PREDICTION OF COLLEGE PERFORMANCE

FOR FRESHMEN AT VIRGINIA POLYTECHNIC INSTITUTE

by

Janet (Yates) Rutherford

Thesis submitted to the Graduate Faculty of the

Virginia Polytechnic Institute

in candidacy for the degree of

MASTER OF SCIENCE

in

STATISTICS

October 1963

Blacksburg, Virginia

TABLE OF CONTENTS

List of Tables

List of Expectancy Tables

List of Figures

I

INTRODUCTION

As colleges throughout the country expand and numbers of applicants multiply, it becomes increasingly desirable to place greater reliance on quantative methods of assessing candidate potential. The high attrition rate among freshmen has also led to some concern about admission requirements and curriculum advising. Admission officers are faced with a medley of test scores and other possibly relevant information for each student. They need some formula to consolidate all this information, together with some knowledge of what bearing each item may be expected to have on the candidate's future performance. Specific efforts have been made in this direction.

Under the auspices of the College Entrance Examination Board, Duggan and Hazlett (1) produced a workbook primarily for the use of college officers. It gives step by step guidance to, and an example of, data collection and computation for a multiple regression analysis, with the freshman grade average as the dependent variable, and Scholastic Aptitude Test scores, verbal and mathematical, and rank in high school as independent variables. Following this workbook, an officer may take a sample of one year's admissions and construct an equation from which he could predict grade averages for subsequent applicants, knowing their SAT scores and high school rank. He may also say with what certainty this prediction is made, by using an expectancy table, constructed on the basis of the standard error of the first sample.

For the sample of one hundred from Fordham college, Duggan and Hazlett obtained an R^2 value of .473, indicating that 47 percent of the variation in first year college grades could be accounted for by the three predictors used. Their prediction equation was

$$G = (.15)V + (.15)M + (.26)H + 46.97$$

(using obvious notation), and the standard error of estimate of an individual prediction was 4.51, where the grade average was measured as a percentage. This means, for example, that there is 95 percent probability that a student will obtain a grade average equal to his predicted average plus or minus 9.02.

Professor Long (2) at the College of William and Mary in Norfolk, reports on a similar, but more extensive, study. His sample was larger than that from Fordham (419) and his predictor variables numbered 31. His criterion for prediction was the College Quality Point Average (QCA), presumably over all four years of undergraduate study. The aim was not so much to produce a prediction equation from this data, but to find out how well the information could predict, and the most efficient set of predictor variables.

His R^2 value of .50 is not so different from the Fordham result, even with so many variables, and using his five "best" predictors, R^2=.46. Long discusses each variable, its significance and contribution, and the complications of intercorrelation among the variables. He reaches the conclusion that the best individual predictor is high school grades, but other variables, mostly academic but including certain personality characteristics, serve to improve the prediction.

In the fall of 1962 a proposal was put forward at Virginia Polytechnic Institute for the initiation of a comprehensive survey of admissions in 1962 and subsequent years. Information to be collected for each student included scores on the following tests: The Otis Gamma test (I.Q.), the Kudar Preference test, a reading comprehension test, chemistry and mathematic placement tests, and the College Board Entrance Examinations. Other data would be high school attended, grade and rank obtained there, department selected at V.P.I., sex, siblings, and father's education and occupation. The student's performance each quarter in the following four years in college would be recorded, and some correlation attempted between prior information and performance. Further investigation among students who withdrew, and top students in Virginia high schools who did not apply to V.P.I. were suggested. Since it was to be a continuing survey, information gained in one year could be used by admission officers to predict performance for the next year's applicants. The study was envisaged as a broad investigation of the relevance of a student's background to his college performance, and also as a scientific tool to help the overworked admission officer in accepting or rejecting future candidates.

The study actually undertaken and reported here is less comprehensive, and of one year's admissions only. It was modelled on the study outline of Duggan and Hazlett (1), but further investigations were made into some subgroups and other sources of variation, and a discriminatory analysis was carried out in an attempt to find more satisfactory predictions.

II

DESCRIPTION OF DATA

Source

The group selected for study comprised all freshman admission to Virginia Polytechnic Institute in the fall of 1961. Included in the group were Cooperative students, and students who had attended summer school, either voluntarily, or on request of the Admissions Officer who had considered part of their background to be weak. Data was collected for both male and female, civilian and cadet. At V.P.I. the Corps is compulsory for male freshmen, with several exceptions, viz., unfit medically, married, transfer, veteran, or over 21 years of age.

For each admission, information was recorded on the following items: sex, high school (Virginia or Out-of-State), rank in high school, verbal and mathematical scores on the College Board examinations, summer school attendance, status at V.P.I. (civilian or cadet), whether or not a Cooperative student, curriculum in the first quarter, changes in curriculum during freshman year, time of withdrawal if the student did not complete freshman year, Quality Credit Average (QCA) at the end of each quarter, and accumulative QCA for winter and spring quarters. The student number was recorded for identification. For students who had attended summer school, the summer credits were included in all the accumulative QCA's and the first quarter QCA. For Cooperative students, the year QCA was a record of performance over one or two quarters only. This was true also for students who dropped out during the year for any reason.

The high school ranks were converted to a standard scale, using the chart given in the computation workbook (1). The conversion gives the highest score to a student with rank 1, that is, top of his class, but the converted rank also depends on the size of class. For example, a student who is top of a class of 85 has a converted rank of 75, while another, first out of 250 students, has a converted rank of 79. Eighty is the maximum and 20 the minimum converted rank obtainable, and the middle student of any size class has a converted rank of 50. If the exact rank was not known, an approximate converted rank was calculated from the partial information such as "in upper third of class". Note was made of whether or not the full information on rank was available.

Every effort was made to procure correct information on all the above items, and only students with every item completed were finally included in the survey, (although a few blanks may have been left inadvertantly). Records for 1060 admissions, of whom 4½ percent were females, were completed and put on I.B.M. punched cards, and the data was then processed using I.B.M. electronic equipment.

Sample Characteristics

The distribution of the sample by sex, status, school and curriculum on admission is given in table 1. Eighty percent of the 1012 males and 88 percent of the 48 females were from Virginia schools; seventy percent of the males were admitted into the Corps of Cadets. Broad groupings of the curricula are made, and it may be seen that distributions among these groups were similar for cadets and civilians. The proportion of cadets was lower than that of civilians in Business, slightly higher in

Table 1. Distribution of sample by sex, status, school, and curriculum on admission.

Sex	Male							Female		
Status	Civilian			Cadet			All			
School / Curriculum	Va.	O-of-S	All	Va.	O-of-S	All	Males	Va.	O-of-S	All
Home Ec.	-	-	-	-	-	-	-	17	1	18
Agriculture	15 (5.6)	1 (2.9)	16 (5.2)	32 (5.9)	2 (1.2)	34 (4.8)	50 (4.9)	3	1	4
Forestry & Wildlife	11 (4.1)	- (0.0)	11 (3.6)	14 (2.6)	4 (2.4)	18 (2.5)	29 (2.9)	-	-	-
Business	32 (11.9)	6 (17.7)	38 (12.5)	43 (8.0)	18 (10.8)	61 (8.6)	99 (9.8)	2	-	2
Aerospace Engin.g	18 (6.7)	2 (5.9)	20 (6.6)	46 (8.5)	18 (10.8)	64 (9.1)	84 (8.3)	-	-	-
Chemical Engin.g	22 (8.2)	1 (2.9)	23 (7.5)	43 (8.0)	21 (12.5)	65# (9.2)	88# (8.7)	1	-	1
Civil Engin.g	20 (7.4)	2 (5.9)	24* (7.9)	71 (13.2)	14 (8.4)	85 (12.0)	109* (10.8)	-	-	-
Electrical Engin.g	44 (16.4)	7 (20.6)	51 (16.7)	96 (17.8)	21 (12.5)	117 (16.6)	168 (16.6)	-	-	-
Mechanical Engin.g	41 (15.2)	3 (8.8)	44 (14.4)	64 (11.9)	22 (13.2)	86 (12.2)	130 (12.8)	-	-	-
Other Engin.g	15 (5.6)	1 (2.9)	16 (5.2)	37 (6.9)	16 (9.6)	53 (7.5)	69 (6.8)	1	-	1
Architecture	20 (7.4)	5 (14.7)	25 (8.2)	32 (5.9)	15 (9.0)	47 (6.6)	72 (7.1)	2	1	3
Science & General studies	28 (10.4)	6 (17.7)	34 (11.2)	60 (11.1)	14 (8.4)	74 (10.5)	108 (10.7)	15	3	18
Distributive Education	3 (1.1)	- (0.0)	3 (1.0)	1 (0.2)	2 (1.2)	3 (0.4)	6 (0.6)	1	-	1
All Curricula	269 100%	34 100%	305* 100%	539 100%	167 100%	707# 100%	1012*# 100%	42	6	48

Numbers in parentheses are percentages of all-curricula totals.
* Includes 2 students with school not known.
# Includes 1 student with school not known.

All Males		
Va.	O-of-S	All
808	201	1012

some Engineering. Both Virginians and Out-of-Staters reflected these slight differences between cadet and civilian curricula distributions, with minor exceptions.

Table 2 gives the means and standard deviations of verbal score for males and females, and for the males this is broken down by status, high school and curriculum. The same information is presented in tables 3 and 4 for math. score and high school rank, respectively.

Several differences between means were tested with the Student's t-test, assuming a common variance in all underlying populations, and using a pooled estimate of that variance from each pair of samples compared.

In this total sample of 1961 admissions, females averaged higher verbal scores and ranks but lower math. scores than the males. The verbal score difference, however, was not significant.

Cadet means for verbal and math. scores were higher than corresponding civilian means. This did not hold within the small group from Out-of-State schools, but the larger group of Virginian students weighted the overall results. Rank averages in the two status groups did not differ significantly.

Out-of-State admissions averaged higher scores and ranks than did Virginians, both among cadets and civilians, the only non-significant comparison being among civilians on rank.

Subdivision by curriculum yielded groups too small for comparison in many cases. However, where comparisons were valid, the results were consistant. No individual curriculum provided inter-status or inter-school comparisons contradictory to those described above, although in many cases there was non-significance.

Table 2. Verbal score - Means and standard deviations by sex, status, school, and curriculum on admission.

Status	Civilian				Cadet			
School	Virginia		Out-of-State		Virginia		Out-of-State	
Curricula	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Agri-culture	405.2 (3)	75.77	640.0*	-	375 5 (2)	67.63	432.0* (4)	53.74
Forestry & Wildlife	457.8 (10)	107.17	-	-	445.6 (4)	61.31	402.3* (2)	71.43
Business	396.0 (2)	80.50	495.3*	57.64	414.6 (3)	72.17	418.1 (3)	70.55
Aerospace Engin.g	471.2 (12)	63.08	487.0*	97.58	494.6 (11)	70.89	532.5 (11)	73.18
Chemical Engin.g	466.5 (11)	92.44	563.0*	-	491.0 (10)	83.29	491.8 (6)	76.58
Civil Engin.g	433.0 (5)	68.99	404.5*	89.80	469.7 (7)	83.54	513.4 (10)	64.01
Electrical Engin.g	457.7 (9)	91.38	539.0*	73.61	476.8 (9)	70.38	513.3 (9)	35.98
Mechanical Engin.g	435.0 (6)	78.58	434.0*	64.55	466.3 (6)	67.98	502.0 (8)	95.64
Other Engin.g	430.7 (4)	76.00	668.0*	-	463.4 (5)	95.44	493.6 (7)	56.89
Arch-itecture	441.5 (8)	62.12	493.2*	114.37	469.7 (7)	74.10	487.9 (5)	61.02
Science & General studies	440.0 (7)	112.39	511.8*	122.22	506.9 (12)	71.55	552.3 (12)	86.91
Distribu-tive Education	271.0* (1)	31.43	-	-	347.0* (1)	-	306.5* (1)	36.06
All Males	437.1	62.71	507.0	94.82	467.3	81.23	493.9	80.90
All male civilians			444.9	90.67	All cadets		473.5	81.84
All females			488.7	107.84	All males		464.9	85.55

Numbers in parentheses are rank of mean in status - school group.

*Means are based on numbers less than 10 .

Table 3. Mathematical score - Means and standard deviations by sex, status, school, and curriculum on admission.

Status	Civilian				Cadet			
School	Virginia		Out-of-State		Virginia		Out-of-State	
Curricula	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Agri-culture	469.9 (4)	81.41	637.0*	-	458.3 (2)	67.61	488.0* (3)	2.83
Forestry & Wildlife	432.5 (2)	83.13	-	-	525.4 (4)	71.21	520.3* (4)	106.28
Business	460.8 (3)	60.97	462.3*	62.91	472.1 (3)	78.68	462.4 (2)	60.72
Aerospace Engin.g	540.4 (9)	64.81	631.5*	7.78	573.4 (12)	64.17	593.7 (10)	56.18
Chemical Engin.g	522.0 (6)	87.47	687.0*	-	570.4 (10)	77.43	566.0 (7)	83.84
Civil Engin.g	528.3 (7)	96.12	538.5*	37.48	558.2 (6)	76.36	567.9 (8)	56.88
Electrical Engin.g	555.3 (12)	82.57	613.4*	84.32	564.4 (9)	65.15	608.0 (11)	76.27
Mechanical Engin.g	542.7 (11)	72.20	586.0*	20.88	564.2 (8)	71.28	583.0 (9)	86.11
Other Engin.g	501.1 (5)	73.37	643.0*	-	527.4 (5)	78.73	564.0 (6)	54.79
Arch-itecture	540.6 (10)	70.16	608.2*	61.27	558.2 (6)	66.77	556.9 (5)	69.31
Science & General studies	536.7 (8)	93.10	558.7	38.94	572.9 (11)	77.00	611.6 (12)	74.96
Distribu-tive Education	290.0* (1)	31.19	-	-	373.0* (1)	-	291.5* (1)	16.26
All Males	517.6	88.59	574.3	80.04	547.8	80.05	563.0	86.47
All male civilians			523.7	89.06	All cadets		551.4	81.75
All females			512.2	85.50	All males		543.0	84.91

Numbers in parentheses are rank of mean in status - school group.

*Means are based on numbers less than 10 .

Table 4. Rank in high school - Means and standard deviations by sex, status, school, and curriculum on admission.

Status	Civilian				Cadet			
School	Virginia		Out-of-State		Virginia		Out-of-State	
Curricula	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Agri-culture	58.9 (11)	8.80	63.0*	-	53.3 (3)	5.81	49.0* (1)	4.24
Forestry & Wildlife	53.2 (4)	7.20	-	-	55.6 (5)	6.06	49.0* (2)	4.90
Business	50.0 (2)	7.64	55.3*	12.60	51.3 (2)	6.40	51.3 (3)	8.02
Aerospace Engin.g	56.3 (8)	5.10	60.5*	.71	57.6 (6)	5.16	59.0 (9)	4.49
Chemical Engin.g	58.9 (12)	6.30	54.0*	-	57.8 (11)	7.10	61.8 (12)	5.76
Civil Engin.g	56.1 (7)	5.66	60.0*	4.24	55.6 (6)	7.30	54.9 (5)	5.88
Electrical Engin.g	55.5 (5)	7.87	58.7*	4.35	55.9 (9)	6.10	59.9 (10)	6.63
Mechanical Engin.g	55.8 (6)	8.10	51.0*	9.17	55.8 (7)	6.37	57.0 (7)	4.31
Other Engin.g	57.8 (10)	7.48	62.0*	-	55.8 (8)	6.27	61.3 (11)	6.62
Arch-itecture	53.0 (3)	5.61	53.6*	6.50	54.5 (4)	6.42	56.0 (6)	4.23
Science & General studies	56.5 (9)	7.03	56.8*	10.29	57.9 (12)	6.88	58.6 (8)	4.55
Distribu-tive Education	42.3* (1)	5.69	-	-	44.0* (1)	-	52.0* (4)	1.41
All Males	55.3	7.65	56.6	7.98	55.7	6.66	57.8	4.94
All male civilians			55.4	7.66	All cadets		56.1	6.66
All females			60.0	8.73	All males		55.7	6.98

Numbers in parentheses are rank of mean in status - school group.

*Means are based on numbers less than 10 .

