, and Ceylon. Questions such as the following are used :

---

*The final form of this report represents a collaboration between members of the staff of the Christian Medical College, Vellore, and members of the staff of the Psychometric Research and Service Unit of the Research and Training School, Indian Statistical Institute, Calcutta. Its antecedents should be acknowledged: The study was originally begun in 1951 by Mrs. Carman (who had been active in the selection programme from its beginning in 1946) and Dr. Harper (then at Ewing Christian College, Allahabad). It was based partly on previous work done by Dr. Frank Lake, Mrs. Carman, and others. The first full report of this work was prepared by Mrs. Carman in 1953, for private circulation, but only a brief non-technical summary was published. The current report draws heavily on this original manuscript, which remained unpublished; but much of the original material has been left out, while new data have been added in certain areas. A first draft was cyclostyled in December 1955 for private circulation, but contained certain inadequacies and inaccuracies. This is the revised and corrected version, to which the major contributions were as follows: Mrs. Carman and Dr. Paul, assisted by Shri P. S. Sunder Rao gathered, sorted, and organized the data, and advised on interpretation. The analysis and reporting were the responsibility of Dr. Harper, with both theoretical and computational assistance from Shri B. Das Gupta, Mr. A. Kudo, Shri S. P. Sangal, and Shri Tapas Kumar Sen. The opinions expressed, however, are those of the authors and do not necessarily reflect the official policy of the College or of the Institute.

Questions. During photosynthesis :

(1) Oxygen and simple sugar combine to form carbon dioxide and water.
(2) Simple sugar oxidizes and releases energy.
(3) Water and carbon dioxide combine to form fats and proteins.
(4) Water and chlorophyl combine to form simple sugar.
(5) Water and carbon dioxide combine to form simple sugar and oxygen.

The student is asked to select the one correct answer from the alternatives listed, and to indicate his selection on a special separate answer sheet. About 250 questions on Physics, Chemistry, Biology, and General Knowledge are answered in about three hours testing time. There is also a General Ability ("intelligence") test, and tests of English Comprehension and Composition. The entire series is completed in one day.* There have been some changes in the specific tests from year to year, but the areas covered have remained substantially the same.

Out of the preliminary applicants who have taken the objective tests, a small group of the most promising are invited to Vellore, for the second stage of the selection programme. The candidates are split up into groups of ten, and a member of the medical college staff is assigned to each group as its "Group Observer." During the next two days each group is put through a series of tests designed to elicit a wide sampling of actual behaviour, under various types of situations. Sometimes the group is assigned a problem-task which can only be solved through active group cooperation. In other tests the individual must perform alone. There are also interviews, physical examinations, and time out for recreation. At the end of this time, all those who have observed candidates in one or more of the various tests, meet together to determine a final over-all rating ("Final Board Grade") for each candidate. Final selections, based on this over-all rating, are made by a Selection Committee. The Selection Committee must, of course, also take into account such factors as all-India representation and the filling of reserved seats.

2. Comparison of old and new selection methods

The question arises, "How good are the tests?" Studies have been made from time to time since the programme began, and are still being made on this question. Here are some of the results.

"Is the new method any better than the one it replaces?" A pilot study, completed in 1952, compared two groups of approximately a hundred students each. The first three classes admitted under the new selection method totalled 98 students. These were compared with the last groups of students admitted under the old methods,

*Objective tests are known to be a more accurate and efficient means of testing knowledge, understanding, thinking, etc., than the traditional type of examination. A further advantage is ease of scoring, even when done by hand. The tests are also adapted to scoring by punch-card machines, with considerable gains in speed, accuracy, and economy. For example, recently 5,400 such tests were marked in a bit over a week, at a cost of less than three cents each.

four classes which totalled 101 students. (The "old methods" were the usual ones of Intermediate marks, written recommendations, etc.). In order to have as large N's as possible to work with, the 1948 class was included, although it was still one year short of graduation (1953). For this reason the first comparison was based on only the first four years for *all* of the 199 cases in the pilot study. The data are presented in Table 1.

TABLE 1. STATUS OF STUDENTS AT THE END OF FOURTH YEAR

	selection method : admission years	
	old :1942–1945	new: 1946–1948
total number of candidates admitted	101	98
left for marriage, health, or transfer to another institution	7	2
number for whom adequate academic record is available	94	96
discontinued for failure	8	2
number available for study in 1952, and on whom the following figures are based	86	94
data on lost years *		
number on schedule at end of fourth year	43	54
number one-half year late	22	29
number one or more years late	21	11
total number of years lost by failure	39.5	27.5
years lost per student	.46	.29
total number of papers failed	130	73
papers failed per student	1.51	.78

*Time lost because of illness was not counted in computing years lost.

As assessed at the end of the Fourth Year of medical college, the introduction of the new selection methods had resulted in a reduction of approximately 40 percent in "lost years", and 50 percent in the number of papers failed, per student. The difference in the number "discontinued for failure" reflects a change in Madras University policy. All of the "old" group, and some of the "new" had to undergo a pre-registration course, and those who failed twice were not admitted to medical studies. This course was discontinued after 1947. Apparently this six-month pre-registration course was a very efficient "selection test" and greatly reduced later failures particularly those in the Fifth Year. It is for this reason that the Fourth-Year results, are presented, for comparison with the remaining Tables in this section of the paper.

