Factors Relating to Persistence in
A Computer Science Major

by

Patricia F. Campbell	George P. McCabe
Department of Computer Sciences	Department of Statistics
Purdue University	Purdue University

Technical Report #82-15

Department of Statistics
Purdue University

May 1982

Factors Relating to Persistence in a Computer Science Major

Patricia F. Campbell
Department of Computer Sciences

George P. McCabe
Department of Statistics

Purdue University
West Lafayette, IN 47907

ABSTRACT

The pre-college entrance variables of freshman computer science majors were reviewed to determine which variables were related to persistence in the major. Students who persisted in computer science, engineering or another science differed from those students who left for an academically dissimilar goal in their SAT-Math and SAT-Verbal scores, their high school rank, and their background in high school mathematics and science. Sex differences were also noted. Discriminant analysis indicated that the students could be classified on the basis of SAT scores, sex and high school mathematics and science background.

Key Words and Phrases: Computer science education, persistence, admission standards

CR Categories: K.3.2, K.3.m

1. Introduction

As the number of undergraduates who wish to major in computer science continues to rise, the problem of identifying factors which influence initial academic success in computer science becomes increasingly important. Limited facilities and faculty size may necessitate a more selective admissions process, one which demands higher standards for some aspects of a university's general admissions criteria. In addition, given specific information on predictive factors, counselors can more effectively advise students as to whether the computer science major realistically suits their interests and abilities.

1.1 Related work

In several studies attempts have been made to predict performance in introductory computing courses. For example, Petersen and Howe [13] concluded that for a service course which reviewed programming topics as well as the impact of the computer on society, prior college grade point average (GPA) and a general intelligence factor (as measured by the General Aptitude Test Battery) explained 40% of the variation in the final grade. In a similar study involving an introductory data processing course for business majors, Fowler and Glorfeld [6] reported that the prior college GPA, the SAT-Math score, and the number of college mathematics courses completed with a grade of C or better were predictive, with age of the student being of marginal importance. Mazlack [11] noted that neither the academic major (arts versus sciences), sex, nor the number of semesters of university coursework correlated with performance in an introductory Fortran course.

Other investigators have attempted to identify programming aptitude using the IBM Programmer Aptitude Test (PAT) [10]. These results are inconclusive as some studies [1,11] have indicated no significant relationship between the PAT and final grade in an introductory programming course while other studies [2,3] have noted a significant association.

1.2 This Study

In contrast to previous research, this study was concerned with the identification of factors which influence success in the first year of study as a computer science major, not simply one programming course. Evaluation of past records indicated that most of the students who continued to the sophomore year as computer science majors (approximately 40-60% of the previous year's freshman majors) went on to earn the bachelor's degree in computer science. Therefore, successful completion of the first year in the computer science program is a useful indicator of success in the major.

Since the focus of this study was progress in the freshman year, the potential predictors which were reviewed were those contained in the students' high school records.

2. Procedure

2.1 The Sample

All first semester freshman computer science majors who were initially enrolled in the first programming course for majors at a large midwestern university during the fall semester, 1979, were studied. A total of 256 students were identified. Other students with prior university background were also enrolled in this course. Generally, these were students who had either transferred to the computer science major from other programs or who wished to undertake more rigorous programming than required in the departmental service courses. Such students were not included in the sample.

2.2 The Freshman Computer Science Program

The typical first two semesters of study in the computer science major consists of two programming courses, two calculus courses, two courses in English composition, and two additional courses chosen from laboratory science, foreign language humanities. The first programming course concentrated on teaching Pascal, completing the textbook by Findlay and Watt [5]. The second programming course involved further work in Pascal including sorting algorithms, programming systems, recursion and information structures as well as an introduction to Fortran. The text by Wirth [14] is followed.

2.3 Method

The pre-college entrance data on each student was available from the registrar. The variables of interest selected for this study were: SAT-Math score, SAT-Verbal score (200-800), rank in high school graduating class (percentile), size of high school graduating class, number of semesters of high school mathematics (6-12), number of semesters of high school science (2-12), number of semesters of high school English (6-9), average grades in high school mathematics, science and English, and sex.