Because of the unequal group sizes involved, simultaneous comparisons of curriculum groups among themselves would be difficult and tedious, but simple inspection and ordering of the means reveals certain interesting consistancies. If the curriculum groups are ranked by average for each variable and status-school group, Distributive Education, Business, Agriculture, and Forestry and Wildlife almost always take the four lowest places, with Distributive Education at the bottom in every case but one. Science and General Studies, Aerospace Engineering, Chemical Engineering, and Electrical Engineering appear to attract, on average, applicants with high qualifications; in particular, the Chemical Engineering group had exceptionally high rank average, but were surpassed, in general, by the Mechanical Engineering group on math. score average.

The median score for a group, more easily obtained than the mean in the absence of high speed computing facilities, is often used by administration to compare that group with another, or with a previous year. Table 5 gives the medians for verbal and math. scores in the female, civilian and cadet groups.

It may be seen that the median values lie reasonably close to the means given in the previous tables. The lowest and highest values for the two scores and three groups are also recorded in table 5, showing that a low score on one test did not, in itself, exclude a candidate from admission.

Other factors which may be connected with QCA are summer school attendance, Cooperative study, and withdrawal or "drop-out". Whether or not any of these factors has an effect on freshman performance or, in

Table 5. Median values and ranges of verbal and mathematical scores for females, civilians, and cadets.

	Median	Low	High
Females (48)			
Verbal score	483	293	724
Mathematical score	512	310	687
Male Civilians (305)			
Verbal score	437	226	688
Mathematical score	521	261	787
Male Cadet (707)			
Verbal score	475	266	721
Mathematical score	556	280	787

Table 6. Incidence of prior summer school attendance, subsequent withdrawal. and Cooperative students.

	No Summer School	Voluntary Summer School	Trial Summer School	Total
No Drop-Out	743	23	82	848
Drop-Out After				
First Quarter	43	-	8	51
Second Quarter	44	1	13	58
Cooperative	53	-	-	53
Part Time (Not Cooperative)	2	-	-	2
Total	885	24	103	1012

the case of drop-out, is a result of poor early performance, they all have a definite effect on the QCA as used in this study, since as explained above, accumulative QCA's include summer school grades and QCA's based on part of the year only. Numbers involved are presented in table 6. The cause of withdrawal was not recorded, but assuming that a certain proportion of withdrawals resulted from poor grades in first or second quarters, it is interesting to note that there was a significantly higher proportion of drop-outs among students who had attended summer school on trial than among other freshmen.

Excluding summer school, Cooperative and drop-out students, there were 654 male admissions, 456 cadets and 198 civilians, for whom full information was recorded on every item. There were a further 86 who had only partial rank information.

Freshman QCA's are presented fully in conjunction with prediction results, but a brief summary is given in table 7. Year QCA's average a little less than first quarter QCA's, but in both cases, the mean for the 48 females is about 1.5, against just over 1.0 for the 1012 males. The civilian yearly average is slightly higher than the cadet, but the difference is not statistically significant. A year QCA of 1.000 is considered the minimum level of satisfactory achievement, and the numbers of freshmen in any group attaining this mean is used as another indicator of group achievement. Table 7 gives these proportions for females, males, civilians and cadets. Chi-square tests show that, by this criterion also, females score better than males, but that the civilian-cadet difference is not significant.

Table 7. Summary of QCA's obtained.

	Mean	Standard Deviation	Number with QCA	
			less than 1.000	1.000 & above
Female				
First Quarter	1.59483	.68032	11	37
Year	1.52290	.59926	9	39
Male				
First Quarter	1.08678	.72137	490	522
Year	1.06584	.59331	512	500
Civilian				
Year	1.09753	.62740	148	157
Cadet				
Year	1.05217	.57834	364	343

About fourteen percent of both males and females changed their curriculum during the freshman year. It proved impossible to do any analysis of cause or effect of this, and it is clearly a complication to any breakdown by curriculum of correlations between qualifications and final performance. In many cases the changes were few enough to justify ignoring possible effects, but the change into Business from various other departments was quite sizeable. The distribution of the 1012 male admissions by status, school and curriculum at the end of freshman year is given in table 8.

Table 8. Distribution of males by status, school, and third quarter curriculum.

Status	Civilian			Cadet			All Males
School	Va.	O-of-S	All civ	Va.	O-of-S	All cad	
Curriculum							
Agriculture	17 (6.3)	1 (2.9)	18 (5.9)	36 (6.7)	3 (1.8)	39 (5.5)	57 (5.6)
Forestry & Wildlife	13 (4.9)	- (0.0)	13 (4.3)	10 (1.8)	2 (1.2)	12 (1.7)	25 (2.5)
Business	44 (16.4)	7 (20.6)	51 (16.7)	64 (11.9)	19 (11.4)	83 (11.7)	134 (13.2)
Aerospace Engineering	14 (5.2)	2 (5.9)	16 (5.2)	42 (7.8)	19 (11.4)	61 (8.6)	77 (7.6)
Chemical Engineering	20 (7.4)	1 (2.9)	21 (6.9)	40 (7.4)	21 (12.5)	62# (8.8)	83# (8.2)
Civil Engineering	15 (5.6)	1 (2.9)	18* (5.9)	71 (13.2)	15 (9.0)	86 (12.2)	104* (10.3)
Electrical Engineering	41 (15.2)	7 (20.6)	48 (15.7)	85 (15.8)	18 (10.8)	103 (14.6)	151 (14.9)
Mechanical Engineering	39 (14.5)	3 (8.9)	42 (13.8)	65 (12.0)	22 (13.2)	87 (12.3)	129 (12.8)
Other Engineering	13 (4.8)	1 (2.9)	14 (4.6)	36 (6.7)	14 (8.4)	50 (7.1)	64 (6.3)
Architecture	20 (7.4)	4 (11.8)	24 (7.9)	28 (5.2)	13 (7.7)	41 (5.8)	65 (6.4)
Science & General studies	29 (10.8)	6 (17.7)	35 (11.5)	57 (10.6)	15 (9.0)	72 (10.2)	107 (10.6)
Distributive Education	4 (1.5)	1 (2.9)	5 (1.6)	5 (0.9)	6 (3.6)	11 (1.5)	16 (1.6)
All curricula	269 100%	34 100%	305* 100%	539 100%	167 100%	707# 100%	1012*# 100%

Numbers in parentheses are percentages of all-curricula totals.

* Includes 2 students with school not known.
\# Includes 1 student with school not known.

III

PREDICTION BY MULTIPLE REGRESSION

Procedure

In the standard multiple regression model,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_k X_k ,$$

y is the QCA, either at the end of the first quarter or at the end of freshman year. There are three independent X variables; verbal score, math. score and high school rank.

For 1060 admissions we have the values of the y's and X's, and our aim is to find the best estimates of the partial regression coefficients, β's, which would most nearly satisfy the equation simultaneously for all admissions. This involves a regular multiple regression analysis, described by many statistical authors, for example, Kramer (3).

Since the analysis of sample characteristics showed the male and female performances to be so dissimilar, a multiple regression analysis was done for each sex. These male and female analyses were carried out for two sets of QCA's, first quarter and accumulative year, and for the year QCA two more analyses were done for the two sections of the male admissions, civilian and cadet. Finally, a group which was considered to be homogeneous in as many respects as possible was subjected to analysis. This was the cadet group excluding summer school students, drop-outs, Cooperative students, and those with only partial rank information.

Most of the computations of the analyses were done on the I.B.M. 1620 machine, using a prepared program, 6.0.007, Stepwise Multiple Linear Regression (4). This program gives sums of variables, uncorrected sums

of squares and cross-products, averages, square root of the corrected sums of squares, and simple correlations between variables. It then selects the X variable, say X_i, which has the highest simple correlation with the y variable and calculates the best estimates, b_0 and b_i of β_0 and β_i, in the equation,

$$y = \beta_0 + \beta_i X_i ,$$

also the standard error of y, and the Student t value to test the significance of b_i. The next most correlated variable, say X_j, is introduced and the process repeated for the equation,

$$y = \beta_0 + \beta_i X_i + \beta_j X_j ,$$

getting new estimates, b_0, b_i and b_j, a new standard error and new significance levels.

The process is continued until all the X variables are included, in this case just three, and final values for R^2 and the standard partials are typed out. At this point it is possible to choose the regression equation including only the significant X variables, or as many as required, to use in predicting y values for the same, or a new set of data.

With the exception of this step-by-step introduction of the variables, the intermediate and final results produced by the computer correspond to most of those set out in the workbook (1). There is a slight discrepancy in the definition of standard error; calculated by Duggan and Hazlett's method, the standard error of the estimate is actually $\sqrt{\frac{n-k-1}{n}}$ times the standard error of y calculated by the computer, where n is the number of obversations and k is the number of X variables. With the large samples involved, however, this difference is negligible.

Duggan and Hazlett give no basis for their interpretation of the prediction weights by means of the "proportional contribution of each predictor to the prediction" (1). Without explanation, the calculations appear rather misleading, and are therefore omitted from these analyses.

Their second suggested interpretation of the prediction weights is the calculation of the correlations between each predictor and the predicted QCA. These values, r_{VP}, r_{MP} and r_{HP}, are computed here for each sample under analysis at the three stages of the step-wise regression, using the simple formula, $r_{VP}=r_{VY}/R$, etc..

Notation and Definitions

N = number in sample

X_V = verbal score

X_M = mathematical score

X_H = high school rank

y_1 = first quarter QCA

Y = year QCA

$\bar{X}_V = \sum X_V/N$ = mean verbal score of sample ($\sum$ denotes summation over all members of sample)

Similarly for $\bar{X}_M$, $\bar{X}_H$, $\bar{y}_1$, and $\bar{Y}$.

$SD_V = \sqrt{\frac{\sum(X_V-\bar{X}_V)^2}{(N-1)}}$ = standard deviation of X_V about mean of sample

Similarly for SD_M, SD_H, SD_y, and SD_Y.

$r_{VM} = \frac{\sum(X_V-\bar{X}_V)(X_M-\bar{X}_M)}{\sqrt{\sum(X_V-X_V)^2\sum(X_M-X_M)^2}}$ = simple correlation between verbal and mathematical scores

Similarly for r_{VH}, r_{MH}, r_{Vy}, r_{VY}, r_{My}, r_{MY}, r_{Hy}, r_{HY}.

β_0 = constant term in prediction equation

$\beta_V, \beta_M, \beta_H$, are partial regression coefficients of X_V, X_M, X_H.

b_0, b_V, b_M, b_H, are best estimates of β_0, β_V, β_M, β_H.

R = multiple regression coefficient

SE_Y = standard error of Y = $\sqrt{\frac{\Sigma(\text{observed Y - predicted Y})^2}{(N-k-1)}}$

where k= number of X variables

SS_Y = corrected sum of squares of Y

SS_E = sum of squares due to error = $(SE_Y)^2/(N-k-1)$

$r_{VP} = r_{VY}/R$ = correlation between predictor for V and over-all prediction

Similarly for r_{MP}, r_{HP}.

$\hat{y}_1$, $\hat{Y}$ = predicted values for first quarter QCA and year QCA, respectively

b_V', b_M', b_H', standard partials or normal coefficients

Prediction Equations

1.10 First Quarter QCA Prediction for Females

N = 48

Means and Standard Deviations

$\bar{X}_V$ = 488.67	SD_V = 107.84	
$\bar{X}_M$ = 512.21	SD_M = 85.49	
$\bar{X}_H$ = 60.04	SD_H = 8.73	
$\bar{y}_1$ = 1.59483	SD_y = .68032	SS_y(corrected) = 21.753362

Simple Correlations

r_{VM} = .723	r_{VH} = .278	r_{Vy} = .281
	r_{MH} = .290	r_{My} = .363
		r_{Hy} = .570

Step-by-Step Regression

Using rank: k=1 $\hat{y}_1 = b_0 + b_H X_H$ (1)

SE_y = .56521 SS_E = 14.695268 R^2 = .32446 R = .5696 r_{HP} = 1.000

b_0 = -1.06964

b_H = .04437 SE of b_H = .00944 t_{46} = 4.700**

Using rank and math. score: k=2 $\hat{y}_1 = b_0 + b_M X_M + b_H X_H$ (2)

SE_y = .55309 SS_E = 13.765885 R^2 = .36718 R = .6060 r_{HP} = .940

r_{MP} = .599

b_0 = -1.65738

b_H = .03950 SE of b_H = .00965 t_{45} = 4.093**

b_M = .00171 SE of b_M = .00098 t_{45} = 1.743

Using all X variables: k=3 $\hat{y}_1 = b_0 + b_V X_V + b_M X_M + b_H X_H$ (3)

SE_y = .55909 SS_E = 13.753592 R^2 = .36774 R = .6064 r_{HP} = .940

r_{MP} = .599

r_{VP} = .463

b_0 = -1.66174

b_H = .03970 SE of b_H = .00981 t_{44} = 4.047** b_H' = .50965

b_M = .00191 SE of b_M = .00139 t_{44} = 1.372 b_M' = .24015

b_V = -.00021 SE of b_V = .00110 t_{44} = -.198 b_V' = -.03449

Best prediction equation is (2) :

$$\hat{y}_1 = -1.65738 + (.00171)X_M + (.03950)X_H$$

with SE_y = .55309

1.20 First Quarter QCA Prediction for Males

N = 1012

Means and Standard Deviations

$\bar{X}_V$ = 464.88 SD_V = 85.55

$\bar{X}_M$ = 543.04 SD_M = 84.91

$\bar{X}_H$ = 55.90 SD_H = 6.98

$\bar{y}_1$ = 1.08678 SD_y = .72137 SS_y(corrected) = 526.094042

Simple Correlations

r_{VM} = .552 r_{VH} = .309 r_{Vy} = .324

r_{MH} = .319 r_{My} = .453

r_{Hy} = .488

Step-by-Step Regression

Using rank: k=1 $\hat{y}_1 = b_0 + b_H X_H$ (1)

SE_y = .63005 SS_E = 400.932632 R^2 = .23791 R= .4878 r_{HP}=1.000

b_0 = -1.73124

b_H = .05041 SE of b_H = .00283 t_{1010} = 17.757**

Using rank and math. score: k=2 $\hat{y}_1 = b_0 + b_M X_M + b_H X_H$ (2)

SE_y = .58830 SS_E = 349.211762 R^2 = .33622 R= .5798 r_{HP} = .842

r_{MP} = .781

b_0 = -2.64777

b_H = .03950 SE of b_H = .00279 t_{1009} = 14.125**

b_M = .00281 SE of b_M = .00022 t_{1009} = 12.223**

Using all X variables: k=3 $\hat{y}_1 = b_0 + b_V X_V + b_M X_M + b_H X_H$ (3)

SE_y = .58824 SS_E = 348.794508 R^2 = .33701 R = .5805 r_{HP} = .841

r_{MP} = .780

r_{VP} = .558

b_0 = -2.67312

b_H =	.03897	SE of b_H = .00283	t_{1008} = 13.738**	b_H' = .37713
b_M =	.00266	SE of b_M = .00026	t_{1008} = 10.005**	b_M' = .31332
b_V =	.00029	SE of b_V = .00026	t_{1008} = 1.108	b_V' = .03457

Best prediction equation is (3) :

$$\hat{y}_1 = -2.67312 + (.00029)X_V + (.00266)X_M + (.03897)X_H$$

with SE_y = .58824

2.10 Year QCA Prediction for Females

N = 48

Means and Standard Deviations

$\bar{X}_V$ = 488.67	SD_V = 107.84	
$\bar{X}_M$ = 512.21	SD_M = 85.49	
$\bar{X}_H$ = 60.04	SD_H = 8.73	
$\bar{Y}$ = 1.52290	SD_Y = .59926	SS_Y(corrected) = 16.878375

Simple Correlations

r_{VM} = .723	r_{VH} = .278	r_{VY} = .212
	r_{MH} = .290	r_{MY} = .254
		r_{HY} = .562

Step-by-Step Regression

Using rank: k=1 $\hat{Y}=b_0+b_H X_H$ (1)

SE_Y= .50111 SS_E= 11.551117 R^2= .31563 R= .5618 r_{HP}= 1.000

b_0= -.79197

b_H= .03855 SE of b_H = .00837 t_{46}= 4.606**

Using rank and math. score: k=2 $\hat{Y}=b_0+b_M X_M+b_H X_H$ (2)

SE_Y= .50326 SS_E= 11.397178 R^2= .32475 R= .5699 r_{HP}= .986

r_{MP}= .446

b_0= -1.03091

b_H= .03657 SE of b_H = .00878 t_{45}= 4.164**

b_M= .00069 SE of b_M = .00089 t_{45}= .779

Using all X variables: k=3 $\hat{Y}=b_0+b_V X_V+b_M X_M+b_H X_H$ (3)