Tables 2, 3, 4, and 5 are based on the full five year medical course, for which the analysis is presented in more detail. Every student included had either completed his or her medical course, or had been discontinued for failure, by the time this paper was written.

One of the bases for comparison of old and new selection methods was the number of *extra* years (beyond the minimum possible of five years) that the students took to complete their course. Table 2 presents these data.

TABLE 2. YEARS LOST BY ALL STUDENTS, UNDER OLD (1942–1945) AND NEW (1946–1948) SELECTION METHODS

		cost to the Institution	
method of selection	number of students	total years lost	years lost per student
old	86[1]	72	.84
new	94[2]	58	.62

[1] Out of 101 admitted 7 left for marriage, health, or transfer to another institution; 8 discontinued for failure in pre-registration course. Time lost for illness not counted.

[2] Out of 98 admitted 2 discontinued for failure (not in pre-registration, but much later in the medical course), and 2 left for marriage or transfer. Years lost for illness not counted.

The criterion here is a very practical one. Extra years spent in the medical college cost the student heavily in both time and money. They also cost the institution a great deal. Table 2 shows that 94 students chosen by the new methods lost (i.e. had to repeat, because of failure) an average of .62 years apiece, while the previous group lost an average of .84 years per student.

If we drop the decimal point in the last column of Table 2, (i.e. multiply "per student" rates by 100), we find that the net gain per hundred students is 22 years, i.e., two years more than the time it takes to give a 5-year medical education to each of four students. In other words, *with a given amount of facilities and finance*, a little better than *four more students per hundred* are being given a medical education since these modern selection methods were substituted for the older ones. (It should also be noted that this is probably an underestimate of the actual savings, as the students "discontinued for failure" have not been included in Table 2. This point is discussed in more detail later in this paper).

A second criterion for the success of a selection programme is its reduction in the number of failures. There were various reasons why we did not take examination marks as a criterion; one was that these marks often represent the student's second or third "try." A better standard seemed to be the total number of failures per individual. A paper failed once was counted once, but a paper failed repeatedly was counted each time it was failed. Table 3 shows that the new selection methods *cut down failures by more than one quarter*.

TABLE 3. NUMBER OF PAPERS FAILED BY STUDENTS SELECTED UNDER OLD AND NEW METHODS

method	number of students	total number failed	failures per student
old	86*	220	2.86
new	94*	181	1.93

* See footnotes to Table 2.

It was necessary to exclude those "discontinued for failure" from Tables 2 and 3, in order to compare the groups more accurately. This is because the eight failures in the "old" group were all dropped for twice failing in the pre-registration course, and so were given no opportunity to lose years or fail in papers. If they passed pre-registration, however, they were allowed to continue almost indefinitely. In the "new" group, on the other hand, the barrier came much later, and also was more lenient. Only those who failed in the First M. B. examinations three times were dropped. Thus the difference in total number of years failed is partially a function of changes in University and College policy, as well as of changes in selection methods.

It does not seem unreasonable to suppose that, had the eight "old method" selectees not been eliminated by the pre-registration course barrier, their failures by the end of the Fifth Year would have been much higher than the average of their classmates, thus increasing the rates "per student". Thus a comparison of Tables 2 and 3 with Table 1 suggests that the latter are conservative "minimum estimates." The early elimination of students "discontinued for failure" has probably reduced the number of failures under "old method" far more than under "new method." (It may also be that "left for marriage, health, or transfer" is an index of the efficiency of selection—although, since this is difficult to prove, its effect has been discounted in all of the tables). Thus it seems safe to conclude that the "true" figures lie somewhere between $\frac{1}{4}$ and $\frac{3}{8}$ reduction in the number of years lost per candidate; and between 27 percent and 50 percent reduction in the number of papers failed.

There is one method by which we can by-pass this problem of the changes in University policy, although the results will not be as directly interpretable in economic terms as was in Table 2. Instead of attempting to add up the exact number of years lost for failure, we can classify the students in several groups. The changes in University policy discussed above do not affect those who complete their course on time, or are delayed, say, half a year. They affect only those who are delayed at least one year. A very strict policy drops students after one year lost in failure; a very lenient policy allows them to continue even as much as five years. So we can reasonably classify the students dropped for failure with the group who completed the course one or more years late, and thus be able to make use of the data which had to be discarded for Tables 2 and 3. Table 4 gives the results of such an analysis.

TABLE 4. EFFICIENCY OF SELECTION METHOD WHEN EFFECT OF CHANGE OF UNIVERSITY POLICY ON FAILURES IS DISCOUNTED

method of selection	number of students	per cent finished less than 1 year late	per cent finished one or more years late or discontinued for failure
old	94	55	45
new	96	70	30

In the group on the right of Table 4 are those who contributed to "waste"* of educational facilities. Those who eventually graduated, contributed to "waste" by using more than the minimum of five years to complete their medical course. Those who were discontinued for failure contributed to "waste" in two ways: the number of years that they spent in the college, and the empty seats that they left. If these places remain unfilled (as they usually must) then they represent facilities that *could* have been used for educating doctors, had the right candidates been selected. Table 4 shows that the percentage of students belonging to this group who contribute seriously to "waste" of facilities has been reduced by one-third, by the introduction of the new selection methods. The percentage of students in the "successful" group, who contributed little or nothing to "waste", increased by about one quarter.