The academic records of the initial majors were then reviewed during the middle of their third semester (sophomore year). The declared major of each student was noted at that time. Of the 256 initial majors, 103 were listed as sophomore computer science majors, 31 were listed as majors in engineering or some other science (including mathematics), 94 were listed in a non-science, non-engineering major, and 28 had left the university. Thus, the 256 freshmen computer science majors were classified as being in one of three groups during the sophomore year: Computer Science, Engineering or Other Science, and Other. Statistical analysis was then completed to determine:

1) Is there a significant difference between these three groups of students on any of the pre-college entrance variables?

2) Which combinations of the entrance variables may be used to classify these groups of students?

3. Analysis of Group Characteristics

3.1 Attributes of the Three Groups

The means and standard deviations of the pre-college entrance variables were computed for each of the three groups of students (Table I). Using analysis of variance the means of the three groups were compared for each of these variables. The resulting F statistics are noted in Table I. Because the variable of sex has only two values (male or female) rather than a range of values, the percent of males was calculated for each group and compared using the Chi-square statistic. The p-value noted in the Table is the probability of obtaining the statistic noted or one larger if, in fact, the three groups are identical with respect to the variable in question. No significant difference in mean values between the three groups is denoted n.s.

Review of these results indicated that in each case the significant difference in mean values found among the three groups was due to the contrast of the group Other with the remaining two groups. For no variable was the group Computer Science distinguishable from the group Engineering and Other Science. A multivariate t-test comparing these two groups simultaneously on all 11 variables also indicated no significant difference. Therefore, for the remaining analysis, these two groups were combined. Thus, in all that follows the comparison of interest is Computer Science, Engineering and Other Science (CS+) versus Other.

3.2. Attributes of the Two Groups

The means and standard deviations of the pre-college entrance variables were computed for the two groups of students (Table II). Because only two groups were to be compared and the question was whether the group CS+ had significantly greater mean values for the variables than the group Other, one-sided t-tests were used. The resulting t statistics and p-values are reported in Table II.

The students who persisted in a major in computer science, engineering or science had significantly higher SAT-Math and SAT-Verbal scores, ranked higher in their high school graduating class, and completed more semesters of high school mathematics and science with higher grades than did those students who left the computer science major for a dissimilar academic discipline. In addition male students were more likely to persist than female students.

A multivariate analysis comparing the two groups on all 11 variables collectively also indicated a significant difference with a Wilks' lambda test statistic of .80, yielding a p-value less than .001.

The finding of a sex difference was both interesting and alarming. Therefore, the means and standard deviations of the entrance variables were computed for the male and female groups. These values along with the resulting t-statistics are presented in Table III. These results indicate that the males had higher SAT-Math scores, completed more semesters of high school science, ranked lower in their high school graduating class, and had lower average grades in high school mathematics and English than did the females.

4. Discriminant Analysis

From the above analysis, it is evident that the two groups, CS+ and Other, have different average characteristics at the time when the students initially enroll as computer science majors. The finding of statistically significant differences is not sufficient, however, to conclude that these differences are useful or important. To investigate this latter point, an attempt to use the available variables to discriminate between the two groups is presented in this section. In other words, the question as to whether or not the pre-college variables can be used to classify students into the CS+ and Other groups is addressed.

4.1 Variable Selection

Different collections of variables can give essentially the same information. One measure of this phenomenon can be seen in the correlation matrix (Table IV). The correlation coefficient is a measure of the strength of the linear relationship between two variables. It is a value between +1 and -1, the sign indicating whether the association is positive or negative, with a zero correlation coefficient indicating no linear relationship between the variables. For example, it can be seen in Table IV that in this sample, high school rank was highly correlated with average grades in high school mathematics, science and English.

Discriminant analysis is a statistical classification technique which can be used to assign an individual of unknown group origin to the group which he or she most closely resembles on the basis of a set of predictor variables [9]. For this study all possible combinations of the 11 entrance variables forming subsets of size 1 through 11 were examined to determine which subsets were most predictive [12].

In Table V the Wilks lambda statistics for each of the four best subsets of size 1 through 4 are given. The Wilks lambda statistic is a measure of how well a particular set of variables separates the two groups. Within a given subset size, smaller values indicate a stronger discrimination between the groups.