SE_Y= .50889 SS_E= 11.394637 R^2= .32490 R= .5700 r_{HP}= .986

r_{MP}= .446

r_{VP}= .372

b_0= -1.03285

b_H= .03666 SE of b_H = .00892 t_{44}= 4.106** b_H'= .53425

b_M= .00078 SE of b_M = .00126 t_{44}= .619 b_M'= .11193

b_V= -.00010 SE of b_V = .00100 t_{44}= -.097 b_V'=-.01747

Best prediction equation is (1) :

$\hat{Y}$ = -.79197 + (.03855)X_H

with SE_Y= .50111

2.20 Year QCA Prediction for Males

N = 1012

Means and Standard Deviations

$\bar{X}_V$ = 464.88 SD_V= 85.55

$\bar{X}_M$ = 543.04 SD_M= 84.91

$\bar{X}_H$ = 55.90 SD_H= 6.98

$\bar{Y}$ = 1.06584 SD_Y= .59360 SS_Y(corrected) = 356.238068

Simple Correlations

r_{VM} = .552 r_{VH} = .309 r_{VY} = .335

r_{MH} = .319 r_{MY} = .422

r_{HY} = .517

Step-by-Step Regression

Using rank: k=1 $\hat{Y}=b_0+b_H X_H$ (1)

SE_Y= .50821 SS_E= 260.860178 R^2= .26774 R= .5174 r_{HP}= 1.000

b_0= -1.39414

b_H= .04400 SE of b_H = .00228 t_{1010}= 19.217**

Using rank and math. score: k=2 $\hat{Y}=b_0+b_M X_M+b_H X_H$ (2)

SE_Y= .48223 SS_E= 234.638685 R^2= .34134 R= .5842 r_{HP}= .885

r_{MP}= .722

b_0= -2.04678

b_H= .03624 SE of b_H = .00229 t_{1009}= 15.808**

b_M= .00200 SE of b_M = .00018 t_{1009}= 10.618**

Using all X variables: k=3 $\hat{Y}=b_0+b_V X_V+b_M X_M+b_H X_H$ (3)

SE_Y= .48135 SS_E= 233.551405 R^2= .34437 R= .5868 r_{HP}= .881

r_{MP}= .719

r_{VP}= .571

b_0= -2.08729

b_H= .03539	SE of b_H = .00232	t_{1008}= 15.246**	b_H'= .41618
b_M= .00176	SE of b_M = .00021	t_{1008}= 8.103**	b_M'= .25233
b_V= .00046	SE of b_V = .00021	t_{1008}= 2.163*	b_V'= .06714

Best prediction equation is (3) :

$$\hat{Y} = -2.08729 + (.00046)X_V + (.00176)X_M + (.03539)X_H$$

with SE_Y= .48135

<u>Simplification</u> (See page 41)

Coefficients used in computer:

b_0= -2.0872948 b_V= .0004658820 b_M= .0017639799 b_H= .035394927

	<u>Simplified b's</u>	<u>(X_V, X_M, X_H)</u>	<u>Error in $\hat{Y}$</u>
1a	Cut off at 5 dec.	(465, 543, 56)	.005 (Av.)
1b	Cut off at 5 dec.	(721, 787, 78)	.008 (Max.)
2	Rounded to 5 dec.	(721, 787, 78)	.008 (Max.)
3	Rounded to 4 dec.	(721, 787, 78)	.05 (Max.)
4	Round b_V, b_M to 4; b_H to 3 dec.	(721, 787, 30)	.04 (Max.)
5	Round b_V, b_M to 3; b_H to 2 dec.	(226, 787, 78)	.4 (Max.)

Using simplification 4: $(5)X_V + (18)X_M + (350)X_H - 20{,}900 = (10{,}000)\hat{Y}$

Using simplification 5: $(2)X_M + (40)X_H - 2{,}100 = (1{,}000)\hat{Y}$

2.21 Year QCA Prediction for Male Civilians

N = 305

Means and Standard Deviations

$\bar{X}_V$ = 444.85 SD_V= 90.66

$\bar{X}_M$ = 523.65 SD_M= 89.06

$\bar{X}_H$ = 55.36 SD_H= 7.65

$\bar{Y}$ = 1.09753 SD_Y= .62741 SS_Y(corrected) = 119.668284

Simple Correlations

r_{VM}= .525 r_{VH} = .249 r_{VY} = .364

r_{MH} = .325 r_{MY} = .435

r_{HY} = .490

Step-by-Step Regression

Using rank: k=1 $\hat{Y}=b_0+b_H X_H$ (1)

SE_Y= .54794 SS_E= 90.972188 R^2= .23980 R= .4897 r_{HP}= 1.000

b_0= -1.12416

b_H= .04013 SE of b_H = .00410 t_{303}= 9.776**

Using rank and math. score: k=2 $\hat{Y}=b_0+b_M X_M+b_H X_H$ (2)

SE_Y= .51731 SS_E= 80.818110 R^2= .32465 R= .5698 r_{HP}= .860

r_{MP}= .763

b_0= -1.80605

b_H= .03192 SE of b_H = .00409 t_{302}= 6.160**

b_M= .00217 SE of b_M = .00035 t_{302}= 7.788**

Using all X variables: k=3 $\hat{Y}=b_0+b_V X_V+b_M X_M+b_H X_H$ (3)

SE_Y= .51226 SS_E= 78.985502 R^2= .33996 R= .5831 r_{HP}= .840

r_{MP}= .746

r_{VP}= .624

b_0= -1.93011

b_H= .03087	SE of b_H = .00407	t_{301}= 7.571**	b_H'= .37670
b_M= .00165	SE of b_M = .00039	t_{301}= 4.158**	b_M'= .23544
b_V= .00101	SE of b_V = .00038	t_{301}= 2.643**	b_V'= .14614

Best prediction equation is (3) :

$$\hat{Y} = -1.93011 + (.00101)X_V + (.00165)X_M + (.03087)X_H$$

with SE_Y= .51226

Simplification (See page 41)

Coefficients used in computer:

b_0= -1.930117h b_V= .0010113h20 b_M= .0016586687 b_H= .030875524

	Simplified b's	(X_V, X_M, X_H)	Error in $\hat{Y}$
1a	Cut off at 5 dec.	(445, 524, 56)	.005 (Av.)
1b	Cut off at 5 dec.	(688, 787, 76)	.008 (Max.)
2	Rounded to 5 dec.	(226, 787, 76)	.001 (Max.)
3	Rounded to 4 dec.	(688, 787, 30)	.05 (Max.)
4	Round b_V, b_M to 4; b_H to 3 dec.	(688, 787, 30)	.05 (Max.)
5	Round b_V, b_M to 3; b_H to 2 dec.	(226, 787, 30)	.2 (Max.)

Using simplification 4: $(10)X_V + (16)X_M + (310)X_H - 193300 = (10,000)\hat{Y}$

Using simplification 5: $X_V + (2)X_M + (30)X_H - 1,900 = (1,000)\hat{Y}$

2.22 Year QCA Prediction for Male Cadets

N = 707

Means and Standard Deviations

$\bar{X}_V$ = 473.52 SD_V= 81.83

$\bar{X}_M$ = 551.40 SD_M= 81.73

$\bar{X}_H$ = 56.14 SD_H= 6.66

$\bar{Y}$ = 1.05217 SD_Y= .57834 SS_Y(corrected) = 236.143767

Simple Correlations

r_{VM} = .550 r_{VH} = .336 r_{VY} = .335

r_{MH} = .310 r_{MY} = .431

r_{HY} = .537

Step-by-Step Regression

Using rank: k=1 $\hat{Y}=b_0+b_H X_H$ (1)

SE_Y= .48824 SS_E= 168.056700 R^2= .28833 R= .5370 r_{HP}= 1.000

b_0= -1.56564

b_H= .04663 SE of b_H = .00275 t_{705}= 16.900**

Using rank and math. score: k=2 $\hat{Y}=b_0+b_M X_M+b_H X_H$ (2)

SE_Y= .46114 SS_E= 149.705670 R^2= .36604 R= .6050 r_{HP}= .888

r_{MP}= .712

b_0= -2.26661

b_H= .03874 SE of b_H = .00274 t_{704}= 14.134**

b_M= .00207 SE of b_M = .00022 t_{704}= 9.288**

Using all X variables: k=3 $\hat{Y}=b_0+b_V X_V+b_M X_M+b_H X_H$ (3)

SE_Y= .46116 SS_E= 149.505988 R^2= .36688 R= .6057 r_{HP}= .886

r_{MP}= .712

r_{VP}= .553

b_0= -2.28543

b_H= .03817 SE of b_H = .00280 t_{703}= 13.618** b_H'= .43953

b_M= .00194 SE of b_M = .00025 t_{703}= 7.567** b_M'= .27551

b_V= .00025 SE of b_V = .00025 t_{703}= .973 b_V'= .03576

Best prediction equation is (2), but equation used is (3) :

$$\hat{Y} = -2.28543 + (.00025)X_V + (.00194)X_M + (.03817)X_H$$

with SE_Y= .46116

Simplification (See page 41)

Coefficients used in computer:

b_0= -2.2854301 b_V= .0002527685 b_M= .0019496882 b_H= .038171769

	Simplified b's	(X_V, X_M, X_H)	Error in $\hat{Y}$
1a	Cut off at 5 dec.	(474, 551, 56)	.007 (Av.)
1b	Cut off at 5 dec.	(721, 787, 78)	.01 (Max.)
2	Rounded to 5 dec.	(721, 280, 78)	.002 (Max.)
3	Rounded to 4 dec.	(721, 787, 35)	.08 (Max.)
4	Round b_V, b_M to 4; b_H to 3 dec.	(721, 787, 78)	.09 (Max.)
5	Round b_V, b_M to 3; b_H to 2 dec.	(721, 280, 35)	.1 (Max.)

Using simplification 5: $(2)X_M + (30)X_H - 2{,}300 = (1{,}000)\hat{Y}$

2.23 Year QCA Prediction for Male Cadets, Excluding Summer School, Drop-Out and Cooperative Students, and Students With Only Partial Rank Information.

$N = 456$

Means and Standard Deviations

$\bar{X}_V = 481.73$ $SD_V = 77.26$

$\bar{X}_M = 558.64$ $SD_M = 75.56$

$\bar{X}_H = 56.48$ $SD_H = 6.68$

$\bar{Y} = 1.08148$ $SD_Y = .54137$ $SS_Y(\text{corrected}) = 133.352147$

Simple Correlations

$r_{VM} = .452$ $r_{VH} = .289$ $r_{VY} = .282$

$r_{MH} = .224$ $r_{MY} = .396$

$r_{HY} = .534$

Step-by-Step Regression

Using rank: k=1 $\hat{Y}=b_0+b_H X_H$ (1)

$SE_Y = .45832$ $SS_E = 95.365979$ $R^2 = .28486$ $R = .5337$ $r_{HP} = 1.000$

$b_0 = -1.35945$

$b_H = .04321$ SE of $b_H = .00321$ $t_{454} = 13.448^{**}$

Using rank and math. score; k=2 $\hat{Y}=b_0+b_M X_M+b_H X_H$ (2)

$SE_Y = .43229$ $SS_E = 84.654214$ $R^2 = .36518$ $R = .6043$ $r_{HP} = .884$

$r_{MP} = .655$

$b_0 = -2.22526$

$b_H = .03793$ SE of $b_H = .00311$ $t_{453} = 12.196^{**}$

$b_M = .00208$ SE of $b_M = .00027$ $t_{453} = 7.571^{**}$

Using all X variables: k=3 $\hat{Y}=b_0+b_V X_V+b_M X_M+b_H X_H$ (3)

SE_Y= .43266 SS_E= 84.611993 R^2= .36550 R= .6046 r_{HP}= .883

r_{MP}= .655

r_{VP}= .465

b_0= -2.24247

b_H= .03760	SE of b_H = .00318	t_{452}= 11.794**	b_H'= .46444
b_M= .00202	SE of b_M = .00030	t_{452}= 6.685**	b_M'= .28245
b_V= .00014	SE of b_V = .00030	t_{452}= .477	b_V'= .02050

Best prediction equation is (2) :

$$\hat{Y} = -2.22526 + (.00208)X_M + (.03793)X_H$$

with SE_Y= .43229

Simplification (See page 41)

Round b_V, b_M to 3; b_H to 2 dec. (No error in $\hat{Y}$ calculated)

$$(2)X_M + (40)X_H - 2{,}200 = (1{,}000)\hat{Y}$$

Discussion of Prediction Equations

One consistant result of the regression analyses undertaken is that high school rank is by far the most efficient predictor, while verbal score makes comparatively small contribution to the over-all prediction. In some cases the reliability of the prediction is actually worsened by the inclusion of verbal score. In both the first quarter and year QCA female analyses this is so; in fact, for the female year QCA, math. score

should also be omitted to get the best least squares prediction line, and in the first quarter b_M is not significant, even at the .05 level.

For the all-males first quarter QCA analysis the standard error of y is virtually unchanged by the inclusion of verbal score, the estimated coefficient of X_V is not significant at the .05 level, and the increase in R^2 is almost negligible. These facts would justify the omission of the variable from the prediction equation, and were the equation to be used for many future predictions, the omission would be recommended, as the extra calculation would produce no closer prediction. However, predictions for the sample admissions themselves were more easily produced using all three variables, and so the predicted QCA's used in the subsequent investigations are values based on the three X variables, even where X_V is redundant.

For the male year QCA the value of b_V is small but significant at the .05 level. This positive contribution of verbal score appears to come from the civilian admissions rather than the cadet, as shown by the two seperate analyses. Prediction for the cadets is not improved at all by inclusion of verbal score, whereas in the civilian prediction equatiom it has a significant coefficient at the .01 level, increases the R^2, and decreases the standard error.

The correlations between each predictor and the predicted QCA show the relative importance of each predictor to the over-all prediction. Obviously, when only one X variable is being used, its correlation with the prediction is 1.00, but even when all three X variables are included, the correlation between rank and prediction exceeds .94 for the females,

and .84 for the males. The unimportance of verbal score is reflected in correlations of .55 and below; values slightly above this correspond to cases where verbal score does contribute to the prediction. Math. score correlations with prediction range from .66 to .78 for males, but are lower for females. The special homogeneous cadet group yields the lowest values of any male group for correlations between verbal score and prediction, and between math. score and prediction.

Table 9 summarizes the R^2 values and standard errors for all the analyses at the inclusion of each variable. It shows that the three variables used are better predictors of the over-all year's performance than of the first quarter's results, in the sense that standard errors are less, that is, predictions are made with a little more certainty. In the same sense, the cadet prediction equation is more efficient than the civilian one, which is not too surprising as the civilians are a somewhat heterogeneous group. The R^2 is a measure of the proportion of variation in the y's which is accounted for by the regression on the X's. All the analyses give relatively low values for R^2, but again, the cadet equation is seen to be a little higher than the civilian. Results for the special homogeneous cadet group are very similar to those for all cadets, except that the standard error is reduced slightly.

Simplification of the Equation

The information gained from the foregoing analyses is interesting and maybe valuable in comparing the different groups of admissions, but one main object of this survey was to construct an equation which could

Table 9. Summary of R^2 values and standard errors for each analysis at the inclusion of each X-variable.

	First Quarter		Year	
Females	R^2	SE_y	R^2	SE_Y
With X_H	.32446	.56521	.31563	.50111
With X_H & X_M	.36718	.55309	.32475	.50326
With X_H, X_M & X_V	.36774	.55909	.32490	.50889
Males				
With X_H	.23791	.63005	.26774	.50821
With X_H & X_M	.33622	.58830	.34134	.48223
With X_H, X_M & X_V	.33701	.58824	.34437	.48135
Civilians				
With X_H			.23980	.54794
With X_H & X_M			.32465	.51731
With X_H, X_M & X_V			.33996	.51226
Cadets				
With X_H			.28833	.48824
With X_H & X_M			.36604	.46114
With X_H, X_M & X_V			.36688	.46116
Special Cadet Group				
With X_H			.28486	.45832
With X_H & X_M			.36518	.43229
With X_H, X_M & X_V			.36550	.43266

be recommended for future prediction purposes, and this now presents many problems.