Table 5 gives a more detailed breakdown of Table 4 and serves two functions. Firstly, the question is frequently asked, "Could the differences between the 1942-1945 group and the 1946-1948 group be due to a change in the quality of the candidates, from whom the selections were made?" There is no direct answer to this. However, accepting Intermediate Division as at least a rough indicator of "quality", and assuming that each selected class represents the best of the candidates, we can give a fairly confident answer. From Table 5 we read that 36 First Divisioners were selected for the 1942-1945 classes, and 39 for the 1946-1948 classes. The number of Second and Third Divisioners was 58 and 57 respectively. (Many of the universities concerned do not award Third Divisions, so the number of Thirds is too small to warrant separate consideration). Thus the difference between the two groups is due to qualities other than those measured by the I.Sc. examinations.

*We are indebted to Dr. E. Douglass Burdick for the help in clarifying this concept of "waste". We hasten to add that we recognize that this term cannot be used in any absolute or dogmatic fashion. The partial medical education of a student discontinued for failure may not be a total loss to society. And occasionally a student who has repeated several years may, by his later career, justify the extra expense.

TABLE 5. RELATIVE EFFICIENCY OF SELECTION METHODS FOR CANDIDATES WITH DIFFERENT MARKS IN I.SC. EXAMINATION

division	method of selection	number	percent finishing		
			on time	1 year late	1+ or fail*
I	old	36	43	36	21
	new	39	49	36	15
II & III	old	58	15	26	59
	new	57	30	30	40

*The third column includes those who finished the course one or more years late, and also those who were discontinued because of failure.

The second function of Table 5 is to point out one of the reasons for the success of the new selection programme. The selection criterion which is probably most widely used by professional colleges, as well as various employers, is the division secured in Intermediate Examinations. This is a valuable criterion, as can be seen by contrasting the two groups selected by the "old" method (i.e. with heavy reliance placed on the marks in I.Sc. Examination). For example, only about one-third as many Seconds as Firsts were able to complete their course in the minimum five years (15% vs. 43%), and nearly three times as many (59% vs. 21%) were seriously detained. (Although it is also notable that one-fifth of the Firsts (21%) did quite poorly, when selection was based mainly on that criterion). The new selection programme, in which Intermediate Division played a minor role, selected slightly better First Divisioners than previously—although this difference is not statistically significant, and so we cannot generalize from it. The main contribution of the new selection methods, however, is in the improvement in the quality of Second and Third Divisioners admitted. As many have suspected, the *lack* of a First Division mark is *not* necessarily proof of a lack of ability. Able students frequently receive lower marks, due to the unreliability of traditional examinations,* ill health, or other irrelevant reasons. The objective tests, with their higher reliability and validity, help to select the more able of the Second and Third Divisioners. The proportion of this group completing the course on time has been doubled (15% to 30%). The proportion seriously detained has been reduced by nearly one-third (59% to 40%). When we consider the fact that the total testing time in both stages of the new selection programme is less than half the number of hours of examination on which the I.Sc. Division is based—and that this shorter testing time assesses not only knowledge but intelligence, special aptitudes, and various aspects of personality as well—the contrast in efficiency is worth noting.

*See, especially, *The Report of the University Education Commission*, Dec. 1948–Aug. 1949, Vol. I, Chapter X, pp. 328, 336.

3. OVER-ALL VALIDITY OF THE TWO-STAGE SELECTION METHODS

In section 1 we described the two-stage selection process: (1) a preliminary selection of a group of promising candidates is made on the basis of objective tests administered to all qualified applicants and (2) this smaller "group of promising candidates" is called to Vellore for intensive assessment, and final rating. It should be obvious that the second stage is far more difficult to organize and administer, and far more expensive (in terms of requiring three full days of time from top-level medical staff) than the first stage. Two questions arise: (1) "Is it worthwhile?" and (2) "Is it necessary?"

The answer to these questions should be based on empirical evidence; even though many other factors must also be taken into account. But even in quantifying empirical evidence, difficult questions arise. Perhaps the hardest to solve is the question, "What should be our criterion, against which we will judge the success of the tests?" Ideally, we would like to find out whether our tests predict whether or not a candidate will become a good doctor. This is not quite the same as predicting whether or not he will do well in medical college. Experience shows that there is by no means a one to one relationship between medical college marks and how successful a doctor the person becomes. But how can we find out how good a doctor the ex-student has become? How shall we define "success"? How many years do we have to wait to know whether he is "successful" or not? These problems seem almost insurmountable. So we did the next best thing; we asked the teachers who knew each candidate best to *rate* him on how good a potential doctor they thought he was. (This rating was done in 1952, at the end of two to five years of the medical course of each student). These ratings took into account ability and achievement, as well as personality, motivation, and drive. Since each student was rated by several doctors who knew him well, these ratings represent a comprehensive judgement on each student. They are, at very least, an interesting criterion against which to test our tests.

