RESEARCH REPORT No. 87

Using Demographic Information in Predicting College Freshman Grades

Richard Sawyer

February 1985

ACT

USING DEMOGRAPHIC INFORMATION IN PREDICTING COLLEGE FRESHMAN GRADES

ABSTRACT

In this study we determined whether adjustments for differential prediction observed among sex, racial/ethnic, or age subgroups in one freshman class at a college could be used to improve prediction accuracy for these subgroups in future freshman classes. The study is based on the ACT test scores, high school grades, and college freshman grade averages of students from national samples of colleges.

For older students, dummy variable and separate subgroup prediction equations were found, on cross-validation, to be more accurate than the total group equations. For sex subgroups, dummy variable and separate subgroup equations were only moderately effective in improving prediction accuracy. For racial/ethnic subgroups, dummy variable and separate subgroup equations were more often than not less accurate, on cross-validation, than total group equations. Among all three kinds of demographic subgroupings, shifts over time in colleges' mean grades were found to be a much more important source of prediction bias than differential prediction. Moreover, prediction bias itself, from whatever source, was typically much smaller than error variance.

USING DEMOGRAPHIC INFORMATION IN PREDICTING COLLEGE FRESHMAN GRADES

Richard Sawyer

At some colleges, the relationship between college freshman grades, standardized test scores, and high school grades may differ among various demographic subgroups of students. When this happens, a prediction equation developed from the total group of students may result in systematic over- or underprediction for different subgroups. In recent years a consensus has been developing among educational researchers that total group prediction equations tend to overpredict the freshman grades of racial/ethnic minorities, to underpredict the grades of women, and to underpredict the grades of nontraditional-aged freshmen (Linn, 1978; Breland, 1979; Levitz, 1982).

It is reasonable, on discovering differential prediction, to inquire whether using subgroup membership to adjust predictions increases prediction accuracy. It would seem that systematic over- or underprediction by a total group equation would necessarily imply this to be so. Few studies, however, have shown evidence that a pattern of differential prediction observed in one freshman class persists in subsequent classes, and whether, therefore, it is possible to reduce prediction bias by making such adjustments. Determining the answer to this question requires cross-validating prediction equations over time, which is a primary theme of this paper.

In the cross-validation paradigm, prediction equations developed from the data of one freshman class are applied to the test scores and high school grades of a future freshman class, and the predicted and actual grades of the future students are compared. This procedure models the actual use of prediction equations by colleges, and it avoids the tendency of estimates of prediction accuracy derived from a single year's data to be overoptimistic. We shall show that prediction equations incorporating sex, racial/ethnic, or age information often are less accurate, on cross-validation, than the total group prediction equations based solely on test scores and high school grades.

As in any use of test scores, the ethical consequences of using demographic information to improve the prediction of college freshman grades should be studied (Cole, 1981). For example, many people would object to using variables like race or sex as predictors if the predictions were part of a highly selective admissions procedure. In a situation like this, candidates would be competing among each other, but some candidates would be put at a disadvantage solely on the basis of background factors that some social norms require not be considered. When freshman grade predictions are used for noncompetitive purposes such as general counseling, sectioning, or placement, this problem would seem to be mitigated. Of the colleges that use ACT Assessment data, a large majority do, in fact, use them for such noncompetitive purposes (Levitz, 1980).

It should be emphasized that the subject of this report is prediction bias (systematic under- or overprediction of a criterion variable). Prediction bias is conceptually different from selection bias caused by systematically under- or overstating a group's true qualifications for purposes of selection. Linn (1984) discussed this concept of selection bias; he showed that when there are differences in the average true qualifications of two groups, and when there is measurement error in the predictor variable, then there will be differences in the groups' regressions of criterion variable on predictor variable, even in the absence of selection bias. As a result, moderately under- or overpredicting the criterion variable for a subgroup does not necessarily indicate selection bias as formulated by Linn. Prediction bias itself, though, is a serious practical concern for students as well as for institutions. In sectioning and placement, for example, systematically overpredicting the grades of a particular subgroup could imply that an institution's sectioning and placement procedure is ineffective for that subgroup.

Gamache and Novick (in press) investigated an alternative to using demographic information explicitly in prediction. In situations where there are multiple predictors, some subset of the predictors might retain most of the predictive validity of the full set, yet differ less among the subgroups in their relationships with the criterion. In this case, using the subset of predictors would reduce systematic under- and overprediction while avoiding the ethical or political difficulties of explicitly using demographic information. It is not yet known how frequently this approach is feasible among colleges generally.

The major purpose of this study was to determine whether using demographic information in prediction

The author thanks David Jarjoura, Michael Kane, Jim Maxey, and Robert Brennan for comments on an earlier draft.

results in improved prediction accuracy, and if so, whether the improvements had practical significance. The demographic variables investigated (sex, race, and age) were chosen because of perceived general interest in them.

A second purpose of the study was to evaluate alternative statistical methods for using demographic information in grade predictions. The most common method for using demographic information is to develop separate prediction equations for each demographic subgroup; ACT's Basic and Standard Research Services (The American College Testing Program, 1983a) provide colleges with the capability of doing this. Another method, using demographic dummy variables, was also evaluated.

A third purpose of this study was to describe the statistical characteristics of colleges at which using sex, racial/ethnic, or age information results in more accurate grade predictions on cross-validation. Statistical criteria for predicting the gain in prediction accuracy were evaluated.

Prediction Errors

The relationship between college grades and predictor variables may differ among subgroups of a population. Therefore, prediction equations which make allowances for these differences could potentially be more accurate than an overall total group prediction equation. There are statistical problems, however, which may prevent this from occurring in practice.