4.2 Classification Results

Following identification of the subsets of interest, discriminant analysis based on each of these sets of predictor variables was completed. Jackknifing was used in order to validate the results [4]. Jackknifing proceeds in the following manner. First, a student is removed from the sample. Then a classification rule is derived using all remaining data. This rule is then used to assign the omitted student to one of the two groups. The student is then returned to the sample. This procedure is repeated for all students in the sample, one at a time, until all are classified. The errors for a given classification model may be estimated by determining the number of mis-assignments [8]. The classification models reported for this study (Table V) were all verified by jackknifing.

As implied by the Wilks lambda values, many classification models produced very similar results. The overall correct classification rate ranged from 58.6% (HSSCG) to 68.4% (SATM, HSMAG, SEX). In Table VI the classification table is given for the model using SATM, HSMAG and SEX. Note that there are two types of errors: a true CS+ may be classified as an Other and a true Other may be classified as a CS+. In this table the two error rates are approximately equal. That is, 43 out of 134 or 32.1% of the CS+ group is misclassified and 38 out 122 or 31.1% of the Other group is misclassified. The classification rule could be modified to decrease one of these error rates with a consequent increase in the other rate.

From a humanistic point of view, the more serious error occurred when the classification model predicted a student's membership in the Other group, when in fact the student was in CS+. For the classification models reported here, that type or error rate ranged from 32.1% (SATM, HSMAG, SEX) to 48.5% (HSSCG, SEX). To further investigate this error, the computer science grades of those students who were misclassified in that manner were reviewed. About 40% of these students received C or lower grades (including withdrawals) in their computer science courses. Thus, it may be argued that although these models misclassify these students as first semester sophomores, the models in fact are identifying the stronger computer science majors.

Another view of these results is obtained by considering the retention rates. Overall 134 out of 256 or 52.3% of the students were in the CS+ group. If only those who were predicted CS+ were admitted, the retention rate would jump dramatically to 91 out of 129 or 70.5%.

5. Discussion

This study was concerned with identifying pre-college entrance variables which were associated with success in the first year of study as a computer science major. The results indicated that those students who persisted in a major in computer science, engineering or other science differed from those students who left computer science for an academically dissimilar goal in their SAT-Math and SAT-Verbal scores, their high school rank, and their background in high school mathematics and science.

Differences based on sex were also noted. Males were more persistent in the scientific and engineering majors than were females. Of the 98 females, only 38 (39%) persisted; of the 158 males, 96 (61%) persisted. Males tended to have higher SAT-Math scores and more semesters of high school science. Females ranked higher in their high school graduating class with better average grades in mathematics and English. Even after conditioning on the other variables, a statistically significant sex difference is detectable. These results indicate that sex-role socialization is probably having a negative effect. Mathematics education research [7,15] suggests that in our society adolescent females are learning to avoid situations where they may fail, while adolescent boys are learning to try harder. The initial programming courses are very demanding, requiring, on the average, 20 to 30 hours of work outside of class each week. This work is typically completed through the evening into the early morning hours. Due to societal influences, females may be more likely to decide that the major is not worth the effort. Thus if the field is serious in its goal of sex equity, a lack of bias within computer science classrooms is probably not sufficient. Rather, overt evidence of support for females is probably necessary to modify existing social forces.

Discriminant analysis indicated that subsets of the pre-college entrance variables can be used to classify the students. SAT scores along with high school mathematics and science background as well as sex seemed to be key variables. The emphasis on mathematics ability in the classification models may be due to a combination of two facts. First, problem solving ability is clearly crucial for success in a scientific or engineering field. In addition, this analysis was designed to distinguish between the groups CS+ and Other. All of the academic majors represented in CS+ require two courses in calculus (designed for science and engineering majors) during their first year of study. Scientific calculus is not required for the majors grouped as Other, although calculus for social science or business students may be expected. Thus the program of study may be influencing this result.

An optimal classification rule could be determined using ideas from statistical decision theory. This procedure would require (1) a prior estimate of the probability that a student is in the CS+ group and (2) a numerical valued loss function specifying the consequences of misclassification for each type of error. Clearly, the second is the most difficult to obtain. On the other hand, if there were a fixed limited number of spaces for applicants, the classification rule could be easily adjusted to select those who appear to be most qualified. Finally, it should be kept in mind that any statistical procedure has serious limitations when applied indiscriminately to people.