Table 6 summarizes the results. Complete data were available on four different classes (1947-1950) who have finished (or should have finished) their medical course. These have been analyzed separately for each year, and then combined. (See Appendix A). The coefficients of correlation in Table 6 represent, therefore, the relationship between selection methods and two different criteria for a total of 130 students. The two criteria used were (1) whether or not the student finished his medical course within five years ("On Time") and (2) his average rating as a "Potential Doctor."

There are two rather striking facts that emerge from Table 6. One is that the methods used at both stages of selection predict the candidates' rating (given two to five years later) as a "potential doctor" better than they do his ability to complete his medical education in five years (+.55 and +.56 vs.+.48 and +.36). What does this mean? The possibility that the ratings were influenced by the original marks is

TABLE 6 RELATIVE AND COMBINED VALIDITIES OF FIRST STAGE AND SECOND STAGE OF SELECTION
Summary of Table A.1

	criterion		
Average of tests used at	N	"on time"	potential doctor
first stage of selection	120	44	55
second stage of selection	120	36	38
multiple R of first and second stages combined*		49	63

* It should be pointed out that the interpretation of these R's is not entirely clear since the selection Board had the First Stage results before them, taking with the remaining marks and ratings when determining the Final Board Rating. Thus the Second Stage is not entirely independent of the First Stage although it is by no means closely dependent either. The inter-correlation of the two Stages, which does not their interdependence and the fact that they are two different estimates of the same underlying traits is .30. There seems adequate reason to believe that the two multiple R's, even though they cannot be interpreted in absolute terms, are not entirely spurious, and also that they give us a fairly accurate index of the relative values of the various combinations. In fact a spuriously high inter-correlation between the two estimates might actually lower the value of R under certain conditions.

largely ruled out as (1) ratings were made two to five years after admissions tests were graded and (2) in any case if a doctor remembered any original gradings at all they would only be those of the ten candidates for whom he was the Group Observer — not all of whom were admitted. A second possible interpretation is that the tests, especially those of the Second Stage, have been designed to measure primarily those factors which are believed to make a good doctor regardless of whether or not he takes more than five years to finish his medical college education. A third possible interpretation (not necessarily inconsistent with the second) is that the "potential doctor" ratings are a more reliable and valid criterion than is "keeping up to schedule," which is based on the passing of university examinations. If the hundreds of research findings on the unreliability of traditional examinations in Europe, the U. K., and the U.S.A., are at all applicable to India (as the Radhakrishnan report seemed to think they were),** then this second interpretation is fairly plausible. Its plausibility is further enhanced by an examination of the detailed data (Table A.1, near the end of this paper). The correlations against the ratings of "potential doctor" vary much less from year to year than do the correlations with the academic criterion. This is not only true for the "judgements" at the second stage of selection, but also for the objective tests of knowledge and ability used at the first stage. The range between highest and lowest correlation averages more than twice as much for the academic criterion as for the ratings.

**The Report of the University Education Commission, Dec. 1948—Aug. 1949, Vol. I, Chapter X, pp. 333-6, etc.

The second striking thing in Table 6 is that the tests at the Second Stage seem to do a considerably poorer job of predicting academic achievement (+.36 vs. +.48), than do the First Stage tests. Both stages do about equally well (+.56 vs. +.55) at predicting "potential doctor". The second stage is significantly better at predicting "potential doctor" than it is in predicting mere academic success. These facts seem to lay emphasis on the second interpretation mentioned in the last paragraph (i.e. that the tests measure something other than mere academic ability); though they do not rule out the third interpretation (i.e. the relative unreliability of university examinations) either.

An examination of the multiple correlation of First plus Second Stage, against each of the criteria, is also revealing.* Ordinarily in adding up examination marks, equal weight is given to each mark. The multiple correlation method gives us the optimum weight for each part, to produce the largest possible correlation coefficient. The multiple R for First Stage plus Second Stage, against the criterion of finishing "on time", is +.49; since the First Stage alone predicts this criterion +.48, the Second Stage does not seem to add anything to purely adademic prediction. However, for the "potential doctor" criterion, the picture is somewhat different. Here the multiple R is +.63, as against +.55 for the First Stage alone. The Second Stage seems to be making some contribution to maximum possible prediction. (The exact nature, and extent of this contribution will be a subject for a later report).

We started this section with two questions. The answers to these two questions seem to be thus: (1) "Is the Second Stage selection process worthwhile?" If you consider the rating by medical staff, after two to five years of contact with the student, of how good a doctor they think he will be—if you consider this an *important* criterion, then the Second Stage is probably worthwhile. (2) "Is the Second Stage in the selection process absolutely necessary?" The answer seems to be definitely "No". The First Stage alone can carry the burden of selection, even if the Second is dropped. This is even more true if you consider academic achievement (and the financial and other implications of "finishing on time") to be the most important criterion. The preliminary battery of objective tests of knowledge, understanding, and ability seem to give a pretty good prediction of academic success. (In relation to studies of this kind done in the U.S.A., +.48 is a reasonably good figure). A multiple correlation weighting each of the tests optimally (instead of just adding and averaging, as is now done) should raise their predictive efficiency even higher.