It is convenient, in discussing sources of prediction error, to use mean squared error as a measure of prediction accuracy. Mean squared error, in this discussion, is the expected squared difference between the grade predicted for a student and the grade actually earned by the student; the expectation is taken with respect to repeated sampling from the hypothetical population or populations of students from which consecutive freshman classes at a college are drawn. Mean squared error is a convenient measure of prediction accuracy because it is the sum of prediction error variance and squared prediction bias. This fact follows from the standard identity $E[e^2] = E[(e-\mu_e)^2] + \mu_e^2$, where e is the prediction error for a student, E denotes expected value, and $\mu_e = E[e]$. The quantity μ_e corresponds to prediction bias, and the quantity $E[(e-\mu_e)^2]$ corresponds to error variance.

Prediction bias is that part of mean squared error due to systematic under- or overprediction. One type of bias occurs when there are differences in slope or intercept among population subgroups. In this situation, total group regression equations are biased, while separate subgroup regression equations may be unbiased. Another type of bias, not often acknowledged, occurs when there are systematic differences between the population of students from whose data the prediction equations are developed and the population of students for whom the predictions are made. In this case, both total group and separate subgroup equations may be biased. We shall show that in predicting college freshman grade average, this latter bias is usually larger than that due to differential prediction among subgroups.

Prediction error variance is the portion of mean squared error due to random errors, i.e., not due to systematic under- or overprediction. Error variance can be thought of as reflecting two sources of random error. One source is the inherent limitation of the predictors in predicting the criterion; it is often expressed by the "residual variance" associated with the regression equation. The other source of random error is sampling error due to estimating regression coefficients. Sawyer (1982) studied the error variance associated with prediction equations developed from and applied to a multivariate normal population. He found that the proportionate increase in error variance due to estimating the regression coefficients can be approximated by a simple function of the base sample size, the number of predictors, and the population residual variance.

The relative importance of bias and error variance depends on the use being made of a prediction equation. If a prediction equation is used to select applicants to a program with a restrictive admissions policy, then prediction bias may be more important than error variance. When a prediction equation is used for general counseling or for course sectioning and placement, then both components could be of roughly equal importance; in such a situation it would be appropriate to use mean squared error or some similar overall measure of prediction accuracy.

In practice, prediction accuracy as measured by mean squared error is a trade-off between bias and error variance. A total group equation may have larger bias than separate subgroup equations, but, because it is based on a larger sample, may have smaller error variance. The net result may be that a total group prediction equation would have smaller mean squared error than separate subgroup equations. Furthermore, some prediction methods may be more sensitive than others to biases caused by differences between the population of students from whose data the prediction equations are developed and the population of students for whom the predictions are made.

Dummy Variables

There are other methods for using demographic information in predictions besides developing separate subgroup equations. A simple but often effective alternative is to use subgroup membership dummy variables as additional predictors. Zedeck (1971) proposed that this should, in fact, be checked before developing separate subgroup equations. With regard to bias and error variance, using dummy variables is a compromise between using total group and separate subgroup equations. Because using dummy variables adjusts the intercept of the fitted regression surface, it potentially results in smaller bias than using a total group equation; because it requires estimating more parameters, however, it results in a larger error variance. On the other hand, a prediction equation with S-1 dummy variables has effectively far fewer predictor variables than S separate subgroup equations; a dummy variable prediction equation, therefore, will tend to result in smaller error variance than separate subgroup equations.

In principle, one could also include interactions of the dummy variables with the other predictors in the prediction equation, thereby approximating even more closely the separate subgroup equations. In practice, this would be difficult to do effectively unless one had cross-validation evidence suggesting which interactions to use.

One should, of course, keep in mind that using dummy variables in prediction does not imply any direct causal relationship between subgroup membership and the criterion variable. The dummy variables can be thought of as proxies for the many complex background, social, and educational characteristics that are related to performance in college, but that are not measured by high school grades or test scores. Viewed this way, using dummy variables is merely a statistical tool for reducing prediction bias.

Data for This Study

The American College Testing Program (ACT) offers to colleges research services for measuring the local predictive validity of the ACT Assessment (ACT, 1983a). These predictive research services summarize the relationships between the ACT scores, high school grades, and college freshman grades of students at a postsecondary institution. These services can also be used to generate weights for predicting the freshman grades of future applicants.

The study is based on three data sets constructed from freshman grade information submitted by colleges to the ACT predictive research services:

- Data Set A consists of student records from 200 colleges. These data were used to evaluate the usefulness of sex (gender) information in prediction.
- Data Set B consists of student records from 256 colleges. These data were used to evaluate the usefulness of racial/ethnic information in prediction.
- Data Set C consists of student records from 216 colleges. These data were used to evaluate the usefulness of age information in prediction.

These three data sets were constructed for earlier predictive validity studies at ACT. The colleges represented in them are stratified random samples of colleges that participated in ACT's predictive research services in two or more years. Although the data sets were constructed separately, they are not mutually exclusive. For a detailed description of the sampling methods used to construct the data sets, see Sawyer and Maxey (1979a), Maxey and Sawyer (1981), and Levitz (1982), respectively. The three data sets are summarized in Table 1.

The data for each college consist of "base year" data used to develop prediction equations, and "cross-validation year" data, against which the prediction equations were cross-validated. To increase the number of colleges with data for minority and nontraditional age freshmen, the base year data for colleges in Data Sets B and C were allowed to be from any one of the three freshman class years 1973-74, 1974-75, or 1975-76. From colleges which submitted freshman grade data for more than one of these three years, only the latest year's data were used.