Table I

Variable	Mean (SD) CS (n=103)	 Eng. & Other Sci. (n=31)	 Other (n=122)	F (2,253)	p
SAT Math	619 (86)	629 (67)	575 (83)	10.35	<.001
SAT Verbal	530 (94)	512 (95)	486 (88)	6.53	.002
HS Rank	88.0 (10.5)	89.2 (10.8)	85.8 (10.8)	1.95	n.s.
HS Size	380 (264)	388 (251)	348 (214)	0.63	n.s.
HS Math Semesters	8.74 (1.28)	8.65 (1.31)	8.25 (1.17)	4.56	.012
HS Science Semesters	6.29 (1.86)	6.29 (1.62)	5.59 (1.89)	4.60	.011
HS English Semesters	7.81 (0.85)	7.84 (0.52)	7.68 (0.84)	0.88	n.s.
HS Math Grades*	3.61 (0.46)	3.62 (0.40)	3.35 (0.55)	9.38	<.001
HS Science Grades*	3.53 (0.50)	3.53 (0.46)	3.26 (0.57)	8.25	<.001
HS English Grades*	3.40 (0.48)	3.41 (0.61)	3.39 (0.47)	0.02	n.s.
% Male	75.0	61.3	50.8	x_2^2=13.92	<.001

*Average grades were coded: 7(A), 6(A-), 5(B+), 4(B), 3(B-), 2(C+), 1(C), 0(C-).

Table II

Variable	Mean (SD) CS, Eng. & Other Sci. (n=134)	Mean (SD) Other (n=122)	t (df=255)	p*
SAT Math	621 (82)	575 (83)	4.51	<.001
SAT Verbal	526 (94)	486 (88)	3.48	<.001
HS Rank	88.3 (10.5)	85.8 (10.8)	1.90	.029
HS Size	381 (260)	348 (214)	1.12	n.s.
HS Math Semesters	8.72 (1.28)	825 (1.17)	3.00	.001
HS Science Semesters	6.29 (1.80)	5.59 (1.89)	3.04	.001
HS English Semesters	7.81 (0.79)	7.68 (0.84)	1.31	n.s.
HS Math Grades	3.61 (0.44)	3.35 (0.55)	4.34	<.001
HS Science Grades	3.53 (0.49)	3.26 (0.57)	4.07	<.001
HS English Grades	3.40 (0.51)	3.39 (0.47)	0.19	n.s.
% Male	71.9	50.8	$\chi_1^2=11.14$	<.001

*The p-value given is for a one-sided test with the alternative hypothesis favoring the CS, Eng. and Other Sci. Group.

Table III

	Mean (SD)			
Variable	Men (n=158)	Women (n=98)	t	p
SAT Math	615 (85)	575 (82)	3.78	<.001
SAT Verbal	509 (96)	505 (90)	0.30	n.s.
HS Rank	85.0 (11.5)	90.5 (8.3)	-4.16*	<.001
HS Size	352 (234)	388 (248)	-1.15	n.s.
HS Math Semesters	8.56 (1.27)	8.40 (1.21)	1.00	n.s.
HS Science Semesters	6.39 (1.72)	5.26 (1.90)	4.82	<.001
HS English Semesters	7.71 (0.84)	7.82 (0.77)	-1.05	n.s.
HS Math Grades	3.44 (0.52)	3.57 (0.48)	2.03	.044
HS Science Grades	3.38 (0.55)	3.44 (0.54)	0.96	n.s.
HS English Grades	3.27 (0.52)	3.61 (0.36)	5.75*	<.001

*In these cases the variance for the men was significantly ($p < .05$) larger than that for the women and the t-statistic reported has been calculated using separate variance estimates.