It must not be forgotten, however, that we have studied only two out of the large number of possible criteria. There is not much reason to believe that other types of purely academic criteria would give radically different results. However, if we selected such criteria as "absence of discipline problems," "leadership," "ability to cooperate with others," or even "beside manner", the Second Stage might prove a much better predictor than the First.

* See footnote to Table 6.

4. Which tests are best?

This question will be most adequately answered by the full correlational study which is in progress. A rough indication of the power of some of the tests, however, has been given in Tables 7-12. For this analysis, data were gathered for three contrasting groups: (1) 50 students who finished in five years; (2) 50 students who were delayed half a year; (3) 50 students who were delayed one or more years (some of whom had not yet finished). Each of the following tables shows the relationship between marks on a particular test, or group of tests, and the time taken to finish the medical college course. (For simplicity, we have not reported the full range of admission test scores, but rather divided the candidates into only three grades: High, Moderate, and Low on each admission test).

One thing which should be kept in mind in evaluating these tables is that they are necessarily based on students actually admitted to the medical college, not on all who took the tests. No correction for this restriction of range has been applied in making the tables. Furthermore, since some tests are restricted more than others, the more valid measures may end up with *apparently* lower validities than some of the measures which actually discriminate less. An example will make one of the reasons for this clear. Candidates are given grades ranging from 1 (for the highest) to 9 (for the lowest) on each test, and in the Final Board Rating. When the Selection Committee sits, they select candidates primarily on the basis of the Final Board Rating. Thus no candidate with a Final Board grade lower than 6 is likely to be admitted. Candidates with Final Board grades of 7, 8, or 9 are automatically excluded from the class. However, several candidates with marks as low as 8 or 9 on the General Ability test may be admitted, because they did well enough on other tests to be graded 5 or 6 by the Final Board. Thus it is obvious that, other things being equal, the difference in later achievement between students rated highest and lowest on General Ability (range 1 to 9) is bound to be greater than the difference between students rated highest and lowest (range 1 to 6) on Final Board Grade. (The actual situation is, of course, usually more complex than this. A more technical discussion of this problem, and its solution, are given in the Appendix to this paper).

Table 7 shows the relationship between the objective (new type) pre-medical tests, (Physics and Chemistry, with Biology added last year) and performance in medical college. Of the 49 students rated High on the pre-medical tests, 24 (49%) finished the five-year course on time, 17 (35%) finished half a year late, and the remaining 8 (16%) finished or will finish one or more years late. The group of students who rated Moderate in the pre-medical battery seemed to contain about equal numbers of good, medium and poor students. But of the 52 students who rated Low, only 21 percent finished on time, while 50 percent fell one or more years behind.

The relationship between test and performance is even more marked for English comprehension (Table 8). Of the 58 students who rated High in their comprehension of written and spoken scientific English, 47 percent did well and 21 percent finished

TABLE 7. RELATIONSHIP BETWEEN OBJECTIVE PRE-MEDICAL TESTS AND PERFORMANCE IN MEDICAL COLLEGE

rating in pre-medical	N	group of fifty	percent falling in each group
high	49	on time	49
		½ year late	35
		1 or more years late	16
moderate	49	on time	31
		½ year late	37
		1 or more years late	33
low	52	on time	21
		½ year late	29
		1 or more years late	50

(or will finish) late. But of those rated Low in English comprehension, only five percent were able to keep up to schedule—and a phenomenal 68 percent (15 out of 22) fell badly behind.

TABLE 8. RELATIONSHIP BETWEEN TEST AND PERFORMANCE (ENGLISH COMPREHENSION)

rating in English comprehension	N	group of fifty	percent falling in each group
high	58	On time	47
		½ year late	33
		1 or more years late	21
moderate	70	on time	31
		½ year late	36
		1 or more years late	33
low	22	on time	5
		½ year late	27
		1 or more years late	68

Another test which discriminates very well is the General Ability test (Table 9). This is a measure of various types of mental ability required for educational success, in general. The table shows that only nine percent of those rated High in General Ability fell more than one-half year behind. But of those who scored Low in General Ability, 65 percent took one or more years to finish their medical education. This relationship has been supported by several other analyses that we have done. In one earlier study (of 124 students), the 26 who scored Low in General Ability lost more than three times as many years per student as did the 40 who were High on this test (1.23 vs. .38 years lost per student).

TABLE 9. GENERAL ABILITY TEST

rating in general ability	*N*	group of fifty	percent falling in each group
high	44	on time	48
		½ year late	43
		1 or more years late	9
moderate	72	on time	33
		½ year late	33
		1 or more years late	33
low	34	on time	15
		½ year late	20
		1 or more years late	65

The Pre-Medical Tests (Physics, Chemistry, Biology), most of the English Comprehension test, and the General Ability test are all group tests. They are administered to large number of candidates, in many different centres all over the country. The First Stage selection is based on an average of these tests, along with one or two others. Table 10 shows that the composite score on the tests mentioned above is a useful predictor of success in this medical college.