The two to four year lag between the base year data and the cross-validation year data, as shown in Table 1, was chosen to reflect the typical frequency of participation of colleges in the ACT predictive research services. Sawyer and Maxey (1979b) found that at most colleges the accuracy of predictions of freshman grade average is very stable over a two- to four-year age in the prediction equations.

The subgroup information on sex, race, and age was reported by students when they registered to take the ACT Assessment. The three racial/ethnic categories used in Data Set B are "Afro-American/Black," "Caucasian-American/White," and "Mexican-American/Chicano," as they are the most frequently reported. For brevity, they will be referred to as "black," "white," and "Chicano," respectively.

TABLE 1

Summary of Data Sets for Cross-Validation Study

Data set	Base year(s)	Cross-validation year	Subgroups	Size of total sample		Size of sample used to calculate statistics	
				Colleges	Students	Colleges	Students
A. Sex	1974-75	1976-77	Female	192	51,437	172	50,370
			Male	178	43,765	170	43,138
			Total group	200	105,502	—	—
B. Race	1973-74, 1974-75, & 1975-76	1977-78	Minority	112	12,007	89	9,351
			White	228	81,645	99	58,805
			Total group	256	134,601	—	—
C. Age	1973-74, 1974-75, & 1975-76	1977-78	Age 17-19	207	78,598	83	46,589
			Older	84	7,521	70	6,735
			Total group	216	96,522	—	—

Note. The sample sizes refer to the number of colleges and number of student records associated with the cross-validation year data for each data set.

Age, in Data Set C, is defined as age at time of matriculation. Age was calculated by subtracting the student-reported year of birth from the year of matriculation. Months were not used to calculate age. The three age subgroups in this study are those used by Levitz (1982): Age 17-19, Age 20-25, and Age 26 or older.

The number of colleges and the total number of student records, by subgroup, are shown in the fifth and sixth columns of Table 1. Because not every student reported demographic information, the subgroup sample sizes in Table 1 do not sum to the total group sample sizes. Moreover, for reasons to be explained below, the cross-validation statistics were not calculated at every college; the sample sizes for the cross-validation statistics are shown in the last two columns of Table 1.

The predictor variables used in this study are the ACT Assessment subtest scores (in English, mathematics, social studies, and natural sciences), and the four self-reported high school grades in the subject areas of the ACT subtests. Alternative predictor variables studied were the ACT Composite Score (the average of the four subtest scores) and HSA (the average of the four self-reported high school grades). For information about the ACT Assessment and self-reported high school grades, see the ACT Technical Report (1973).

The criterion variable used in this study is college freshman grade average reported on a 0.0-4.0 scale. Most of the grade averages are from the first semester of the freshman year. Colleges participating in ACT's research services do have the option of pooling grades from previous years, or reporting grade averages based on the entire freshman year. ACT does not maintain records of individual colleges' choices of criteria. However, it is estimated that over 60% of the colleges in the data bases for this study reported first semester grade average and most of the rest reported first year cumulative grade average.

It should be noted that, as is usual in predictive validity studies, the criterion measure for this study reflects any treatment or selection made on the basis of the predictor variables. Thus, at colleges which use test scores and high school grades for sectioning, the students' grade averages reflect these interventions; at colleges which use test scores and high school grades for selecting applicants, the enrolled freshmen may not be representative of the larger applicant pool.

One should also keep in mind that because the data in this study were collected from colleges participating in the ACT predictive research services, they are in some respects not representative of students nationally:

- Colleges using the ACT Assessment are located mainly in the Rocky Mountains, Great Plains, Southwest, Midwest, and Southeast with comparatively fewer in the Northeast and West Coast.
- Privately controlled institutions are relatively underrepresented among colleges that use the ACT Assessment, and publicly controlled institutions are overrepresented.
- Participation in ACT's research services is voluntary, so that the data base is self-selected even among colleges that use the ACT Assessment.

The results of the study therefore cannot be claimed to represent precisely the results that would be obtained if data from all colleges in the United States could somehow be collected.

Method

At each college, multiple linear regression prediction equations were calculated from the freshman grade averages, ACT test scores, self-reported high school grades, and demographic characteristics of the students in its base year sample. The ACT test score and high school grade information was used in two alternative combinations:

(a) four ACT test scores and four high school grades (8V), and

(b) ACT Composite score and HSA (2V).

ACT routinely uses in its prediction services a slight modification of the 8V multiple linear regression. The 2V predictions in this study were calculated to determine whether decreasing the number of predictor variables would improve prediction accuracy by decreasing error variance.

ACT routinely calculates for each college a total group equation (TG), i.e., one that does not use student demographic information. In this study, demographic information was used in the form of:

(a) demographic dummy variables (DV), and

(b) separate subgroup equations (SG).

Therefore, five different kinds of prediction equations were calculated at each college: the standard (8V-TG), and four alternatives using demographic information (8V-DV, 8V-SG, 2V-DV, 2V-SG).

Data Set B was constructed to maximize the number of minority students. Nevertheless, only 21 colleges had enough Chicanos to permit reporting DV or SG predictions for that group, and only 4 colleges had enough blacks and Chicanos to permit reporting DV or SG prediction equations for all three racial/ethnic groups simultaneously. Some compromise in the scope of this study was therefore required.

According to ACT's enrolled freshmen norms (The American College Testing Program, 1983b), the mean ACT Composite score and HSA of blacks are 12.7 and 2.66, respectively; of Chicanos, 14.7 and 2.83, respectively; and of whites, 19.9 and 3.05, respectively. Therefore, the ACT scores and self-reported high school grades of Chicanos are more similar to those of blacks than they are to those of whites. Moreover, the same is true of the validity of ACT test scores and self-reported high school grades in predicting college freshman grade average; Maxey and Sawyer (1981), for example, reported cross-validated mean absolute errors of .59, .59, and .53 grade units for blacks, Chicanos, and whites, respectively, using total group prediction equations. One would prefer to study prediction equations for blacks and Chicanos separately. Given the limitations imposed by sample size, though, the preceding considerations make combining the two groups into a single "minority" subgroup for developing DV and SG prediction equations a sensible compromise.