It would not be difficult for a clerk, with the aid of a desk calculator, to use one of the equations in the form given, that is, with coefficients of five decimal places, but it is a little cumbersome, and an admissions officer may wish to have it in a form which would enable him to do a quick mental calculation. Accordingly, some investigations were made into the problem of simplification and possible loss of accuracy in predictions. It will be seen later that, in view of the large standard errors involved, concern over rounding errors is purely academic. However, it raises some practical points which may have relevance to any future survey.

One difficulty arises from the use of the computer and the prepared program with variables of such different sizes. The program instructs the machine to work with eight significant figures, regardless of the original number size, and larger numbers are cut off without rounding at eight figures. This does mean that all variables have results with the same relative accuracy, and in general, these results would be correct to six significant figures. However, in printing the results in final form the program allows for five decimal places for all numbers of the order 10^{-3} or larger, but prints numbers smaller than 10^{-3} in floating point form.

Since X_V and X_M values are approximately five hundred times as large as the Y variable, the corresponding coefficients are necessarily very small and, as a result, only three significant figures are printed out.

The remaining five figures were found by printing out and deciphering intermediate output from the computer, and it was clear that all eight figures had been used by the machine in calculating the predicted values for the sample admissions. The possible discrepancies arising from use of the coefficients cut off at five decimal places are demonstrated below for the year civilian equation.

The complete coefficients as used by the computer are: $b_0 = -1.9301174$, $b_V = .0010113420$, $b_M = .0016586687$, and $b_H = .03087524$. Using the printed coefficients, the error in the predicted Y for average X values, (445, 524, 56), is five in the third decimal place, or for an extreme case, ($X_V = 688$, $X_M = 787$, $X_H = 76$), is eight in the third decimal place. If the coefficients had been rounded, rather than cut off, to five decimal places, the error in the Y prediction would be no more than one in the third decimal place.

Since the five figures given are not corrected anyway, it seems sensible to suggest rounding to four decimals. This would involve errors of maximum magnitude five in the second decimal, while a further approximation of b_H to three decimals does not increase this error because b_H is multiplied by numbers ten times less than X_V and X_M. There would be no point in expressing b_0 to greater accuracy than two decimals, and so the equation could be written:

$$(10)X_V + (16)X_M + (310)X_H - 19{,}300 = (10{,}000)\hat{Y}$$

Although less formidable to the layman than the original, this equation may still present difficulties to quick calculations, and tempts even greater approximations. Rounding of b_V and b_M to three decimals,

and b_H to two, would produce maximum errors of two in the first decimal place, and give the prediction in the form:

$$X_V + (2)X_M + (30)X_H - 1{,}900 = (1{,}000)\hat{Y}$$

Simplification of each of the other prediction equations requires investigation of the specific errors involved, and this has been carried out for the male year QCA analysis and the cadet year QCA analysis, as well as for the civilian case described above. A summary of simplification is given after the appropriate analysis result. The simplest forms could only be used as guides, since if a prediction may be .2 off either way, for example, and a dividing line between admission and non-admission were set at 1.0, there would be considerable danger of misclassification. The equation for females is based on too few numbers to be reliable for prediction in any form.

Expectancy Tables

As pointed out by Duggan and Hazlett, there are many sources of error in predicting college performance. We have seen that, at most, we can account for 37 percent of the variation among the 1961 V.P.I. admissions by using College Board scores and high school rank. We wish then to make some statement about the magnitude of error in our predictions, arising from the nature of our variables and samples.

An assumption basic to the validity of these multiple regression analyses is that, for any set of X's, the Y's are distributed normally with variance, σ_E^2, and mean, $\mu_Y = \beta_0 + \beta_V X_V + \beta_M X_M + \beta_H X_H$. Now $\hat{Y} = b_0 + b_V X_V + b_M X_M + b_H X_H$ is an unbiased estimate of μ_Y, and SE^2, the standard error of estimate, is an unbiased estimate of σ_E^2.

Following the workbook, we make the further assumption that, our sample being sufficiently large, $\hat{Y}$ and SE^2 are accurate estimates of μ_Y and σ^2, and have no sampling distribution. In considering future samples of admissions, or samples other than the one used, we may select students with a definite set of X values and, relating them to the corresponding Y probability distribution, may make statements such as, "a proportion, p, of these are expected to get QCA's greater than Y_0". In terms of the individual, this may be expressed as, "the probability of this student getting a QCA higher than Y_0 is p".

We are not setting confidence limits on a prediction, nor yet on the estimate of μ_Y: this may be done, but involves results not easily accessible from the machine calculations, and in the case of individual predictions, is impractical for more than one or two. The method used is sufficient to describe the uncertainty involved when a prediction is made, and to caution the administrator against indiscriminant interpretation.

To save admissions officers the task of using the Normal Curve table for each prediction, Duggan and Hazlett suggest the construction of an expectancy table, which presents the required probabilities for certain values of Y, namely, at intervals of half the standard error, (½SE). These are easy values for which to use the Normal table, but not very systematic for reference, For the range of Y values in these analyses, 0.00 to 3.00, intervals of .25 are suggested here, and for m= 0,1,..,12 and n= 0,1,..,12 , we calculate probabilities that a student with a predicted Y value of (0.00 + (.25)m) will get an actual QCA of (0.00 + (.25)n) or higher. Expectancy tables for each analysis have

Expectancy table 1.10 First Quarter QCA for Females

Chances in 100 that a student will obtain a first quarter QCA equal to or higher than:

Predicted freshman first quarter QCA	0.000	0.276	0.552	0.828	1.104	1.380	1.656	1.932	2.208	2.484	2.760	3.0+
3.0+							99	98	93	84	69	50
2.760						99	98	93	84	69	50	31
2.484					99	98	93	84	69	50	31	16
2.208				99	98	93	84	69	50	31	16	7
1.932			99	98	93	84	69	50	31	16	7	2
1.656		99	98	93	84	69	50	31	16	7	2	1
1.380	99	98	93	84	69	50	31	16	7	2	1	
1.104	98	93	84	69	50	31	16	7	2	1		
0.828	93	84	69	50	31	16	7	2	1			
0.552	84	69	50	31	16	7	2	1				
0.276	69	50	31	16	7	2	1					
0.000	50	31	16	7	2	1						

Expectancy table 1.20 First Quarter QCA for Males

Chances in 100 that a student will obtain a first quarter QCA equal to or higher than:

Predicted freshman first quarter QCA	0.000	0.294	0.588	0.882	1.176	1.470	1.764	2.058	2.352	2.646	2.940	3.0+
3.0+							99	98	93	84	69	50
2.940						99	98	93	84	69	50	31
2.646					99	98	93	84	69	50	31	16
2.352				99	98	93	84	69	50	31	16	7
2.058			99	98	93	84	69	50	31	16	7	2
1.764		99	98	93	84	69	50	31	16	7	2	1
1.470	99	98	93	84	69	50	31	16	7	2	1	
1.176	98	93	84	69	50	31	16	7	2	1		
0.882	93	84	69	50	31	16	7	2	1			
0.588	84	69	50	31	16	7	2	1				
0.294	69	50	31	16	7	2	1					
0.000	50	31	16	7	2	1						

Expectancy table 2.10 Year QCA for Females

Chances in 100 that a student will obtain a year QCA equal to or higher than:

Predicted year QCA	0.000	0.250	0.500	0.750	1.000	1.250	1.500	1.750	2.000	2.250	2.500	2.750	3.0+
3.000								99	98	93	84	69	50
2.750							99	98	93	84	69	50	31
2.500						99	98	93	84	69	50	31	16
2.250					99	98	93	84	69	50	31	16	7
2.000				99	98	93	84	69	50	31	16	7	2
1.750			99	98	93	84	69	50	31	16	7	2	1
1.500		99	98	93	84	69	50	31	16	7	2	1	
1.250	99	98	93	84	69	50	31	16	7	2	1		
1.000	98	93	84	69	50	31	16	7	2	1			
0.750	93	84	69	50	31	16	7	2	1				
0.500	84	69	50	31	16	7	2	1					
0.250	69	50	31	16	7	2	1						
0.000	50	31	16	7	2	1							

Expectancy table 2.20 Year QCA for Males

Chances in 100 that a student will obtain a year QCA equal to or higher than:

Predicted year QCA	0.000	0.250	0.500	0.750	1.000	1.250	1.500	1.750	2.000	2.250	2.500	2.750	3.0+
3.000					100		99.9	99.5	98	94	85	70	50
2.750				100		99.9	99.5	98	94	85	70	50	30
2.500			100		99.9	99.5	98	94	85	70	50	30	15
2.250		100		99.9	99.5	98	94	85	70	50	30	15	6
2.000	100		99.9	99.5	98	94	85	70	50	30	15	6	2
1.750		99.9	99.5	98	94	85	70	50	30	15	6	2	0.5
1.500	99.9	99.5	98	94	85	70	50	30	15	6	2	0.5	0.1
1.250	99.5	98	94	85	70	50	30	15	6	2	0.5	0.1	
1.000	98	94	85	70	50	30	15	6	2	0.5	0.1		0
0.750	94	85	70	50	30	15	6	2	0.5	0.1		0	
0.500	85	70	50	30	15	6	2	0.5	0.1		0		
0.250	70	50	30	15	6	2	0.5	0.1		0			
0.000	50	30	15	6	2	0.5	0.1		0				

Expectancy table 2.21 Year QCA for Civilians

Chances in 100 that a student will obtain a year QCA equal to or higher than:

Predicted year QCA	0.000	0.250	0.500	0.750	1.000	1.250	1.500	1.750	2.000	2.250	2.500	2.750	3.0+
3.000					100		99.8	99.3	97	93	84	69	50
2.750				100		99.8	99.3	97	93	84	69	50	31
2.500			100		99.8	99.3	97	93	84	69	50	31	16
2.250		100		99.8	99.3	97	93	84	69	50	31	16	7
2.000	100		99.8	99.3	97	93	84	69	50	31	16	7	3
1.750		99.8	99.3	97	93	84	69	50	31	16	7	3	0.7
1.500	99.8	99.3	97	93	84	69	50	31	16	7	3	0.7	0.2
1.250	99.3	97	93	84	69	50	31	16	7	3	0.7	0.2	
1.000	97	93	84	69	50	31	16	7	3	0.7	0.2		0
0.750	93	84	69	50	31	16	7	3	0.7	0.2		0	
0.500	84	69	50	31	16	7	3	0.7	0.2		0		
0.250	69	50	31	16	7	3	0.7	0.2		0			
0.000	50	31	16	7	3	0.7	0.2		0				

Expectancy table 2.22 Year QCA for Cadets

Chances in 100 that a student will obtain a year QCA equal to or higher than:

Predicted year QCA	0.000	0.250	0.500	0.750	1.000	1.250	1.500	1.750	2.000	2.250	2.500	2.750	3.0+
3.000						100		99.7	98	95	86	71	50
2.750					100		99.7	98	95	86	71	50	29
2.500				100		99.7	98	95	86	71	50	29	14
2.250			100		99.7	98	95	86	71	50	29	14	5
2.000		100		99.7	98	95	86	71	50	29	14	5	2
1.750	100		99.7	98	95	86	71	50	29	14	5	2	0.3
1.500		99.7	98	95	86	71	50	29	14	5	2	0.3	
1.250	99.7	98	95	86	71	50	29	14	5	2	0.3		0
1.000	98	95	86	71	50	29	14	5	2	0.3		0	
0.750	95	86	71	50	29	14	5	2	0.3		0		
0.500	86	71	50	29	14	5	2	0.3		0			
0.250	71	50	29	14	5	2	0.3		0				
0.000	50	29	14	5	2	0.3		0					

Expectancy table 2.23 Year QCA for Cadets, Excluding Summer School, Drop-Out, and Cooperative Students, and Partial Rank Information.

Chance in 100 that a student will obtain a year QCA equal to or higher than:

Predicted year QCA	0.000	0.250	0.500	0.750	1.000	1.250	1.500	1.750	2.000	2.250	2.500	2.750	3.0+
3.000						100		99.8	99	96	88	72	50
2.750					100		99.8	99	96	88	72	50	28
2.500				100		99.8	99	96	88	72	50	28	12
2.250			100		99.8	99	96	88	72	50	28	12	4
2.000		100		99.8	99	96	88	72	50	28	12	4	1
1.750	100		99.8	99	96	88	72	50	28	12	4	1	0.2
1.500		99.8	99	96	88	72	50	28	12	4	1	0.2	
1.250	99.8	99	96	88	72	50	28	12	4	1	0.2		0
1.000	99	96	88	72	50	28	12	4	1	0.2		0	
0.750	96	88	72	50	28	12	4	1	0.2		0		
0.500	88	72	50	28	12	4	1	0.2		0			
0.250	72	50	28	12	4	1	0.2		0				
0.000	50	28	12	4	1	0.2		0					

been constructed with scales in intervals of ($\frac{1}{2}$SE) for first quarter predictions, (1.10, 1.20), and scales in intervals of .25 for year predictions, (2.10, 2.20, 2.21, 2.22, 2.23).

Thus, suppose a male applicant has a predicted year QCA of .75, and the admission registrar wishes to know what risk is involved in admitting him. Reference to table 2.20 shows that the boy has a chance of 30 in 100 of attaining the minimum year QCA of 1.000. Suppose now that this candidate was definitely eligible for the Corps, had not attended summer school, was not to be a Cooperative, and had full information on rank. Then expectancy table 2.23 might be used, and the boy's chance of getting 1.000 or over is reduced slightly to 28 in 100. For another example, consider a married candidate with a predicted QCA of 2.30. Is there any question about his success? Using the civilian table, 2.21, it is seen that his chances of obtaining a year QCA of 1.000 or more lies between 99.3 and 99.8. So there is only less than one in a hundred probability that he will not meet requirements, and he could certainly be considered a good risk academically.

Discussion of Prediction Errors

It has been noted that even the largest R^2 value obtained is low, and that the standard errors of estimate are high. In such a study as this we expect large errors but comparison with the Fordham study shows that V.P.I. standard errors are relatively almost twice as large.

Taking 60 to be the zero of Fordham grade scale, the coefficient of variation (CV) is 4.51/16.65 = .27. For the cadet results here,

CV = .46116/1.05217 = .44 , and pulling out the irregular admissions, (summer school, drop-outs, etc.) only reduces the CV to .40 .

A more fair comparison may be the standard error expressed as a percentage of the range of the scale. Thus for Fordham, 4.51/40 = 11.3%; for cadets at V.P.I., .46116/3.000 = 15.4% ; and for cadets exclusive of irregulars, .43266/3.000 = 14.4% .

Either way, the present results compare unfavorably with the sample study. It may be that the variation unaccounted for is entirely due to individual differences, response to college work and new environment, teacher-pupil relationship, motivation and other personal problems. These may lead to greater variation than for a group of admissions to Fordham, because V.P.I. may admit a more diverse section of candidates. If this were so, and we could postulate a stability of situation from year to year, (no large scale change in policy of admission, grading or student administration), then there may be some value in the cautious application of the obtained prediction equations, since no better ones could be found.

It is possible, however, that there are some other measurable factors which bear on a student's college performance. Results show, for instance, that civilians and cadets perform differently as groups. If one prediction equation were to be used for both, then this status should be introduced as a new variable. Other possible sources of variation are place of residence (dormitory of town), marital status, veteran or non-veteran, home state (Virginia or other). Factors such as these may have been discounted elsewhere as having no predictive value, but could

be problems peculiar to a large technical college, predominantly male and military, in a small rural town. Even if we agree to lump such possible variation under individual response, we should still justify our hypothesis of a stable situation before predicting with any confidence at all.

Choice of Equation

There would then be one further problem concerning the choice of the most suitable equation. Choice between the all-male equation or separate civilian and cadet ones is really a problem for the admissions officer. If he is sure, in advance, what the status of the student would be, the appropriate equation would be more valuable, but in absence of this prior knowledge, the all-male equation would have to be a substitute.

It is the job of the statistician, however, to decide whether it is more ethical to base an equation on as homogeneous a group as possible, or to include obvious misfits such as drop-outs, trial summer students, and Cooperatives. Each of these may be a different problem, but it is useful, first of all, to investigate whether their performance is, in fact, different from that of the ordinary students. This is done, to a limited extent, in the next chapter. The problem for drop-outs and Cooperatives, and in this case, summer school students, is that the Y value for them is not measuring true year QCA. For drop-outs, particularly, we would like to overcome this, as their withdrawal may intercept a potentially low grade, and an average of their QCA to date may overestimate their "year's" performance.