Table IV

Correlations Among Variables

	SATV	HSR	HSS	HSMAS	HSSCS	HSENS	HSMAG	HSSCG	HSENG	SEX
SAT Math	.38	.17	.05	.27	.22	-.04	.35	.18	.07	-.18
SAT Verbal		.22	-.01	.12	.10	.07	.14	.24	.28	.03
HS Rank			.02	.08	.03	-.04	.58	.60	.68	.29
HS Size				-.02	-.07	.06	-.03	-.07	-.11	.09
HS Math Semesters					.10	-.01	.19	.11	.13	-.02
HS Science Semesters						.04	.09	.15	-.01	-.27
HS English Semesters							.05	.06	.03	.08
HS Math Grades								.53	.45	.19
HS Science Grades									.56	.12
HS English Grades										.35

Note: Correlations greater than .12 (.16) in absolute value are significantly different from zero with $p < .05$ ($p < .01$).

Table V

Wilks Lambda for Best Subsets in Discriminant Analysis

Subset Size	Wilks Lambda	Subset
1	.925	SATM
1	.933	HSMAG
1	.941	HSSCG
1	.953	SATV
2	.871	HSMAG,SEX
2	.888	HSSCG,SEX
2	.891	SATM,HSSCG
2	.898	SATM,HSMAG
3	.845	SATV,HSMAG,SEX
3	.857	HSMAG,HSSCG,SEX
3	.857	SATM,HSMAG,SEX
3	.858	SATM,HSSCG,SEX
4	.836	SATV,HSS,HSMAG,SEX
4	.837	SATV,HSMAS,HSMAG,SEX
4	.838	SATV,HSMAG,HSSCG,SEX
4	.838	SATV,HSMAG,HSENG,SEX

Table VI

Classification Table (Model: SATM, HSMAG, SEX)

		Predicted Group		
		CS+	Other	Total
	CS+	91 (35.5%)	43 (16.8%)	134
Actual Group	Other	38 (14.8%)	84 (32.8%)	122
	Total	129	127	256

References

1. Alspaugh, C.A. Identification of some components of computer programming aptitude. Jrnl. for Res. in Math. Ed., 3, 2 (March, 1972), 89-98.

2. Bateman, C.R. Predicting performance in a basic computer course. Proc. of the Fifth Annual Meeting of the Amer. Inst. for Decision Sciences, Boston, 1973.

3. Capstick, C.K., Gordon, J.D., and Salvadori, A. Predicting performance by university students in introductory computing courses. SIGCSE Bull. 7, 3 (Sept., 1975), 21-29.

4. Dixon, W.J., and Brown, M.B. (Eds.). BMDP Biomedical Computer Programs: P-Series 1979. University of California Press, Berkeley, 1979.

5. Findlay, W., and Watt, D.A. PASCAL: An Introduction to Methodical Programming. Computer Science Press, Potomac, Maryland, 1978.

6. Fowler, G.C., and Glorfeld, L.W. Predicting aptitude in introductory computing: A classification model. AEDS Jrnl., 14, 2 (Winter, 1981), 96-109.

7. Fox. L.H. The effects of sex-role socialization on mathematics participation and achievement. In Fox, L.H., Fennema, E., and Sherman, J. (Eds.). Women and Mathematics: Research Perspectives for Change (NIE Papers in Education and Work: No. 8). National Institute of Education, Washington, D.C., 1977, 1-77.

8. Gray, H.L., and Shucany, W.R. The Generalized Jackknife Statistic. Marcel Dekker, Inc., New York, 1972.

9. Lachenbruch, P.A. Discriminant Analysis. Hafner Press, New York, 1975.

10. Manual for Administrating and Scoring the Aptitude Test for Programmer Personnel. IBM Technical Publications Department, White Plains, New York, 1964.

11. Mazlack, L.J. Identifying potential to acquire programming skill. Comm. ACM, 23, 1 (Jan., 1980), 14-17.

12. McCabe, G. Computations for variable selection in discriminant analysis. Technometrics, 17, 1 (Feb., 1975), 103-109.

13. Petersen, C.G., and Howe, T.G. Predicting academic success in introduction to computers. AEDS Journal, 12, 4, (Sum., 1979), 182-191.

14. Wirth, N. *Algorithms ± Data Structures = Programs*. Prentice Hall, Englewood Cliffs, New Jersey, 1976.

15. Wolleat, P.L., Pedro, J.D., Becker, A.D., and Fennema, E. Sex differences in high school students' causal attributions of performance in mathematics. *Jrnl. for Res. in Math. Ed.*, *11*, 5 (Nov., 1980), 356-366.