A racial/ethnic dummy variable was therefore created to differentiate black and Chicano students from white students. Data from students who did not report a racial/ethnic category or who reported a category other than black, white, or Chicano were not used to develop DV and SG equations.

Sample size considerations also made it necessary to follow a similar procedure in developing the age DV and SG predictions from Data Set C. Those students age 20 or older will be referred to as "older" students.

DV and SG prediction equations calculated from very small numbers of students could be subject to large sampling errors. To avoid cluttering the results with statistics that primarily reflect such sampling errors, a minimum sample size of 25 students from each appropriate subgroup (male/female; minority/white; age 17/older) was required to calculate DV or SG prediction equations. Similarly, cross-validation statistics calculated from very small numbers of students could be subject to large sampling errors. Therefore, a minimum sample size of 25 student records from a subgroup was required to calculate cross-validation statistics. Thus, cross-validation statistics were calculated on subsets of the colleges in the three data sets, as shown in the last two columns of Table 1.

From the cross-validation year data in each data set, the following measures of prediction accuracy were calculated for each college, prediction method, and subgroup:

- MSE, the observed mean squared error, i.e., the average squared difference between predicted and earned grade average. Smaller values of MSE correspond to more accurate prediction than do larger values of MSE.
- BIAS, the average observed difference between predicted and earned grade average. Positive values of BIAS correspond to overprediction, and negative values correspond to underprediction.

Because its algebraic properties make MSE convenient to use in discussing sources of prediction error, we have chosen to report it here. In practical applications, however, mean absolute error is a more intuitively appealing measure of prediction accuracy because it is expressed in the same unit of measurement as the criterion. Mean absolute error results for the demographic subgroups in this study are given in the references cited in the discussion of the three data bases A, B, and C. However, root mean squared error, the square root of MSE, does have the same unit as the criterion; when sampling from a multivariate normal population, mean absolute error is approximately $\sqrt{2/\pi}$ times root mean squared error.

The cross-validation statistics for the TG predictions for each subgroup were calculated only from those data used to calculate cross-validation statistics for the DV and SG predictions. For example, the TG cross-validation statistics for students age 17-19 were computed from 46,589 records from 83 colleges, rather than from 78,598 records at 207 colleges. This was done to permit making a direct comparison of the accuracies of the TG, DV, and SG prediction methods.

Results

This section focuses on overall trends in the cross-validation statistics. A discussion of the relationships between the cross-validation statistics and other institutional characteristics is contained in later sections. The distributions of the cross-validation statistics across colleges are summarized below. All frequency distributions have been weighted to reflect the sampling designs used to create Data Sets A, B, and C; the weighted results refer to the population of colleges that participate in the ACT predictive research services.

Sex

The distribution of cross-validated BIAS across colleges in Data Set A (Sex) is summarized in Tables 2 and 3. The distribution of cross-validated MSE is summarized in Tables 4 and 5. The symbols Q_1 and Q_3 in these tables denote the first and third quartiles.

Bias. According to Table 2 the 8V-TG predictions for females were typically somewhat negatively biased (median BIAS = –.05 grade units), and the 8V-TG predictions for males were typically somewhat positively biased (median BIAS = .05 grade units). For both sexes, though, there was a large range of BIAS values among colleges (from about –.4 to .4).

None of the alternative prediction methods considered in this study tended to reduce simultaneously the magnitudes of both positive and negative prediction biases. For example, the minimum, median, and maximum BIAS values for females were –.44, –.05, and .55, respectively, for the 8V-TG predictions; for the 8V-DV predictions, they were –.34, .00, and .66, respectively. The alternative prediction methods therefore tended to reduce the negative biases but enlarge the positive biases. A corresponding effect, opposite in sign, occurred in the predictions for males.

The difference $\Delta BIAS^2$(alternative) = $BIAS^2$(8V-TG) – $BIAS^2$(alternative) is an indicator of the degree to which an alternative prediction method reduced squared BIAS. Positive values of $\Delta BIAS^2$ indicate that the alternative prediction method was successful in

TABLE 2

Distribution of Cross-Validated BIAS Across Colleges, by Sex Subgroup and Prediction Method

Sex subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Female	Min.	–.44	–.34	–.36	–.31	–.30
	Q_1	–.13	–.08	–.08	–.06	–.06
	Med.	–.05	.00	.01	.02	.02
	Q_3	.04	.08	.09	.10	.11
	Max.	.55	.66	.66	.66	.65
Male	Min.	–.43	–.50	–.52	–.50	–.51
	Q_1	–.03	–.09	–.11	–.10	–.10
	Med.	.05	–.01	–.02	–.01	–.01
	Q_3	.14	.08	.08	.07	.07
	Max.	.38	.34	.29	.33	.32

reducing squared prediction bias; negative values of this difference indicate that the alternative was not successful in reducing squared bias.

Table 3 shows the distributions of $\Delta BIAS^2$ for the alternative prediction methods. The category limits of .01 and .04 in this table were chosen because they correspond to changes of about 15% and 30%, respectively, in root mean squared error, given the median MSE(8V-TG) of .46. Table 3 shows that the alternative prediction methods were able to reduce squared BIAS in only a small majority of colleges. The most successful alternative was 8V-DV, which reduced squared bias in only 56% and 54% of the colleges for females and males, respectively. Thus, the alternatives to 8V-TG prediction were only marginally successful at what they were intended to do.