Prediction for the Sample Admissions

The last computer step in each analysis was the calculation of predicted QCA for each student included in that sample, together with the residuals of actual minus predicted. As mentioned earlier, prediction was carried out using all three X variables, with coefficients to eight significant figures. Actual and predicted performances were then used to compare various subgroups, in an effort to isolate other sources of variation, and the next chapter is devoted to these investigations.

IV

VARIATION OF SAMPLE SUBGROUPS

Initial Disparity of Status and School Groups

It has been suggested that freshmen in the Corps tend to have lower grades than their civilian classmates. Accordingly, the survey was designed to test this theory, and to discover all the possible differences between them. It has already been noted that the two groups differ in their initial make-up; the civilian group is composed of all successful candidates who are exempt from the Corps for one of many reasons. Until they become civilian students at V.P.I. these candidates have nothing in common as a group. The freshmen who enter as cadets, on the other hand, are more likely to be within a narrow age range, non-veteran, straight out of high school, unmarried and with no physical disability, in fact, a reasonably homogeneous group. Both groups draw on applicants from Virginia and from other states, but not in equal proportions. Figures given in table 1 show that 24 percent of cadets are from Out-of-State, compared with only 11 percent of the civilians. A one degree of freedom chi-square test shows this to be a significant difference, and so it is as well to consider Virginia - Out-of-State differences along with civilian - cadet investigations. The four groups will be referred to by the following self-explanatory abreviations; CivVa., CivOS., CadVa., and CadOS..

Admission policy is partly responsible for the group differences in qualifications. Minimum requirements are more strictly enforced for Out-of-Staters, and if College Board scores are given as much consideration

as high school rank, this may account for the higher averages of Out-of-State admissions on verbal and math. scores. One reason the ranks of Out-of-State applicants are not correspondingly higher than ranks for Virginians is that competition may be greater in Out-of-State schools. Civilian College Board scores may be weighted by some older applicants who, taking the tests prior to entry, could not perform as well as recent school-leavers, although their school records were as good.

However debatable these rationalizations of differences may be, the actual differences must be taken into account when comparisons of QCA's are made among the groups. This can only be done by trusting the prediction equations to the extent of using predicted values as true standards against which to measure the actual QCA's obtained.

Variation of Actual With Predicted Performance

The differences between actual and predicted QCA's may be assessed for a group in various ways. The degree of agreement may be expressed by a correlation coefficient, or the differences tested with a paired t-test. The group may be described by the proportion of its number which have actual QCA's higher than predicted, and vice versa. Finally, a dividing QCA may be chosen, say 1.000, and the group divided into students with predicted and actual QCA's less than 1.000, students with predicted less than 1.000 and actual 1.000 or higher, and so on.

The last three methods have been used to compare civilians and cadets, and Virginians and Out-of-Staters. Results are summarized in table 10, and tests of group differences discussed below.

Paired t-tests Since the sample under investigation is the same as that used to construct the prediction equation, the mean difference between actual and predicted QCA's over all the sample is obviously zero. For the same reason, any large subgroup of the sample, such as cadets, will have weighted the predictions to the extent that residuals of actual minus predicted QCA's for that group may average out to near zero.

In fact, using one all-male equation, the mean residuals are .094 for civilians and -.041 for cadets, and the Student t-values on a paired test are respectively, 3.204, significant at the one percent level for 304 degrees of freedom, and -2.366, significant at the five percent level for 706 degrees of freedom. This means that, in spite of the tendency towards zero of the residuals, we can say that actual performance is better than predicted for the civilian group, and worse than predicted for the cadet group. The same is true for Virginians only in each group, but in the Out-of-State subgroups paired t-tests between actual and predicted were not significant.

Positive and Negative Residuals A count of admissions in CivOS with actual QCA's greater than predicted shows that 64 percent of them fall into this category, compared with only 54 percent for the CivVa group, but when tested with a one degree of freedom chi-square, no significance is established for the difference. A small difference is detected between Virginians and Out-of-Staters in the cadet group. Here percentages with actual QCA's exceeding predicted are 40 and 51 respectively, and the probability of the chi-square value is between .025 and .01, that is, the difference is significant at the five percent level.

Table 10. Summary of predicted and actual QCA characteristics by status and school.

Group characteristic	CivVa	CivOS	CivNK	CadVa	CadOS	CadNK
Number (N)	269	34	2	539	167	1
Mean QCA:						
Actual	1.072	1.324	.651	1.021	1.155	.673
Predicted	.985	1.166	.670	1.069	1.171	1.084
Residual	.087	.149		-.048	-.018	
SD of Resid.	.479	.617		.452	.488	
t_{N-1}	2.868**	1.449		-2.465*	-.478	
Mean Resid.	.094			-.041		
SD of Resid.	.512			.461		
t	$t_{304} = 3.204$**			$t_{706} = -2.366$*		
By All-male equation:						
Act.> Pred.	146	23		216	86	
Act.< Pred.	123	13		323	82	
By separate equations:						
Act.> Pred.	124	20		237	92	
Act.< Pred.	145	14		302	75	
Pred,Act group						
00	95 (35.3)	3 (8.8)	2	199 (35.4)	31 (18.6)	0
01	40 (14.9)	7 (20.6)	0	42 (7.8)	17 (10.2)	0
10	42 (15.6)	6 (17.6)	0	107 (19.8)	34 (20.4)	1
11	92 (34.2)	18 (52.9)	0	199 (36.9)	85 (50.9)	0

It seems that, even allowing for their higher initial qualifications, Out-of-Staters do perform better than expected in relatively more instances than Virginians, but that the effect is masked in the civilian group by the small size of the CivOS group. It may be that a student who does much better than expected by the overall standard because he is a civilian will not have an even higher standard because he is also from Out-of-State. This sort of thing is, in effect, an interaction between school and status, and thus status differences may be shown up better by consideration of separate civilian and cadet predictions.

Positive residuals resulting from the separate equations were counted, and the new proportions in the appropriate Virginia and Out-of-State subgroups tested with chi-squares. The tests gave the same answer as in the one equation situation, non-significance for the civilians and significance at the five percent level for the cadets. Examination of percentages in the subgroups of students with actual performance better than predicted shows they are now very similar for civilians and cadets, 46 and 54 compared with 44 and 55. This does suggest that there is a real, though maybe small, difference between Virginians and Out-of-Staters on freshman performance, which affects cadets and civilians alike, but that the CivOS group is so small that the difference cannot be determined as statistically significant, for the civilians.

1.000 QCA Division There are many different tests which could be done on the final set of results in table 10, where admissions are classified into one of four groups, 00, 01, 10 or 11, 0 denoting a QCA less than 1.000, 1 denoting a QCA of 1.000 or more, and the first digit referring to predicted, the second to actual QCA.

Tests between subgroups on the predicted or actual dichotomies alone give very little information beyond previous results. Summarizing, these are that the proportion of predictions less than 1.000 is just lower for cadets than for civilians, but proportions of actual QCA's less than 1.000 is about the same for each group. Out-of-Staters have proportionally **less** predictions less than 1.000 than do Virginians, and also more actual QCA's less than 1.000. These differences are significant, and hold true for both civilians and cadets.

It was found, using three degree of freedom chi-square tests on two subgroups and four QCA categories, that QCA differences were significant between CadOS and CadVa at one percent level, between CadVa and CivVa at one percent level, and between CivVa and CivOS at five percent level. The two smallest groups, CivOS and CadOS, did not have significantly different distributions.

The percentage distributions show where the differences occur, and not surprisingly, it is found that the proportion in the two categories 01 and 10, which may be considered as the total misclassification, is highest, 38 percent, for the smallest group CivOS, and lowest, 28 percent, for the largest group CadVa. Although the total misclassification for any group tends to be inversely proportional to its size, actual subgroup characteristics do show up in the type of misclassification, mainly that cadet predictions are over-optimistic.

Using the separate civilian and cadet prediction equations instead of the one all-male equation, the total misclassification is only reduced from 29 percent to 28 percent, but the actual - predicted distributions

are somewhat altered, especially for the civilians. This group, too, now has a high proportion of misclassifications in the 10 category.

The appropriate distribution is used in the next section to assess differences from the norm of small groups such as Cooperatives and drop-outs. Further discussion of misclassification is included in the final chapter.

Investigation of Cooperative, Drop-Out and Summer School Groups

Two methods were used to see how certain special groups differed from the overall average on actual and predicted performance. Table 11 summarizes the paired t-tests and the distributions into the four actual - predicted categories, 00, 01, 10 and 11.

Since only those freshmen with good qualifications are allowed on a Cooperative basis, it is not surprising that the majority of Cooperative students fell into the 11 QCA category. There may have been some prospective Cooperatives who did not do well enough in the first quarter to qualify and so stayed in school all year. These are not included, or they might have increased the 10 group. However, those who did complete the program did even better in college work than expected, whether predictions are made from the all-male equation or from the separate civilian and cadet equations. It may be argued that it is easier to sustain a high QCA over one or two quarters than over the whole year, but differences in the first quarter and year QCA averages for the total sample are very slight compared with the positive residuals averaged by the Cooperatives.

Table 11. Summary of predicted and actual QCA characteristics for Cooperative, drop-out, and summer school groups.

Group characteristics	Cooperative			D-O	D-O	No D-O	No D-O
	All	Civ.	Cad.	Trial SS	No SS	Vol SS	Trial SS
N	53	10	43	21	87	23	82
By All-male equation:							
Mean (A-P)	.327			-.215	.307	.092	.052
SD of Resid	.416			.493	.545	.472	.374
t_{N-1}	5.714**			2.002	-5.246**	.938	1.261
By separate equations:							
Mean (A-P)		.550	.294				
SD of Resid		.293	.244				
t_{N-1}		5.930**	7.904**				
By All-male equation: P/A group							
00	0			16 (76.2)	31 (35.6)	9 (39.1)	57 (69.5)
01	2 (3.8)			2 (9.5)	2 (2.3)	2 (8.7)	14 (17.1)
10	2 (3.8)			2 (9.5)	31 (35.6)	3 (13.0)	3 (3.6)
11	49 (92.4)			1 (4.8)	23 (26.4)	9 (39.1)	8 (9.8)
By separate equations: P/A group							
00		0	0				
01		0	4 (9.3)				
10		0	2 (4.6)				
11		10	37 (86.0)				

N.B. (A-P) = Resid.

Drop-out and summer school groupings overlap and so the drop-outs are considered in two groups, those who attended summer school on trial, and others. The first group were expected to do badly and did not exceed expectations, in fact, the mean residual is very negative, but no statistical significance can be attached to it for such a small group.

The larger section of drop-out students, the 87 who had not attended summer school, had about the same ratio of 0 and 1 predictions as did the cadets by the all-male equation, but they only had two percent in the 01 category, and a high proportion (36 percent) in the 10 category. By the paired t-test, their actual performance was indisputably worse than predicted. It is college policy to retain a student on trial for his freshman year even though he has very low grades, but more drop-outs occur in the low QCA groups, and results suggest that students tend to withdraw when their college performance falls much below expectation.

In the small group of students who had attended summer school voluntarily proportions above and below 1.000 for actual and predicted QCA's were not remarkable, nor was the mean residual significant.

The majority of students who had attended summer school on trial entered with low predictions, but excluding the drop-outs already discussed, the actual QCA's of the others averaged a little higher than predicted, although the bulk of the group was still in the 00 category.

It may be concluded that Cooperative, trial summer school and drop-out freshmen do perform differently from the norm, and should probably be separated from any sample used to construct a prediction equation. Multiple regression predictions and administration agreed to the extent

that 21 percent of the "bad" predictions were asked to attend summer school, and only two percent of the "good" predictions. Finally, in general, drop-outs after the first or second quarters occurred among students whose QCA's were lower than predicted.

Investigation of Curriculum Groups

There is very little basis for suspecting that curricula groups might differ from one another in their relationship of actual to predicted performance. Freshmen in all departments are required to take similar programs, and so are subjected to teaching and grading from most other departments besides their own. Business majors, for example, may find the compulsory mathematics courses difficult, but this is because students with high mathematical ability are not likely to enter the Business curriculum, and this will be reflected in predicted as well as in actual grades. In general, a student's interest and ability, together with guidance from placement officers, will lead him to the curriculum where he knows that, after freshman year, he will be best suited. By the same argument, the few course hours which differ between departments should present the same relative challenge to the students concerned. There are a few possible exceptions; for example, Agriculture, including Forestry and Wildlife, students are required to take chemistry, which many of them may never have had in school. Performance on this is not included in their predicted grade, and unlike Engineering or Chemistry majors, they are not likely to excel in the subject. Another example is within the School of Engineering where all Engineers have the same

freshman program, with sixteen hours of chemistry. It is possible that the majority of Chemical Engineering prospectives have already some interest and tuition in the subject and will do better than their Civil Engineering classmates, for example, whose actual QCA's may be lowered by a mark in a subject which played no part in their predicted QCA's.

This is suggesting that the three variables used for prediction are not tests of general ability or intelligence, but only markers of mathematics, English and the subjects emphasized in high school. What the variables really measure is certainly a matter for discussion, and this is one argument for investigating curricula differences, in spite of the difficulties. This, however, can only be an incidental side-line of the present survey which has insufficient information collected to deal fully with the matter.

The questions of whether students really do choose their curricula wisely, and receive sufficient placement guidance, and whether the wrong choice of curriculum at onset jeopardizes their eventual progress, are also beyond the range of this data analysis. Initial curriculum data on first quarter QCA may be relatively free from confusion with transfers, but then neither does it give any information about them. Similarly, year QCA's, grouped by third quarter curriculum, include the whole record of the transfers in their new curricula, but give no separate analysis of them.

A significant difference, by one measure or another, between predicted and actual performance in any one curriculum group may be interpreted, then, as unwise choice or bad placement, on the one hand, or inefficient predictions, on the other. The two causes may act together

on first quarter results and curriculum break-down, but the year analysis with third quarter curriculum break-down, is possibly a fair indication of the extent of the latter, since by this stage, misfits will mostly all have transferred and will have made up the courses required in their new curriculum.

Another difficulty is the arbitrary grouping of the curricula. This was done partly on the basis of numbers; for instance, each Engineering curriculum with a large number of admissions was kept as a separate group, while the varied majors, such as English and Physics, in Science and General Studies were grouped together. Upon the advice of the Admissions Director, curricula such as Forestry and Wild Life and Distributive Education were investigated as seperate groups.

It was found that, when the sample was divided by status, school and curriculum, most of the resulting groups were too small to give much useful information. Observations have already been made about relative sizes of average scores and ranks in the groups, and for the record, mean actual and predicted scores are presented in tables 12 and 13, first quarter results for initial curriculum break-down, and year results for third quarter curriculum break-down. Paired t-tests were done to test each actual against predicted, but the few significances shown were clearly just reflections of the characteristics of the status-school group within which they were found.

Correlation coefficients between actual and predicted were calculated for the year results. All the coefficients were around .5, and were significant for the larger groups only. The correlations were not improved when separate prediction equations were used.

Table 12. Actual and predicted means, and significant t-values for first quarter QCA by first quarter curriculum, school, and status.

Curriculum	CivVa		CivOS		CadVa		CadOS	
	Act.	Pred.	Act.	Pred.	Act.	Pred.	Act.	Pred.
Agri-culture	1.177	.990	2.118	1.665	.656	.734	1.250	.662
Forestry & Wildlife	.532	.684	-	-	.746	1.021	.511	.739
Business	.708	.619	1.038	.859	.594	.703	.562	.680
Aerospace Engin.g	1.391	1.098 t_{17}= +2.679*	1.500	1.508	1.201	1.244	1.300	1.362
Chemical Engin.g	1.526 t_{21}= +3.736**	1.146	3.000	1.424	1.363	1.240	1.229	1.384
Civil Engin.g	.886	1.044	1.312	1.217	1.082	1.117	.929	1.126
Electrical Engin.g	1.371 t_{41}= +2.616*	1.101	1.968	1.405	1.068	1.146	1.513	1.428
Mechanical Engin.g	1.200	1.071	1.511	1.001	1.080	1.139	1.280	1.247
Other Engin.g	1.009	1.039	2.625	1.650	.847 t_{36}= -2.557*	1.039	1.339	1.362
Arch-itecture	1.133	.958	1.448	1.179	1.039	1.075	1.120	1.134
Science & General studies	1.263	1.087	1.059	1.178	1.029 t_{59}= -3.038**	1.256	1.121	1.402
Distribu-tive Education	.866 t_{2}= +6.331*	.172	-	-	.188	.136	1.000	.219

Table 13. Actual and predicted means, and significant t-values for year QCA by third quarter curriculum, school, and status.