TABLE 10. COMPOSITE SCORE ON THE TESTS
(Tables 7—9)

rating on composite score of PMT+EC+GA	*N*	group of fifty	percent falling in each group
high	45	on time	49
		½ year late	40
		1 or more years late	11
moderate	70	on time	34
		½ year late	32
		1 or more years late	34
low	35	on time	11
		½ year late	29
		1 or more years late	60

Table 11 gives the actual number of students (rather than percentages) in each group, on which Table 10 is based. It is interesting to note that if the college had refused admission to all of the candidates who scored "Low", they would have lost only 4 really good students but would have kept out more than 40 percent of the 50 slowest ones.

TABLE 11. ACTUAL NUMBER OF STUDENTS

rating on composite score of PMT EC + GA	number finishing medical college			
	on time	½ year late	1 or more years late	total
high	22	18	5	45
moderate	24	22	24	70
low	4	10	21	35
total	50	50	50	150

The small group who are called to Vellore, on the basis of the First Stage results, are examined far more intensively and intimately. Many of these Second Stage selection devices are primarily for the assessment of non-cognitive factors—interests, personality, character, leadership qualities, motivation and drive, etc. The validities of the various Second Stage items are not reported here; we will present only the Final Board Rating. In making its final rating, the Board has before it all available information about the candidate—his scores in the objective First Stage tests, as well as ratings, descriptions, and comments given by Group Observer, Test Observer, psychologist or psychiatrist, etc. All of these enter into the Board's final judgement as to the fitness of the candidate. The validity of this final judgement is quite evident from Table 12. It is not a perfect predictor of "ability to finish college in five years"—but this is at least partly due to the fact that this is *not* the only thing the Board is trying to predict. (The relationship is also misleadingly lowered because of the "restriction of range" effect discussed above and in the Appendix). But even so, of the candidates rated High, the number who fall one or more years behind is less than half of the number who keep up to schedule. But among those few who were admitted in spite of being rated Low by the Final Board, the "one or more years behind" group is larger than both the "on time" and "half year late" groups put together.

5. CONCLUSION

The Vellore Christian Medical College selection programme has proved itself a useful method for selecting students. Its validity is higher than that of the more traditional selection methods, which were in use before 1946. Not only the two composite scores, but also the individual tests show an ability to predict academic success in medical education. In addition to what has been presented here, all of the unpublished data—some of it covering several more years, and, therefore, larger numbers of students than this report—point to the same conclusion.

TABLE 12

final board rating	*N*	group of fifty	percent failing in each group
high	70	on time	47
		½ year late	31
		1 or more years late	22
moderate	67	on time	22
		½ year late	36
		1 or more years late	42
low*	9	on time	22
		½ year late	22
		1 or more years late	56

*We cannot, of course, place much reliance in percentages calculated for the "Low" group, because the number of "Low" candidates admitted is necessarily small. However, presumably those admitted are among the better of the "Low" candidates; and, in the context of the table, it seems unlikely that the proportion failing would have been any less if a larger number of "Low" candidates had been admitted. *cf.* also discussion of "restriction of range" effect.

Since these results are in line with those of similar studies done in the U.S.A. and elsewhere, there seems to be no reason why the methods should not be equally applicable to any other medical college in India. This is true even if the Second Stage (personality assessment, individual interview, etc.) is dropped from the programme. Although the Second Stage adds to the value of the selection programme, it is not an absolute necessity. The objective group tests of pre-medical training, general ability, and comprehension are in themselves a valid battery. They can be applied simply and economically to large groups of candidates. Where personality and character are not to be taken into account in selecting students, the objective group tests form an adequate basis for selection.

Appendix A

The restriction of range problem*

The above presentation has tried to avoid the technicalities of statistical analysis, in favour of a more directly meaningful interpretation. But, for those who are aware of some of the knotty problems involved— and also for those who may wish to compare this with other such studies—a brief description is given here. It should be noted that—largely due to complex theoretical problems—the statistical analysis is not yet complete on the detailed tests. When this is finished, a full report will be published.

In any testing programme, we want to measure many attributes of the individual. The tests are our measuring instruments, and the score is the numerical evaluation of the ability or attribute. But before this numerical evaluation can be accepted as a valuable or meaningful measure, it becomes necessary to answer two questions:

(1) Is the measuring instrument reliable?

(2) To what extent is the measuring instrument measuring what it is supposed to measure, i.e. how valid is it?

It is the second question that was faced in this study, and some of its complications will be dealt with in this Appendix.

For dealing with the mathematical part, the tests will be denoted by X and the criterion by Y.

The answer to the second question, "How valid is the test?" is given by the coefficient of correlation between X and Y, which is known as the validity coefficient.

Table A.1 gives the validity coefficients for each Stage of selection against each criterion, separately for each year. Three correlations are reported for the "on time" criterion. The *biserial* r seemed especially relevant for this criterion; variation is allowed on one side of the curve ($\frac{1}{2}$ year late, 1 year late, $1\frac{1}{2}$ years late, etc.) but not on the other (no student can get a score *better* than "on time"). The biserial r corrects for this by "normalizing" the curve—i.e. by telling us what the product-moment r would be if the total group were normally distributed. (The biserial r was not relevant in this way to the "potential doctor" criterion, so was not calculated there).