MSE. The ratio $BIAS^2/MSE$ is an indicator of the relative importance of prediction bias as a component of MSE. Values of this ratio near 1 would suggest that MSE is due mostly to prediction bias, while values near 0 would suggest that MSE is due mostly to error variance. In the distribution of $BIAS^2/MSE$ for the 8V-TG equations for either sex (table not shown), the first quartile is about .005, the median .02, and the third quartile .06. Stated another way, prediction bias represented about 7% to 24% of root mean squared error among the middle half of colleges. Thus while prediction bias was large for a few colleges, error variance usually accounted for a larger proportion of mean squared error.

Table 4 shows the expected result that the MSE typically observed for females was somewhat smaller than that typically observed for males. The median MSE for the 8V-TG predictions for females, for example, was .43, as compared to .50 for the 8V-TG predictions for males. Table 4 also shows that the MSEs associated with the 2V predictions were very similar to those associated with the more complex 8V predictions.

The median MSE for females was virtually the same for both the 8V-TG and the 8V-DV predictions. The median MSE for males was also the same (.50) for both kinds of predictions. The separate subgroup equations typically resulted in slightly larger MSEs for males (median = .54) than did the 8V-TG predictions.

The superiority of dummy variable over separate subgroup predictions in reducing MSE is apparent in the proportions of colleges with given differences MSE(8V-TG) - MSE(8V-DV) and MSE(8V-TG) - MSE(8V-SG), as shown in Table 5. In about 54% of the colleges the 8V-DV predictions for females were more accurate than the 8V-TG predictions; but only in about 39% of the colleges were the 8V-SG predictions more accurate. The 8V-DV predictions for males were more accurate than the 8V-TG predictions in 57% of the colleges; but the 8V-SG predictions for males were more accurate in only about 30% of the colleges.

TABLE 3

Distribution of Differences in Cross-Validated Squared BIAS Across Colleges, by Sex Subgroup

Sex subgroup	Range in difference	$BIAS^2$(8V-TG) - $BIAS^2$(8V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(8V-SG)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-SG)
Female	less than -.04	.03	.05	.04	.05
	-.04 to -.01	.15	.14	.14	.17
	-.01 to .00	.27	.27	.27	.25
	.00 to .01	.29	.31	.27	.27
	.01 to .04	.23	.18	.21	.21
	.04 or more	.04	.06	.06	.06
Male	less than -.04	.07	.10	.05	.08
	-.04 to -.01	.17	.17	.21	.20
	-.01 to .00	.22	.22	.20	.20
	.00 to .01	.30	.26	.32	.28
	.01 to .04	.19	.20	.17	.18
	.04 or more	.05	.04	.05	.06

TABLE 4

Distribution of Cross-Validated MSE Across Colleges, by Sex Subgroup and Prediction Method

Sex subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Female	Min.	.17	.17	.17	.18	.15
	Q_1	.31	.31	.32	.30	.30
	Med.	.43	.42	.43	.41	.41
	Q_3	.54	.53	.54	.53	.53
	Max.	1.22	1.36	1.36	1.34	1.39
Male	Min.	.22	.22	.22	.19	.21
	Q_1	.38	.38	.40	.38	.38
	Med.	.50	.50	.54	.49	.50
	Q_3	.62	.63	.67	.61	.62
	Max.	1.07	1.13	1.48	1.12	1.25

Table 5 also shows that the 2V-DV and 2V-SG predictions more frequently reduced MSE than their 8V counterparts did. In about 59% of the colleges the 2V-DV predictions for females were more accurate than the 8V-TG predictions; in about 56% of the colleges the 2V-SG predictions for females were more accurate than the 8V-TG predictions. The magnitudes of the reductions in MSE, though, typically are not large.

In the distribution of the differences between 8V-TG error variance and 8V-DV, 8V-SG, 2V-DV, and 2V-SG error variance (table not shown), the median differences for females were approximately -.00, -.01, .00, and .00, respectively. The medians of these differences for males across colleges were .00, -.01, .01, and .00, respectively. While the differences among these medians are small, they do suggest that the 2V-DV predictions tended to reduce error variance and that the 8V-SG predictions tended to increase error variance.

In summary, none of the four alternatives to the 8V-TG predictions reduced prediction bias at a large majority of colleges. The simplest alternative method (2V-DV) was the most successful in reducing MSE, but the magnitude of the reduction was typically modest. The most complex alternative (8V-SG) actually tended to increase MSE because it increased the error variance.

TABLE 5

Distribution of Differences in Cross-Validated MSE Across Colleges, by Sex Subgroup

Sex subgroup	Range in difference	MSE(8V-TG) - MSE(8V-DV)	MSE(8V-TG) - MSE(8V-SG)	MSE(8V-TG) - MSE(2V-DV)	MSE(8V-TG) - MSE(2V-SG)
Female	less than -.04	.07	.18	.05	.09
	-.04 to -.01	.16	.25	.18	.17
	-.01 to .00	.23	.17	.18	.18
	.00 to .01	.27	.17	.18	.18
	.01 to .04	.22	.15	.24	.25
	.04 or more	.05	.07	.17	.13
Male	less than -.04	.07	.32	.05	.10
	-.04 to -.01	.15	.24	.16	.23
	-.01 to .00	.20	.14	.19	.16
	.00 to .01	.26	.14	.14	.11
	.01 to .04	.25	.12	.28	.25
	.04 or more	.06	.04	.17	.16

Race

The distribution of cross-validated BIAS across colleges in Data Set B (Race) is summarized in Tables 6 and 7. The distribution of cross-validated MSE is summarized in Tables 8 and 9.