Curriculum	CivVa		CivOS		CadVa		CadOS	
	Act.	Pred.	Act.	Pred.	Act.	Pred.	Act.	Pred.
Agri-culture	.943	.936	2.148	1.564	.756	.790	1.105	.902
Forestry & Wildlife	.766	.915	-	-	.936	1.073	.990	.781
Business	.736	.741	.953	1.081	.763	.850	.737	.783
Aerospace Engin.g	1.245	1.096	1.017	1.395	1.193	1.195	1.107	1.302
Chemical Engin.g	1.320	1.172	2.800	1.298	1.260	1.218	1.289	1.390
Civil Engin.g	.812	.976	.962	1.298	1.025	1.095	.945	1.078
Electrical Engin.g	1.197	1.052	1.622	1.324	1.000	1.123 t_{84}= +2.436*	1.482	1.294
Mechanical Engin.g	1.214	1.099	1.244	.954	1.073	1.118	1.090	1.228
Other Engin.g	1.192	1.074	2.509	1.553	.957	1.078	1.342	1.340
Arch-itecture	1.050	.921	1.556	1.062	1.116	1.055	1.325	1.110
Science & General studies	1.332 t_{28}=	1.100 +2.454*	1.006	1.148	1.196	1.180	1.266	1.228
Distribu-tive Education	.618	.205	.558	.300	.594	.425	.934	.734

The situation is represented in figure 1, where each curriculum group within CivVa, CadVa and CadOS is plotted in a graph whose axes are actual and predicted mean year QCA's. Figure 2 is a graph of the mean values resulting from selecting random groups of various sizes from the total sample. On the graph are plotted also points for the main subgroup averages. Comparison of the two graphs suggests the following remarks. In their deviation from the actual - equals - predicted line, the curriculum groups are not, in general, more widely dispersed than the randomly constructed groups of similar magnitude. (Groups with significant t values are marked.) However, the groups are spread out along the line, with most of the Out-of-State components at the upper end, CivVa and CadVa in the centre, with more CadVa above the line, (predicted greater than actual), and more CivVa below it, (actual greater than predicted). These remarks have already been supported by the foregoing analysis and discussion. In addition, we can see more clearly some trends between curricula. Low on both the actual and predicted scale are Distributive Education, Business, Agriculture, and Forestry and Wildlife. Consistantly ahead of the other corresponding Engineering groups are the Chemical Engineers, while the Civil Engineers score low, as predicted, in each group. To enable further visual comparisons, a new figure, 3, was plotted of actual versus predicted for first quarter results and curriculum break-down. Most of the above remarks hold true, except that these points are more widely scattered about the mean line.

In a final effort to isolate curriculum differences, the four school - status subgroups were combined. This was considered justified

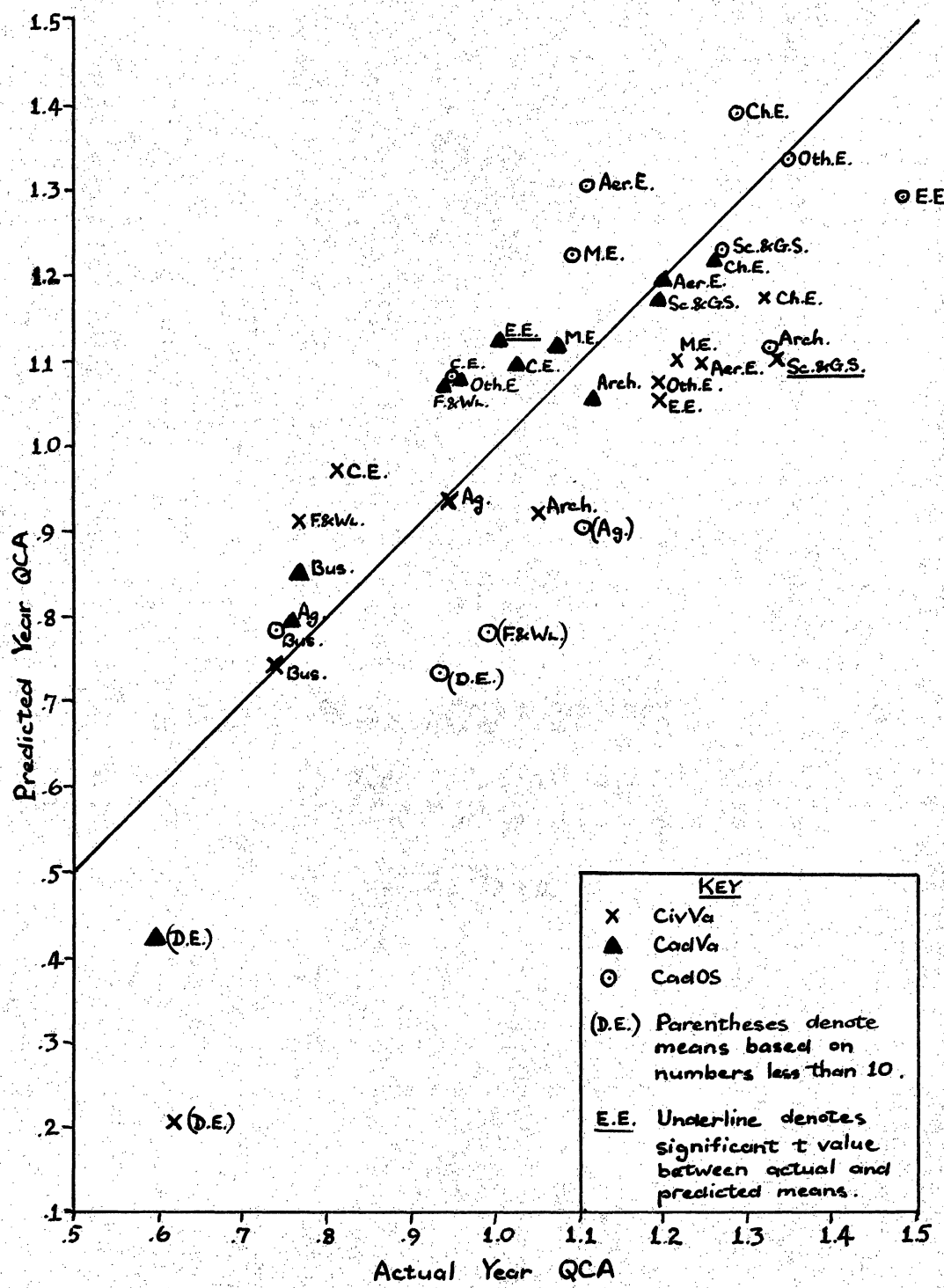

Figure 1. Graph of actual against predicted year QCA means for third quarter curriculum groups.

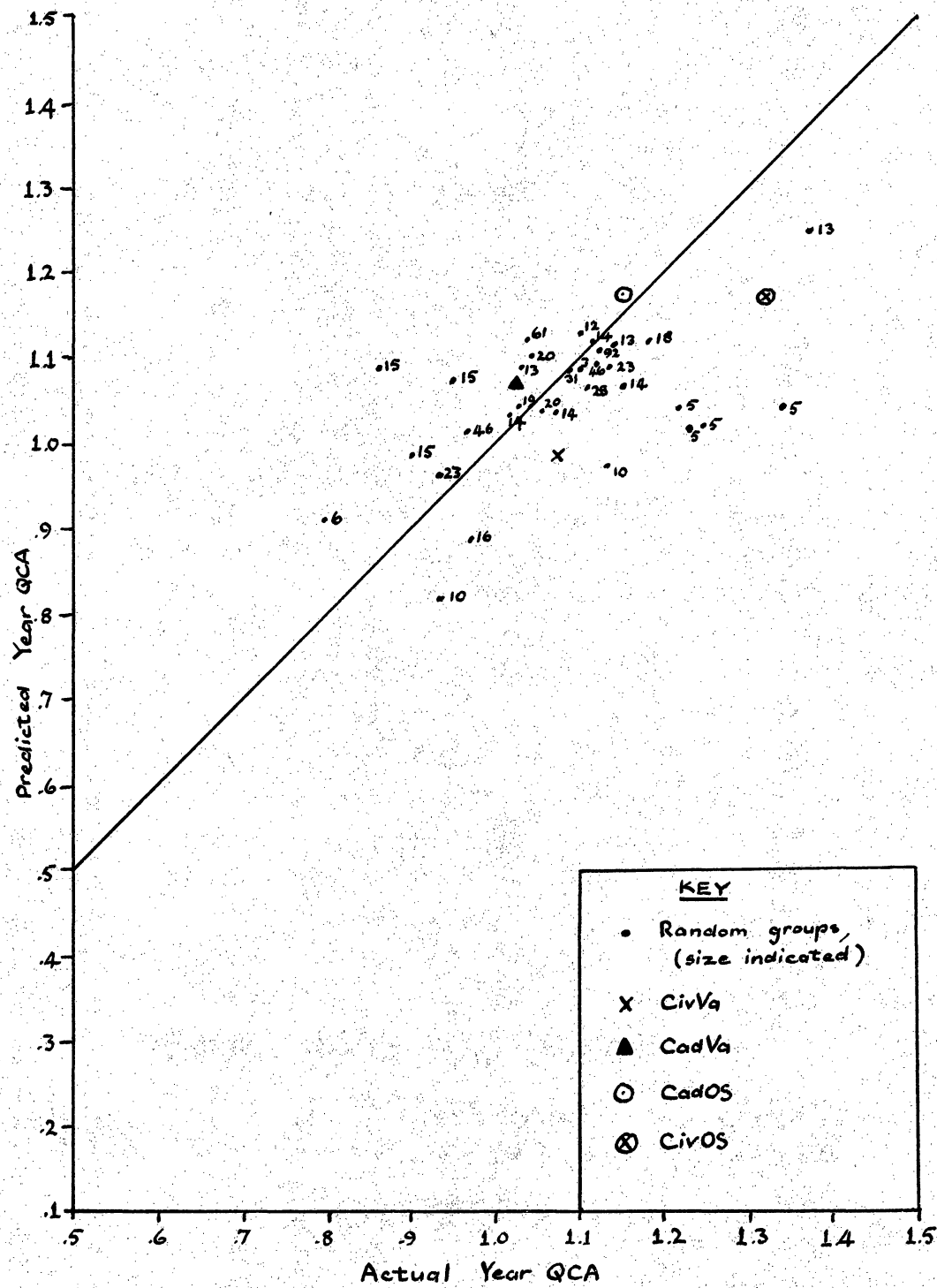

Figure 2. Graph of actual against predicted year QCA means for groups chosen at random from the total sample.

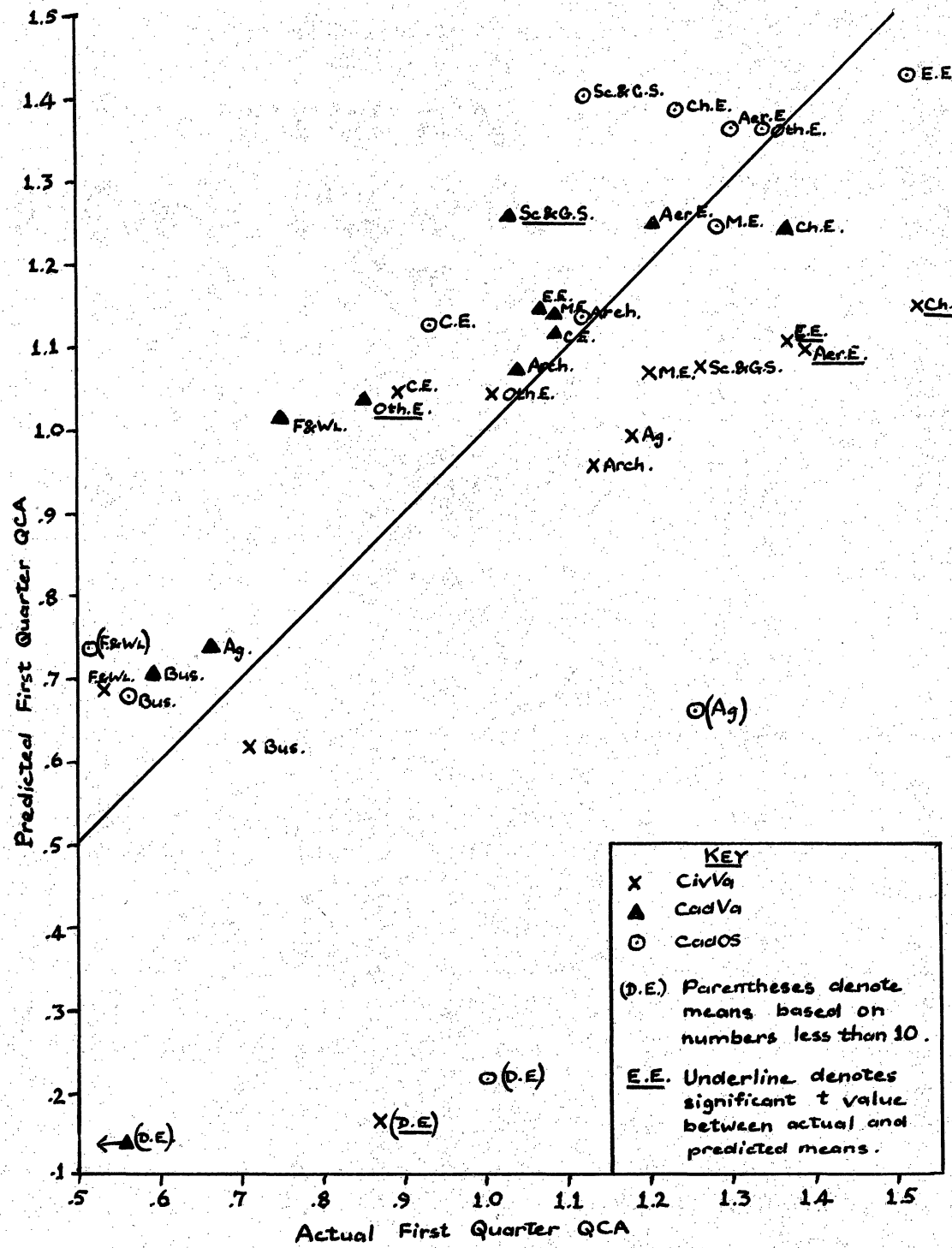

Figure 3. Graph of actual against predicted first quarter QCA means for first quarter curriculum groups.

by the very similar distributions of curricula in each subgroup. Afterwards, however, this assumed similarity was tested by comparing the distribution of subgroups in each curriculum group against the total distribution of subgroups with two by four chi-square tests. The only curricula which contained significantly different distributions were Forestry and Wildlife and Civil Engineering. Two of the expected values in the former, and one in the latter, were really too small for valid chi-square tests, but when these classes were omitted, significant chi-square values were still obtained for the remaining distributions. Thus Civil Engineers are heavily weighted with Virginian cadets, and Forestry and Wildlife have an unduly high proportion of Virginian civilians.

First quarter and year curricula were both investigated by consideration of proportions with actual QCA's greater than predicted. These figures are given in table 14, and were tested with two by two chi-square tests for each curriculum against the rest. Chemical Engineering and Distributive Education were found to have significantly higher proportions of positive residuals than average by first quarter results, while Architecture and Science and General Studies had the two significantly high proportions of the year results.

The year curriculum groups were put to one more test, the paired t-test of actual minus predicted means. By this, Civil Engineers were found to do worse than expected, while Architecture and Distributive Education exceeded prediction by actual performance. Results are summarized in table 15 and represented graphically in figure 4.

Table 14. Frequencies of positive and negative actual-minus-predicted QCA by (a) first quarter curriculum groups (b) final freshman curriculum groups.

Curriculum	Initial				Final			
	Act. > Pred.		A < P	Total	Act. > Pred.		A < P	Total
	No.	Percent.			No.	Percent.		
Agriculture	26	52.0	24	50	23	40.4	34	57
Forestry & Wildlife	12	41.4	17	29	9	36.0	16	25
Business	41	41.4	58	99	57	42.5	77	134
Aerospace Engineering	47	56.0	37	84	31	40.2	46	77
Chemical Engineering	54	61.4*	34	88	43	51.8	40	83
Civil Engineering	45	41.3	64	109	41	39.4	63	104
Electrical Engineering	93	55.4	75	168	72	47.7	79	151
Mechanical Engineering	61	46.9	69	130	57	44.2	72	129
Other Engineering	26	37.7	43	69	27	42.2	37	64
Architecture	38	52.8	34	72	39	60.0*	26	65
Science & General studies	47	43.5	61	108	61	57.0*	46	107
Distributive Education	6	100.0*	0	6	11	68.8	5	16
Total	496	49.0	516	1012	471	46.5	541	1012

* Proportion significantly different from average by chi-square test.