Unfortunately, we cannot apply the correction for "restriction of range" (see below) to a biserial r. The assumptions underlying the biserial (a normal distribution) and the correction (a truncated distribution) are mutually contradictory. Therefore, we also calculated the product-moment r in each case. It was this product-moment r which was "corrected for restriction of range." The r's were then added, using Fisher's z-transformation, and the mean r calculated. It is this mean r which was reported in Table 6.

* This note was prepared mainly by S. P. Sangal, after consultation with R. G. Laha and A. Kudo—the latter suggesting most of the formulae.

TABLE A.1

	year				
	1947	1948	1949	1950	mean
number of cases	33	33	30	34	130
criterion : on time or behind schedule					
first stage of selection					
r_{bis}	+.304	+.719	+.488	+.287	
uncorrected r	+.278	+.520	+.494	+.194	
corrected r	+.342	+.741	+.500	+.213	+.48
second stage of selection					
r_{bis}	+.330	+.402	+.527	*****	
uncorrected r	+.396	+.334	+.473	*****	
corrected r	+.479	+.524	+.491	*****	+.36
criterion : rating as a potential doctor					
first stage of selection					
uncorrected r	+.528	+.431	+.621	+.364	
corrected r	+.480	+.521	+.635	+.444	+.55
second stage of selection					
uncorrected r	+.501	+.488	+.519	+.386	
corrected r	+.628	+.658	+.387	+.557	+.56

Note: Correlations replaced by ***** were not significantly different from zero. While these were taken into account in calculating mean r, the values are not reported to avoid possible misinterpretation.

It is well known that when the range of one or both variables in a correlation problem is restricted, the coefficient of correlation will be reduced. This effect is always present to some degree when we validate tests by correlating them with a criterion in a selected group. Where a large proportion of candidates are selected, the effect is slight; but when we have, as in the Vellore study, selection ratios as stiff as one in ten, the effect may be quite marked. If the lowering of the validity coefficients were equal for all tests, this probably would not matter. What makes the problem really serious is the fact that the lowering of the validity coefficient of a test is *proportional to the extent to which that test has been used as a basis for selection.* Thus, unless a correction is applied, a valid test which was weighted heavily in the selection battery may end up with a coefficient which is *lower* than that of an actually less valid test which was not given much weight in the composite score.

Formulae are available for various simple cases of restriction. (See R. L. Thorndike, *Personnel Selection : Tests and Measurement Techniques* or H. Gulliksen, *Theory of Mental Tests*). These formulae are not, however, directly applicable to our more complex case of two-stage selection. The Psychometric Research and Service Unit of the Indian Statistical Institute has been doing some original theoretical work on this problem. Fuller reports, with proofs of formulae, etc., are to be published later. Briefly, the method developed (and applied for making the corrections in Table A.1) is as follows:

Problem. Suppose we have K tests and L criteria. The selection is on the basis of K tests. On the first test all those whose score is $> X_1$ are selected. Those selected students are given the second test and all those whose score is $> X_2$ are selected, and so on till all K tests have been used.

Let the population variance covariance matrix be Σ. The problem is to estimate Σ from the known values of the variances and covariances for the selected groups.

Σ will be used to denote the population values at any stage, and S the variance-covariance matrix of the uncorrected values at any stage.

Assumptions.

$$(1)\quad \Sigma'_{12}\,\Sigma_{11}^{-1} = S'_{12}\;S_{11}^{-1}$$

$$(2)\quad \Sigma_{22}-\Sigma'_{12}\,\Sigma_{11}^{-1}\Sigma_{12} = S_{22}-S'_{12}\;S_{11}^{-1}S_{12}$$

Where at the n-th stage of selection we have the following notation :

$$\Sigma_{11} \equiv \begin{bmatrix} \sigma_{11}\sigma_{12} \ldots\ldots\ldots \sigma_{1,\,n-1} \\ \sigma_{21}\;\sigma_{22} \ldots\ldots\ldots \sigma_{2,\,n-1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sigma_{n-1,1}\;\sigma_{n-1,\,2} \ldots\ldots \sigma_{n-1,\,n-1} \end{bmatrix}$$

$$\Sigma_{22} \equiv ((\sigma_{nn}))$$

$$\Sigma_{12} \equiv \begin{bmatrix} \sigma_{1n} \\ \sigma_{2n} \\ . \\ . \\ . \\ \sigma_{n-1,n} \end{bmatrix}$$

$$S_{11} \equiv \begin{bmatrix} s_{11}\ s_{12} \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot s_{1n-1} \\ s_{21}\ s_{22} \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot s_{2n-1} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ s_{n-1.1}\ s_{n-1.2} \cdot \cdot \cdot \cdot \cdot s_{n-1.n-1} \end{bmatrix}$$

$$S_{22} \equiv (\!(s_{nn})\!)$$

$$S_{12} \equiv \begin{bmatrix} s_{1n} \\ s_{2n} \\ \vdots \\ \cdot \\ s_{n-1n} \end{bmatrix}$$

Procedure. Σ_{11} and all S's are known quantities.
From assumption (1) we have

$$\Sigma'_{12} = S'_{12}\ S_{11}^{-1}\ \Sigma_{11} \qquad \ldots \quad (1)$$

From assumption (2) we have

$$\Sigma_{22} = S_{22} - S'_{12}S_{11}^{-1}S_{12} + \Sigma'_{12}\Sigma_{11}^{-1}\Sigma_{12} \qquad \ldots \quad (2)$$

Substituting the value of Σ'_{12} from (1) in (2) we get the value of Σ_{22}. Proceeding this way at each stage we can estimate the population matrix Σ.