BIAS. According to Table 6, the grade averages of minority students were typically somewhat overpredicted by the 8V-TG equations (median BIAS = .09 grade units). The median BIAS of the 8V-TG predictions for whites was nearly 0. As is true of the sex subgroups, though, there was a large range of BIAS values among colleges.

The alternative prediction methods considered in this study did not reduce the under- and overprediction for white students. This result is not surprising, as whites constitute a large majority at most colleges. The minimum and maximum BIAS observed for 8V-TG predictions were -.35 and .43, respectively. The corresponding maximum and minimum for the 8V-DV predictions were -.39 and .45; for the 2V-DV predictions, they were -.38 and .42.

TABLE 6

Distribution of Cross-Validated BIAS Across Colleges, by Racial/Ethnic Subgroup and Prediction Method

Racial/ethnic subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Minority	Min.	-.40	-.59	-.61	-.62	-.58
	Q_1	-.04	-.13	-.11	-.10	-.09
	Med.	.09	.01	.01	.01	.01
	Q_3	.22	.13	.14	.12	.13
	Max.	.54	.66	.80	.70	.76
White	Min.	-.35	-.39	-.39	-.38	-.39
	Q_1	-.09	-.07	-.06	-.06	-.05
	Med.	-.00	.02	.02	.02	.01
	Q_3	.08	.10	.09	.11	.11
	Max.	.43	.45	.46	.42	.42

TABLE 7

Distribution of Differences in Cross-Validated Squared BIAS Across Colleges, by Racial/Ethnic Subgroup

Racial/ethnic subgroup	Range in difference	$BIAS^2$(8V-TG) - $BIAS^2$(8V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(8V-SG)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-SG)
Minority	less than -.04	.16	.18	.15	.15
	-.04 to -.01	.20	.18	.14	.16
	-.01 to .00	.16	.14	.20	.13
	.00 to .01	.13	.16	.15	.20
	.01 to .04	.20	.16	.18	.18
	.04 or more	.15	.16	.17	.17
White	less than -.04	.02	.01	.03	.02
	-.04 to -.01	.07	.11	.07	.09
	-.01 to .00	.47	.43	.40	.42
	.00 to .01	.34	.34	.40	.36
	.01 to .04	.08	.08	.08	.08
	.04 or more	.02	.03	.02	.03

The alternative prediction methods did reduce the median BIAS for minority students (from .09 to .01 grade units), but they exaggerated the extremes. The minimum and maximum BIAS for the 8V-TG predictions for minority students were -.40 and .54, respectively. For the 8V-DV predictions the minimum and maximum were -.59 and .66; for the 8V-SG predictions they were -.61 and .80. A similar exaggeration of the extremes occurred with the 2V predictions.

The distributions of the differences in squared BIAS between the 8V-TG predictions and the alternatives indicate that the 8V-DV and 8V-SG predictions actually increased squared BIAS in a small majority of colleges for both minority and white students. The 2V-DV and 2V-SG predictions reduced squared BIAS for minority students in 50% and 55% of the colleges, respectively. One must conclude, therefore, that none of the alternatives to the standard 8V-TG predictions was particularly successful in reducing prediction bias.

MSE. In the distribution of $BIAS^2/MSE$ for the 8V-TG predictions for minority students (table not shown), the first quartile was about .01, the median .04, and the third quartile .12. The corresponding quartiles for white students were .00, .02, and .04, respectively. Thus, as in Data Set A, prediction bias contributed less to mean squared error than did error variance.

Table 8 shows that the MSE for the 8V-TG predictions for minority students was typically somewhat larger (median = .53) than that for white students (median = .50). Table 8 also suggests that in most colleges the alternative prediction methods did little to reduce the MSEs for either minority or white students, and, in fact, usually increased MSE. This effect is apparent in Table 9, which summarizes the distributions of the differences between MSE(8V-TG) and MSE for the alternative prediction methods. The 8V-DV, 8V-SG, 2V-DV, and 2V-SG predictions reduced MSE for minority students in only about 45%, 28%, 45%, and 47% of the colleges, respectively. A similar result occurred with white students: In only about 42%, 42%, 40%, and 44% of the colleges was MSE reduced by the alternative prediction methods.

In the distribution of the differences between 8V-TG error variance and 8V-DV, 8V-SG, 2V-DV, and 2V-SG error variance (table not shown), the median differences for minority students were -.00, -.03, -.01, and -.01, respectively. The median differences for white students were .00, -.00, -.00, and -.00, respectively. Therefore, the alternative prediction methods tended to increase prediction error variance for both minority students and for white students.

In summary, none of the four alternative prediction methods reduced prediction bias for racial/ethnic groups in more than a small majority of colleges; in fact, the 8V alternatives increased prediction bias in a small majority of colleges. Moreover, none of the alternative prediction methods reduced MSE in a majority of colleges. Finally, all four alternative prediction methods tended to increase error variance.

TABLE 8

Distribution of Cross-Validated MSE Across Colleges, by Racial/Ethnic Subgroup and Prediction Method

Racial/ethnic subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Minority	Min.	.22	.20	.24	.22	.21
	Q_1	.43	.45	.47	.44	.45
	Med.	.53	.54	.57	.54	.53
	Q_3	.68	.64	.72	.61	.64
	Max.	.99	1.17	2.03	1.25	1.39
White	Min.	.21	.22	.22	.21	.21
	Q_1	.39	.40	.39	.39	.39
	Med.	.50	.51	.51	.51	.51
	Q_3	.64	.64	.63	.64	.64
	Max.	1.03	1.10	1.09	1.07	1.07

Age

The distribution of cross-validated BIAS across colleges in Data Set C (Age) is summarized in Tables 10 and 11. The distribution of cross-validated MSE is summarized in Tables 12 and 13.