Table 15. Actual and predicted mean year QCA's by third quarter curriculum, tested with paired t-tests.

Curriculum	N	Mean Actual	Mean Predicted	Mean Resid.	SD of Resid.	t_{N-1}
Agriculture	57	.855	.853	.001	.331	.033
Forestry & Wildlife	25	.852	.968	-.116	.605	-.957
Business	134	.760	.817	-.056	.433	-1.501
Aerospace Engineering	77	1.177	1.209	-.032	.475	-.594
Chemical Engineering	83	1.293	1.250	.043	.516	.766
Civil Engineering	104	.975	1.069	-.094	.442	-2.172*
Electrical Engineering	151	1.140	1.133	.007	.553	.147
Mechanical Engineering	129	1.122	1.127	-.004	.480	-.107
Other Engineering	64	1.114	1.142	-.028	.469	-.484
Architecture	65	1.165	1.025	.140	.401	2.804**
Science & General studies	107	1.232	1.163	.069	.514	1.390
Distributive Education	16	.725	.478	.247	.364	2.717*

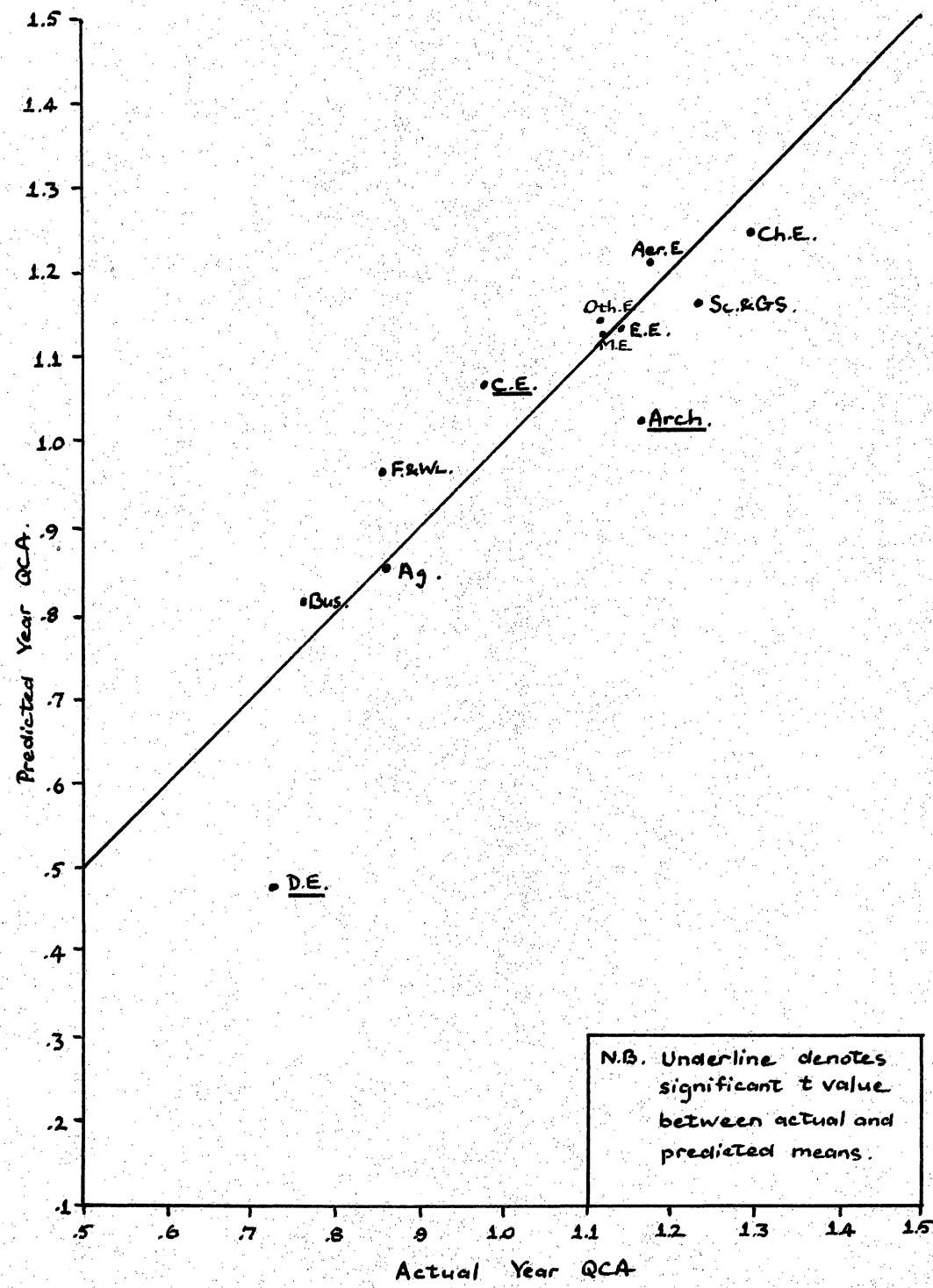

Figure 4. Graph of actual against predicted year QCA means for third quarter curriculum groups, combining all other groups.

It is impossible to be dogmatic in interpretation of these differences, but it is suggested that the Civil Engineering results may be attributed to the preponderance of Virginian cadets in the group. The small numbers in Distributive Education really invalidate a chi-square test, but D.E. majors enter with such extremely low predictions that they obviously must be given special attention or grade concessions. One reason for Chemical Engineers' superiority has already been suggested. Also, other conditions probably do not average out as assumed for each curriculum. For instance, Chemical Engineering has relatively more Cooperative students than any other group, and less trial summer school students.

The higher performance averages of Architecture and Science and General Studies may have similar explanations, but it does seem that, once the many complicating factors are unravelled, such real differences in curriculum groups do exist.

V

PREDICTION BY DISCRIMINANT FUNCTION

Introduction

The first part of this study showed that multiple regression predictions are subject to large errors, and it was thought that a technique which aimed at being less precise, that is, merely classifying candidates into one of two, or possibly more, classes, might yield more conclusive results.

Often, it would be sufficient for admission officers to be able to discriminate between a potential failure with less than 1.000 QCA, and a potential success with QCA 1.000 or higher, using items of prior information known to be pertinent. It is obvious that, if we divide our sample into two groups by QCA, with 1.000 as dividing line, none of the prediction variables used in the multiple regression is dichotomized correspondingly. There is no score on the verbal test, for instance, which would divide our sample candidates into the same two groups as produced by the QCA division. It can be argued that, if all relevant factors could be assessed, then some combination of these would produce a quantitative variable for candidates which would dichotomize exactly as the QCA. This is a theoretical ideal, and in practice, we use what we hope to be the most relevant factors, and produce a combination which will best discriminate between the success and failure groups.

Theory

To establish the best combination we follow the accepted procedure of discriminant analysis, as presented by Rao (5). We consider our

admissions to be two samples, one drawn from the population of "successes", and the other taken from the population of "failures". We select k factors, or prediction variables, and assume that there is a linear combination of these, a quantity y, say, by which the admissions may be identified. We expect the two groups to give reasonably distinct distributions of y, and we choose linear coefficients, c_i, to maximize the distance between the distributions, while minimizing the dispersion of each.

Approaching the problem from different standpoints, statistical authors arrive at the same quantity to be maximized, that is:

$$\frac{N_1N_2(c_1d_1+ \ldots\ldots c_kd_k)^2}{(N_1+N_2)\sum_{i=1}^{k}\sum_{j=1}^{k}c_{ij}w_{ij}}$$ where N_1, N_2 are the two sample sizes.

All the k variables are assumed to be normally correlated, and d_i, the difference between the mean values of variables X_i in the two samples, is supposed to be normally distributed, independantly of w_{ij}.

The two populations are assumed to have a basic population with respect to variances and covariances, and so w_{ij}, (i, j = 1,....k), is an element of the matrix of pooled estimates on (N_1+N_2-2) degrees of freedom of the elements in the underlying dispersion matrix. Fisher (6) essentially uses (N_1+N_2-2-k) degrees of freedom in calculating w_{ij}. This would seem to confuse the problem with a multivariate one, and in this study the formula given in Rao (5) is used for the calculation of w_{ij}, that is:

$$(N_1+N_2-2)w_{ij} = \sum_{t=1}^{N_1}(X_{i1t}-\bar{X}_{i1})(X_{j1t}-\bar{X}_{j1}) + \sum_{t=1}^{N_2}(X_{i2t}-\bar{X}_{i2})(X_{j2t}-\bar{X}_{j2})$$

for i, j = 1,....k .

The maximizing procedure described above resolves into a practical problem of setting up the matrix (w_{ij}) and solving the simultaneous equations, $(w_{ij})\underline{c} = \underline{d}$, to give $\underline{c} = (w^{ij})\underline{d}$, where (w^{ij}) is the reciprocal of (w_{ij}). In fact, only the ratios of the coefficients can be uniquely determined, and so the scale of the coefficients is somewhat arbitrary, but the actual values given by solving the above, are useful in a test of significance for the discriminant function. The test is based on the fact that, with a hypothesis of no real difference between the two populations, considering all k variables simultaneously, the quantity,

$$\frac{N_1N_2(N_1+N_2-k-1)}{k(N_1+N_2)(N_1+N_2-2)} \cdot D^2$$ is known to be distributed as a variance ratio, with k and (N_1+N_2-1-k) degrees of freedom. Mahalanobis' D^2, or the distance between the two populations of y, is just the quantity, $c_1d_1+ \ldots\ldots c_kd_k$, or $\sum\sum w^{ij}d_id_j$, which is also D^2.

It may be assumed, with some justification by the sample admissions, that the two populations, "failure" and "success", are approximately the same size. This simplifies the problem of calculating the probability of error in predictions by discriminant function. The dividing value of the discriminant function will fall midway between the estimated means of the two y populations. Since we are assuming equal variances, the probability of misclassification is the probability of y being greater than $\frac{1}{2} D^2$ for one distribution and less than $\frac{1}{2} D^2$ for the second distribution; that is, considering total misclassification, the probability of $|y|$ being greater than $\frac{1}{2} D^2$. Converting to a standard unit variate, z, it becomes the probability that $|z|$ is greater than $\frac{1}{2} D^2/D$, or Prob $\{|z|>D/2\}$.

Selection of Sample and Variables

The data collected for the multiple regression analysis was used again but, in view of the findings on special groups, (discussed in chapter 4), summer school, Cooperatives and drop-out students were excluded from the sample for the purposes of calculating the discriminant function.

The same three prediction variables, verbal score, X_V, math. score, X_M, and high school rank, X_H, were used, and in addition, status and school were included as prediction factors. Admissions were given the score X_C=1, if they entered V.P.I. as civilians, and X_C=0, if they were cadets. The school variable, X_S, had the value 0 for Virginian students, and 1 for admissions from Out-of-State schools.

When the five-by-five matrix had been inverted, the variable suspected to contribute least to the discriminant function, X_S, was dropped, and the resulting four simultaneous equations solved. Dropping variables in the order of least importance, solutions were found for the three variable case, dropping X_C, for the two variable case, dropping first X_V, then retaining X_V and dropping X_M, and finally, using rank X_H alone.

There is no reason to suppose that freshmen whose high school rank was not given in full might perform differently from the rest of their group, but it was hoped to discover whether the incomplete recording of rank biased predictions in any way. For this purpose, the above calculations were carried out using (a), only those 654 admissions with full information on rank, and (b), including the 86 whose rank was calculated from percentile only.

For each set of solutions the quantities D^2 and $\frac{1}{2} D$ were calculated. The variance ratios, with formula given in the theory section, were computed and tested for significance with the F table. The probabilities of misclassification were found for each value of $\frac{1}{2} D$, by reference to a table of areas under the Normal curve.

Inversion of all matrices, and solutions of the sets of simultaneous equations was done on the IBM 1620 computer.

Results

The mean values of the five variables, X_i, for success and failure samples are given in table 16, together with the differences, d_i, between the pairs of means. There are only slight differences in the results for (a) and (b).

Table 17 gives the corrected sums of squares and cross-products for all the variables. Comparison of success and failure groups shows that, in general, the values are of the same order, which justifies combining them for a pooled estimate of the dispersion elements. However, the cross-products of X_H with X_C and with X_S both show marked dissimilarities between the success and failure groups, which are hard to explain, and may invalidate their pooling. Nevertheless, the sample values were combined, and the elements of the matrices to be inverted are set out in table 18.

The final results are presented below for cases (a) and (b), and using from one to five variables.

Table 16. Mean values of variables, X_i, for success and failure samples, and differences, d_i.

(a) Admissions with complete rank information			
Variable	QCA 1.0 & above	QCA below 1.0	d_i
N	343	311	
$\bar{X}_V$	495.6064	451.8842	43.7222
$\bar{X}_M$	574.2711	525.7042	48.5669
$\bar{X}_H$	59.0000	53.1061	5.8939
$\bar{X}_C$	.3294	.2733	.0561
$\bar{X}_S$	.2799	.1576	.1223
(b) Including admissions with only percentile rank information			
N	386	354	
$\bar{X}_V$	493.4352	452.1808	41.2544
$\bar{X}_M$	574.3782	524.7712	49.6070
$\bar{X}_H$	59.0000	53.2345	5.7655
$\bar{X}_C$	.3290	.2655	.0635
$\bar{X}_S$	.2591	.1469	.1122

Table 17. Corrected sums of squares and cross products of variables, X_i, for success and failure samples.

Values of $(X_{it}-\bar{X}_i)(X_{jt}-\bar{X}_j)$				
Sample	(a) Complete rank information		(b) Including percentile ranks	
QCA	1.0 and above	Less than 1.0	1.0 and above	Less than 1.0
i,j=				
V,V	2,285,263.8659	1,815,103.8328	2,518,912.8808	1,979,166.4294
M,M	1,976,981.7843	1,822,846.7846	2,253,918.7772	1,956,482.4661
H,H	14,918.0000	11,025.4984	15,884.0000	12,705.5396
C,C	75.7726	61.7685	85.2150	69.0396
S,S	69.1312	41.2798	74.0933	44.3616
V,M	835,011.6035	833,953.3505	945,142.4560	889,208.6441
V,H	35,490.0000	21,464.8199	38,842.0000	27,259.9436
M,H	22,611.0000	18,576.7621	28,856.0000	24,425.9915
V,C	-1,420.5248	-2,132.1608	-1,863.2746	-2,383.9944
M,C	-1,311.6385	-2,535.8553	-1,847.0363	-2,647.4915
H,C	-163.0000	-20.0193	-184.0000	-19.0395
V,S	1,520.7843	807.6720	1,639.4767	817.5989
M,S	747.9708	1,000.4952	856.1762	890.8983
H,S	-28.0000	93.8006	-31.0000	90.8079
C,S	-10.6268	-7.3923	-11.9016	-7.8079

Table 18. Pooled estimates of elements in dispersion matrix, (w_{ij}).