Appendix B

Significance tests

The tables in this paper illustrate some of the outstanding features of the data available. They do not, of course, *prove* anything—they merely help the non-statistician to understand the "meaning" of what the statistician finds. The data in them has been summarized and organized primarily for this purpose.

Significance tests have, however, been calculated on the raw data of the first five tables, combined in various ways. The problems involved in combining the raw data—and the reasons why they must be combined—have been discussed earlier in this paper. For example, if we take the full frequency distributions on which Table 4 is based, we find the "number of years late" stretching from 0 to 5. But the extreme right of these distributions is affected by a change in University policy about failures, which took place during the years covered by the New Selection Method groups. Thus we must combine the data of those detained one or more years to cancel out the effect of this change in University policy. At the other extreme, it is generally felt that a candidate may fall half year behind 'just by chance'. Sometimes a new student doesn't know what he is up against until his first examination, and it takes a failure to get him down to work. Sometimes special factors—a temporary illness, a serious emotional upset—may occur just at examination time. And the less-than-perfect reliability of the examinations themselves allow for a certain number of 'chance' failures. Thus, it seems best to combine those who are half year late with those who are on time, in making our statistical analysis. Analyses have, of course, been done with and without this combination—but it is not surprising that the uncombined data sometimes do not show significance.

A more serious problem is faced in the skewness of the distributions. Each of our measures of ability—years detained and papers failed—has a definite limit, better than which no candidate can possibly score. Thus we have a serious skewing, a piling up of 'highest' scores, which is purely an artifact of the measuring (i.e. nobody is allowed to finish in less than five years). It is as though we had a thermometer which could register only up to 100°F, so that all hotter days—whether 101° or 118°—would be recorded as exactly equal in temperature, i.e. just 100°. It would not be possible to calculate, accurately, either the mean or the standard deviation of the temperature. Similarly, we cannot use the usual powerful tests for the significance of the difference between means and distributions, but must use tests which do not require those statistics.

One such test is the p-test for the difference between proportions. This is frequently given in the form:

$$t = \frac{D_p}{\sigma_{D_p}}$$

$$\sigma_{D_p} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}$$

where D = the difference between the proportions

p_1 = the proportion of the first group possessing the characteristic

$q_1 = 1 - p_1$

N_1 = the number of cases in the first group

and p_2, q_2, and N_2 refer to the same characteristics of the second group.

A more precise form of the standard error the difference between two proportions is the following:

$$D = \sqrt{\frac{pq}{N_1} + \frac{pq}{N_2}}$$

where p = the proportion of the combined groups possessing the characteristic

$$q = 1 - p$$

and the remaining symbols have the same meaning as above. This formula is preferable when N is small, or the $p : q$ split far from the median. It is this latter formula which we have used.

The significance tests for the various tables are summarized in Table B.1.

TABLE B.1. SUMMARY OF TESTS OF SIGNIFICANCE ON DIFFERENCE BETWEEN OLD AND NEW SELECTION METHODS

table	part	D_E/σ_{Dp}	probability
1		2.87	.01
4		2.14	.03
5	division I	.08	.48
5	divisions II and III	1.90	.05

Note: All data dichotomized by combining "On Time" with "One-half Year Late", and "Failures" with "One or More Years Late."

Table 1. Data were dichotomized by combining the "On Time" and "One-half Year Late" groups and combining the "Failures" with the "One or More Years Late" group. The difference in proportions was significant at the 1 percent level. (This was also true when "failures" were dropped out, and only those who completed the course compared with each other).

Tables 2 and 3. As explained above, the artificial skewing of the data prevents the calculation of the "true" means and variance, and therefore of significance tests for these tables. It is true that means have been calculated for Table 2, and that a heavy burden of interpretation has been placed on them. This seems justified in

the light of the significance of Table 4. Note, however, that the purposes of Tables 2 and 4 differ. In Table 4 we are using the "detention rate" as a measure of the supposed *ability* of the students to pursue medical studies. We cannot calculate mean number of years required to complete the course as a measure of ability, as no student is permitted to finish in less than five years. In Table 2, however, our primary interest is in the economic cost of *extra* years spent in the institution. Therefore, it is appropriate in this context to calculate means.

Table 4. Data were dichotomized as has been explained. The difference between the proportions who furnished "half a year or less late" who were selected under the Old and the New methods is significant at the 3 percent level.

Table 5. When the proportion of First Divisioners who finish their medical course within half a year of the prescribed time is examined, the difference in this proportion between Old and New selection methods is negligible. For Second and Third Divisioners, however, the difference is significant at the 5 percent level of significance.

Paper received : March, 1959.