Bias. Table 10 shows that the grade averages of older students were typically underpredicted by the standard 8V-TG equations, and often to a substantial degree. The quartiles of BIAS for this subgroup and prediction method were –.33, –.20, and –.08. The grade predictions for students age 17-19 typically had biases of smaller magnitude.

Table 10 shows that the alternative prediction methods tended to reduce the extreme positive values of BIAS for students age 17-19, and to increase the extreme negative values. For older students, the alternative prediction methods reduced the extreme negative values of BIAS and increased the extreme positive values.

TABLE 9

Distribution of Differences in Cross-Validated MSE Across Colleges, by Racial/Ethnic Subgroup

Racial/ethnic subgroup	Range in difference	MSE(8V-TG) - MSE(8V-DV)	MSE(8V-TG) - MSE(8V-SG)	MSE(8V-TG) - MSE(2V-DV)	MSE(8V-TG) - MSE(2V-SG)
Minority	less than –.04	.17	.43	.12	.23
	–.04 to –.01	.26	.20	.32	.21
	–.01 to .00	.11	.09	.11	.09
	.00 to .01	.13	.04	.13	.13
	.01 to .04	.17	.15	.16	.17
	.04 or more	.15	.09	.16	.17
White	less than –.04	.04	.08	.02	.04
	–.04 to –.01	.11	.13	.25	.23
	–.01 to .00	.43	.37	.33	.30
	.00 to .01	.26	.27	.20	.25
	.01 to .04	.12	.12	.18	.16
	.04 or more	.04	.03	.02	.03

TABLE 10

Distribution of Cross-Validated BIAS Across Colleges, by Age Subgroup and Prediction Method

Age subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Age 17-19	Min.	–.34	–.48	–.49	–.41	–.42
	Q_1	–.01	–.08	–.08	–.07	–.07
	Med.	.06	.02	.02	.02	.02
	Q_3	.14	.11	.11	.11	.11
	Max.	.57	.39	.39	.37	.37
Older	Min.	–.89	–.72	–.78	–.77	–.63
	Q_1	–.33	–.14	–.15	–.17	–.16
	Med.	–.20	–.05	–.03	–.03	–.04
	Q_3	–.08	.12	.11	.12	.09
	Max.	.30	.56	.49	.56	.48

The distributions of differences in squared BIAS between 8V-TG and the alternatives are shown in Table 11. All the alternative prediction methods reduced squared BIAS in a majority of colleges. For students age 17-19, they reduced squared BIAS in 55%-61% of the colleges. For older students, they reduced squared BIAS in 64%-70% of the colleges, and in about a third of the colleges they reduced squared BIAS by .04 or more.

MSE. In the distribution of $BIAS^2/MSE$ for students age 17-19 (table not shown), the quartiles were .00, .02, and .05. For older students, the quartiles were .03, .06, and .13. Thus, although prediction bias was a more significant source of prediction error among older students than among students age 17-19, in both subgroups error variance accounted for a larger proportion of mean squared error than did bias.

Table 12 shows that the 8V-TG MSE for older students (median = .78) was considerably larger than that for students age 17-19 (median = .55). The alternative prediction methods tended to reduce slightly the median MSE for older students, but did exaggerate the

TABLE 11

Distribution of Differences in Cross-Validated Squared BIAS Across Colleges, by Age Subgroup

Age subgroup	Range in difference	$BIAS^2$(8V-TG) - $BIAS^2$(8V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(8V-SG)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-DV)	$BIAS^2$(8V-TG) - $BIAS^2$(2V-SG)
Age 17-19	less than -.04	.02	.04	.01	.01
	-.04 to -.01	.10	.06	.10	.10
	-.01 to .00	.27	.34	.28	.28
	.00 to .01	.43	.38	.41	.45
	.01 to .04	.12	.11	.14	.11
	.04 or more	.05	.06	.06	.05
Older	less than -.04	.12	.16	.12	.18
	-.04 to -.01	.12	.12	.14	.08
	-.01 to .00	.07	.07	.03	.03
	.00 to .01	.12	.09	.13	.11
	.01 to .04	.19	.21	.21	.25
	.04 or more	.37	.34	.36	.34

TABLE 12

Distribution of Cross-Validated MSE Across Colleges, by Age Subgroup and Prediction Method

Age subgroup	Quantile	Prediction method				
		8V-TG	8V-DV	8V-SG	2V-DV	2V-SG
Age 17-19	Min.	.26	.27	.27	.25	.26
	Q_1	.46	.46	.46	.46	.46
	Med.	.55	.56	.55	.55	.55
	Q_3	.67	.68	.70	.66	.66
	Max.	1.18	.99	1.05	.98	1.00
Older	Min.	.23	.23	.24	.24	.23
	Q_1	.67	.65	.65	.64	.67
	Med.	.78	.75	.78	.75	.76
	Q_3	1.01	.97	1.07	1.00	.98
	Max.	1.31	1.59	2.58	1.59	1.69

extreme upper values. For example, using the 8V-DV predictions reduced the median MSE to .75, but increased the maximum MSE from 1.31 to 1.59.