(a) Complete rank information: $(N_1+N_2-2) = 652$

	V	M	H	C	S
V	6,288.9075	2,559.7622	87.3540	-5.4489	3.5713
M	2,559.7622	5,827.9579	63.1714	-5.9011	2.6816
H	87.3540	63.1714	39.7906	-.2807	.1009
C	-5.4489	-5.9011	-.2807	.2110	-.0276
S	3.5713	2.6816	.1009	-.0276	.1693

(b) Including percentile ranks: $(N_1+N_2-2) = 738$

	V	M	H	C	S
V	6,094.9584	2,485.5706	89.5690	-5.7551	3.3294
M	2,485.5706	5,705.1507	72.1978	-6.0901	2.3673
H	89.5690	72.1978	38.7392	-.2751	.0811
C	-5.7551	-6.0901	-.2751	.2090	-.0267
S	3.3294	2.3673	.0811	-.0267	.1605

5 Variables: X_V, X_M, X_H, X_C, X_S.

(a) Complete rank information

Discriminant function is:-

$y = (.00291050)X_V+(.00608220)X_M+(.13597618)X_H+(.77175092)X_C+(.60942710)X_S$

D_5^2= 1.34190523 Degrees of freedom= 5, 648 Test statistic= 43.51**

D/2= .5792 Probability of misclassification= .2812

(b) Including partial rank information

$y = (.00257421)X_V+(.00648051)X_M+(.13533264)X_H+(.82066167)X_C+(.61821991)X_S$

D_5^2= 1.32941277 Degrees of freedom= 5,734 Test statistic= 48.83**

D/2= .5765 Probability of misclassification= .2821

4 Variables: X_V, X_M, X_H, X_C.

(a)

$y = (.00314986)X_V+(.00618240)X_M+(.13634146)X_H+(.70150371)X_C$

D_4^2= 1.28091610 Degrees of freedom= 4, 649 Test statistic= 51.94**

D/2= .5659 Probability of misclassification= .2859

(b)

$y = (.00281314)X_V+(.00655689)X_M+(.13543474)X_H+(.75062236)X_C$

D_4^2= 1.26983556 Degrees of freedom= 4, 735 Test statistic= 58.38**

D/2= .5634 Probability of misclassification= .2867

3 Variables: X_V, X_M, X_H.

(a)

$y = (.00280052)X_V+(.00566190)X_M+(.13298601)X_H$

D_3^2= 1.18.23207 Degrees of freedom= 3, 650 Test statistic= 64.03**

D/2= .5434 Probability of misclassification= .2938

(b)

$y = (.00238656)X_V+(.00598290)X_M+(.13216033)X_H$

D_3^2= 1.15722020 Degrees of freedom= 3, 736 Test statistic= 71.04**

D/2= .5379 Probability of misclassification= .2958

2 Variables: X_M, X_H.

(a)

$y = (.00684568)X_M+(.13725475)X_H$

D_2^2= 1.14143923 Degrees of freedom= 2, 651 Test statistic= 92.95**

D/2= .5342 Probability of misclassification= .2972

(b)

$y = (.00697626)X_M+(.13582701)X_H$

D_2^2= 1.12918196 Degrees of freedom= 2, 737 Test statistic= 104.11**

D/2= .5313 Probability of misclassification= .2981

2 Variables: X_V, X_M.

(a)

$y = (.00416123)X_V+(.00685711)X_M$

D_2^2= .51496670 Degrees of freedom= 2, 651 Test statistic= 41.93**

D/2= .3588 Probability of misclassification= .36

1 Variable: X_H.

(a)

$y = (.14812292)X_H$

D/2= .4672 Probability of misclassification= .32

(b)

$y = (.14882856)X_H$

D/2= .4632 Probability of misclassification= .3216

All the values of the test statistic were significant, showing that, in every case, there was a real difference between the "success" and "failure" samples with respect to the variables involved.

As the number of variables decreased, so did the "distance", D^2, between the population means, while the probability of misclassification increased slightly.

For any set of variables, the discriminant function was found to be slightly less efficient when cases with only partial rank information were included in its calculation, in spite of the increase in sample size. However, the differences in results for cases (a) and (b) were always less than the differences caused by dropping a variable.

The probability of misclassification only increased appreciably when X_H or, to a lesser extent, X_M was dropped. As long as these two variables were used, the probability remained just below .30 .

Predictions for Sample Admissions

In order to test the actual success and failure rate against the predicted for various groups, it was decided to use the discriminant function calculated from group (a) admissions, with the five variables, to gain all possible accuracy.

To simplify calculations, the coefficient, c_V, was set equal to unity, and the other coefficients were divided through by c_V.

The mean of the success and failure samples were calculated by the formula, $\bar{y} = \sum_{i=1}^{5} c_i' \bar{X}_i$, where $c_i' = c_i/c_V$, to be 4598.0697 and 4137.0132, respectively. The midpoint, 4367.5414, between the two sample means was taken to be the discriminant value for assigning predictions into the success or failure groups.

A program was written in Fortran language for the IBM 1620 to calculate $y' = X_V + c_M'X_M + c_H'X_H + c_C'X_C + c_S'X_S$ for each sample admission, and to assign him the value 1 or 0, according to whether his prediction was above or below 4367.5414 . The results were then analysed to see how they compared with the expected probable error of misclassification, to investigate curriculum, and to compare special groups against the sample.

Admissions were classified as 00 - predicted and actual QCA's less than 1.000, 01 - predicted less than 1.000 and actual 1.000 or more,

10 - predicted 1.000 and over and actual less than 1.000, or 11 - both predicted and actual QCA's 1.000 or over. The frequencies in each class are given in table 19 for the various groups of sample admissions. The combined numbers in 01 and 10 are taken to be the total misclassification, and are presented as percentages of the total in that group.

The classification was also carried out for each curriculum group, by both first quarter and final freshman curriculum. Percentage total misclassifications are given in table 20. To facilitate curricula comparisons, all admissions who changed curriculum were removed from the sample and the resulting distributions are given in table 21.

Error of Misclassification

In the sample (a), used to construct the discriminant function, the actual total misclassification was 26.8 percent compared with the expected percentage of 28.1 . Considering that the assumptions of symmetry are violated a little, this would seem to be a fair agreement. Errors of both types are not quite evenly represented, however. Of the actual failures, 24 percent are predicted as successes, and 29 percent of the actual successes are wrongly predicted as failures.

Although it was seen that inclusion of candidates without full rank information altered the discriminant function only slightly, it is clear that such a candidate incurs a higher probability of being misclassified himself; 33.7 percent of the 86 admissions with only partial rank information are erroneously classified by the discriminant procedure.

Table 19. Actual and predicted success and failure classification and percentage misclassification.

Admission group	Total	Predicted - Actual Group				Percent misclas.
		00	01	10	11	
Sample (a)	654	236	100	75	243	26.76
Incomplete information on rank	86	29	15	14	28	33.72
Drop-out						
Trial S.S.	21	16	3	2	0	23.81
No S.S.	86	37	6	23	20	33.72
No drop-out						
Vol. S.S.	23	10	3	2	8	21.74
Trial S.S.	82	59	16	1	6	20.73
Cooperative	53	0	6	2	45	15.09

Cooperative, Summer School and Drop-Out Students

It is clear that, for Cooperative and summer school students, prediction by discriminant function is fairly accurate, because they are at the top and bottom of the scale respectively. Cooperative students are generally predicted successes, and it is probable that the few who were wrongly predicted, by the criteria used here, as failures had some other factors in their favor which were known to the admission registrar.

For summer school students, failure is usually predicted before they take the remedial courses, and the percentage of misclassification is lower than in the average class, (although numbers involved are small). There is a certain number, however, who profit from summer school to the extent of being in the C1 group of misclassifications. These are mainly non-drop-out students, and the size of the group supports an argument for reassessment of prediction after attendance at summer school.

Curriculum Investigations

The groupings according to curriculum on entry to V.P.I. show roughly the same misclassification pattern as the final freshman curriculum groupings.

In general, the students remaining in one curriculum throughout the year had slightly higher proportions of misclassifications than did those who changed from one department to another, (a number who transferred out of Architecture form the exception). From this one could infer that misplacement in curriculum was not strongly connected with academic standards, but it is also reasonable to conjecture, since year

Table 20. Percentage misclassification by curriculum.

Curriculum	First quarter curriculum	Final freshman curriculum	Unchanged curriculum
Agriculture	15.38 (26)	12.90 (31)	15.15 (26)
Forestry & Wildlife	44.44 (18)	52.94 (17)	50.00 (16)
Business	26.23 (61)	27.71 (83)	28.85 (52)
Aerospace Engineering	33.33 (63)	31.58 (57)	32.14 (56)
Chemical Engineering	14.00 (50)	14.00 (50)	15.56 (45)
Civil Engineering	29.69 (64)	32.26 (62)	32.73 (55)
Electrical Engineering	21.57 (102)	22.10 (95)	21.05 (95)
Mechanical Engineering	24.39 (82)	24.68 (77)	24.64 (69)
Other Engineering	29.79 (47)	30.23 (43)	31.71 (41)
Architecture	42.62 (61)	36.84 (57)	37.74 (53)
Science & General Studies	23.68 (76)	26.76 (71)	26.56 (64)
Distributive Education	0.00 (4)	9.00 (11)	0.00 (4)
Changes			23.08 (78)
Total	26.76 (654)	26.76 (654)	26.76 (654)

N.B. Numbers in parentheses are the total numbers in the appropriate curriculum groups.

Table 21. Actual and predicted success and failure classification for curriculum groups. (Omitting students who changed curriculum)

Curriculum	Total	Predicted - Actual Group			
		00	01	10	11
Agriculture	26	14	2	2	8
Forestry & Wildlife	16	4	4	4	4
Business	52	31	14	1	6
Aerospace Engineering	56	14	9	9	24
Chemical Engineering	45	12	2	5	26
Civil Engineering	55	24	6	12	13
Electrical Engineering	95	33	12	8	42
Mechanical Engineering	69	21	12	5	31
Other Engineering	41	12	7	6	16
Architecture	53	16	13	7	17
Science & General Studies	64	16	11	6	31
Distributive Education	4	4	0	0	0
Changes	78	35	8	10	25
Total	654	236	100	75	243

QCA's were used as success-failure criteria, that results for changers were better predicted because of this very change to a curriculum for which they were most suited.

Table 20 shows that Agriculture, Chemical Engineering and Electrical Engineering were three curricula groups with exceptionally low percentages of misclassifications, while Forestry and Wildlife, Aerospace Engineering, Civil Engineering and Architecture had rather more than average. These figures could serve as a very rough guide to one more factor in the consideration of probabilities attached to predictions for candidates. The type of error shows up in table 21.

Concerning actual standards in each curriculum, no more can be added to remarks made in the discussion of multiple regression prediction. Apart from the total prediction error, there are a few other prediction results worth noting which show up in table 21. Over two-thirds of the successes in the Business department would have been wrongly predicted as failures by this discriminant function, compared with less than one third in the whole sample. The proportion was also high for Architecture and Forestry and Wildlife. Relatively few errors of this kind would have been made in Chemical Engineering, Agriculture and Electrical Engineering. Very small numbers of actual failures would be misjudged as successes by this discriminant function in Agriculture, Business and Electrical Engineering, but rather more mistakes than average would be made among Forestry and Wildlife, Aerospace Engineering, Civil Engineering and Other Engineering.

It is improbable that this exact pattern would repeat itself in following years, but these curricula differences may well be worth further investigation.

VI

DISCUSSION AND SUMMARY

It is interesting that, although the discriminant function technique is used specifically to divide the "successes" from the "failures", it involves as high a probability of error as does the multiple regression technique, which attempts a more exact prediction.

Although definite differences are shown up between civilian and cadets in the regression analysis, the inclusion of this factor does little to improve prediction by discriminant function. The same is true of the split between Virginia and Out-of-State students.

Results of both analyses testify that the one outstanding factor correlated with college performance is the student's high school record. In fact, for a very quick estimation of a candidate's potential, the admissions officer could find his converted high school rank, and predict him a "failure", (QCA less than 1.000), or a "success", (QCA 1.000 or more), according as his rank was below or above 56, (see results of discriminant function prediction with one variable). He is likely to be wrong 32 times out of 100, but he would make nearly as many mistakes using the full regression prediction. The use of only one predictor may introduce the hazard of a completely ridiculous prediction, but this is likely to be tempered, anyway, by other factors known to the admissions officer.

As pointed out by Duggan and Hazlett (1),

> "Statistical predictions merely give the admissions officer a starting point of information about the applicant. The

> admissions officer may then weigh the predicted QCA with other information about the student."

Another remark made in the workbook (1) cannot be too strongly emphasized:

> "Predictive efficiency of the score and school record is somewhat diminished by two factors:
> 1) the selection process decreases the range of ability in the group, and thereby depresses the validity of the predictors,
> 2) the criterion, college grades, is itself less than perfect as a measure of performance and less than completely stable, i.e. reliable."

These observations would apply equally to discriminant analysis as to multiple regression analysis. The first point refers to the fact that applicants who had very low qualifications were excluded from the study, as of course they were not admitted to college. Thus there has been pre-judgment on some of the very cases which the investigation was set up to learn how to judge. Similarly,* those admissions with low scores and/or rank, who did well in college, show that the admissions officer made a wise choice in admitting them. Presumably, he had other information which helped him to decide that they were not such poor risks as their records indicated. These cases confuse the correlation coefficients and predictions, since the distributions are not even simply truncated, but include a medley of abnormals.

The second point may be added to one made in the discussion of curriculum results: are the College Board scores a valid measure of the student's all-round ability, even at the time of taking the test? Do they measure more than just his ability to do well on tests? It is not surprising, to this author, that high school rank, which is compiled

over a period of time, and by teachers who know the student well, is the most effective variable in prediction. It, too, has obvious pitfalls, similar to those in college grades cautioned by Duggan and Hazlett.

The actual grade system also may have some bearing on prediction accuracy. The point system used at V.P.I. may be more insensitive than a percentage system. For instance, a student who continually just misses an A grade would be penalized against one who usually just made the B class.

The results of this study are not recommended as suitable for future prediction purposes, as there is no proof of a stable situation. As advised by Duggan and Hazlett, results over a number of years should be used for prediction, to average out irregularities, or to show definite trends.

A multiple regression analysis is thought to be more efficient than a discriminant function, and it is recommended that the sample be chosen to exclude as many irregular cases as possible.

However, the two studies here give consistant information concerning the importance of various prediction factors, and of the behaviour of certain groups of students who entered V.P.I. in the fall of 1961. Certain differences between curriculum groups were established, but small numbers caution conclusive statements about them.

ACKNOWLEDGMENTS

The author wishes to thank her supervisor, for his guidance and patience throughout this study. The cooperation of and the Admissions Office in securing the admission data was greatly appreciated.

The author is indebted to and other members of the Statistics department for their continued encouragement and interest.

Much credit must go to friends in the Statistical Laboratory who so willingly sacrificed their spare time, and to and ladies in the Computing Centre for their invaluable help with the machine work.

Finally, the author wishes to express appreciation for the Teaching Assistantship which made it possible for her to do graduate work at Virginia Polytechnic Institute.

BIBLIOGRAPHY

1. Duggan, John M., and Hazlett, Paul H. Jr., Predicting College Grades. 1961. College Entrance Examination Board, New York, N.Y.

2. Long, John. Unpublished report on analysis of admissions, William and Mary College, Norfolk, Va.

3. Kramer, Clyde Y., Multiple Regression. Unpublished. Statistics Department, Virginia Polytechnic Institute, Blacksburg, Va.

4. Wyman, D.G., Stepwise Multiple Linear Regression Analysis for 1620 Card Output. I.B.M. Corporation, 401 Grand Avenue, Oakland, California.

5. Rao, C. R., Advanced Statistical Methods in Biometric Research. 1952. John Wiley and Sons Inc., New York, N.Y.

6. Fisher, R. A., Contributions to Mathematical Statistics, 1949. John Wiley and Sons Inc., New York, N.Y.

Text book for Reference:

Ostle, B., Statistics in Research. 1954. The Iowa State College Press, Ames, Iowa.

The vita has been removed from the scanned document

ABSTRACT

The thesis reports an investigation of 1060 freshman admissions to Virginia Polytechnic Institute in fall, 1961. Multiple regression methods were used to produce equations linking college performance with high school rank and College Board verbal and mathematical scores. Analyses were done for males and females, civilians and cadets. The three predictors accounted for only 34 percent of the variation in first year QCA among males. High school rank contributed most to the prediction; verbal score was found to be a very poor predictor. Predictions were more reliable for accumulative year performance than for first quarter only. Prediction error is discussed and expectancy tables constructed. Actual and predicted quality credit averages were used to investigate subgroups of the sample. Differences were found between civilians and cadets, Virginians and Out-of-Staters. Cooperative, drop-out, and trial summer school groups all had distinct characteristics. Curriculum groups differed widely in actual performance, but in most cases, corresponded to prediction. Exceptions are discussed. The same data was subjected to discriminant analysis, using two extra variables, civilian or cadet status, and Virginia or Out-of-State school. The function divided students by a predicted QCA of 1.000, with 28 percent probability of misclassification. Rank alone was found to be more effective in prediction than verbal and mathematical scores combined. Results of investigation into subgroups using discriminant predictions agreed with regression findings, with different exceptions to curriculum non-significance. Prediction is advised by regression rather than by discriminant analysis, but the present results are not recommended for future application, for reasons discussed.