Table 13 shows that the 8V-DV and 2V-DV predictions reduced MSE for older students at 69% and 64% of colleges, respectively. The 8V-SG and 2V-SG predictions reduced MSE at only 47% and 55% of colleges, respectively. In the distributions of the difference between 8V-TG error variance and 8V-DV, 8V-SG, 2V-DV, and 2V-SG error variance for older students (table not shown), the medians were -.00, -.02, -.00, and -.01, respectively. These results, along with those in Table 11, suggest that the 2V-DV predictions did not reduce error variance, as might be thought, but instead reduced MSE by reducing prediction bias. A similar result is true of students age 17-19.

TABLE 13

Distribution of Differences in Cross-Validated MSE Across Colleges, by Age Subgroup

Age subgroup	Range in difference	MSE(8V-TG) - MSE(8V-DV)	MSE(8V-TG) - MSE(8V-SG)	MSE(8V-TG) - MSE(2V-DV)	MSE(8V-TG) - MSE(2V-SG)
Age 17-19	less than -.04	.02	.09	.02	.05
	-.04 to -.01	.12	.08	.17	.20
	-.01 to .00	.25	.23	.33	.31
	.00 to .01	.35	.38	.08	.05
	.01 to .04	.20	.16	.29	.29
	.04 or more	.05	.05	.11	.09
Older	less than -.04	.18	.38	.16	.24
	-.04 to -.01	.08	.12	.10	.12
	-.01 to .00	.05	.04	.08	.08
	.00 to .01	.14	.04	.09	.10
	.01 to .04	.21	.12	.25	.11
	.04 or more	.34	.31	.30	.34

The Relationship Between Prediction Accuracy and Other Statistical Characteristics of Institutions

We have seen that the alternative prediction methods considered in this study were moderately successful in reducing BIAS and MSE for age subgroups, marginally successful for sex subgroups, and mostly unsuccessful for racial/ethnic subgroups. In a first step toward explaining these results, we determined the statistical characteristics of colleges associated with different levels of prediction accuracy. The following base year statistics, in various combinations, were studied:

Summary

- Base year sample sizes for the subgroups (BASEN). These variables are related to sampling error in estimating regression coefficients, and therefore to prediction error variance. Base year sample sizes could also be proxy variables for the characteristics of students who enroll at different types of colleges.
- Error variance index (VINDX). We used the following estimate of the prediction error variance (Browne, 1975):

$$\text{VINDX}(i) = \frac{(n_i+1)(n_i-p)}{n_i(n_i-p-2)} S_i^2$$

where n_i is the base year sample size for Subgroup i, p is the number of predictors and S_i^2 is the usual unbiased estimate for the residual variance for Subgroup i.
- Differential prediction bias index (DPINDX). An intuitively appealing and common quantification of differential prediction for a subgroup is the difference between the total group and subgroup predictions at the subgroup mean. Specifically, let YTG(i) and YSG(i) be the means of the total group and

separate subgroup predictions in the base year data for Subgroup i. The difference

$$DPINDX(i) = \overline{YTG}(i) - \overline{YSG}(i)$$

was used as an index of prediction bias for Subgroup i.

To determine the effect of changes in grading practices on prediction accuracy, we also considered the difference in mean grade average ($\Delta\overline{Y}$). This variable is the difference between the cross-validation year and base year mean grade averages at a college:

$$\Delta\overline{Y}(i) = \text{Cross-validation year } \overline{Y}(i) - \text{Base year } \overline{Y}(i).$$

Positive values of $\Delta\overline{Y}(i)$ correspond to a trend of higher grades over time. This variable was calculated for every subgroup (i).

The variables BASEN, DPINDX, and VINDX are institutional characteristics computable from base year data. Institutions could, therefore, use these variables to predict the benefit of incorporating demographic information in their predictions. The variable $\Delta\overline{Y}$ cannot, of course, be used this way; but, as will be evident, it is a more important determinant of cross-validated prediction accuracy than the base year statistics.

Other variables reflecting changes in the joint distribution of predictor variables and grades, such as changes in mean ACT Composite or HSA, could also potentially be related to prediction accuracy. These relationships could be caused by differences in predictive validity among students with different ability levels, or they could be proxies for relationships between the cross-validation statistics and other, unspecified variables. In either case, these relationships are likely to be much weaker than the relationship between prediction accuracy and $\Delta\overline{Y}$. In view of the difficulty and expense of collecting data on and computing these other change variables, the analyses were restricted to the four institutional statistics BASEN, VINDX, DPINDX, and $\Delta\overline{Y}$ defined above. These four variables were calculated for every subgroup in the three data sets. Their distributions are summarized in Table 14.

BIAS

At a college where ACT score and high school grade means are stable over time, the expected value of BIAS for the 8V-TG predictions for Subgroup i is equal to $E[DPINDX(i)]-E[\Delta\overline{Y}(i)]$. We therefore modeled observed BIAS as:

$$BIAS = a + b \cdot DPINDX + c \cdot \Delta\overline{Y} + \text{error},$$

where one would anticipate the constant b to be positive and the constant c to be negative. In the fitting this model all variables were standardized to have mean 0 and variance 1. This was done so that regression coefficients for DPINDX and $\Delta\overline{Y}$ could be directly compared. BASEN was used as a third explanatory variable in the model for older students, as preliminary analyses had suggested that BASEN and BIAS were strongly related for this particular subgroup.

To prevent outlier observations from unduly influencing the estimated regression coefficients, observations with large Cook D statistic values (Cook, 1977) were eliminated from the analyses. Observations were eliminated when they fell outside the 20% confidence contours associated with the estimated regression coefficient vectors. About 3-6 cases were deleted from the various data sets.

The regression coefficients are displayed in Table 15. The positive signs for the regression coefficients for DPINDX show that prediction biases observed in the base year data tended to carry over, though in diminished relative magnitude, to future classes. The negative signs of the regression coefficients for $\Delta\overline{Y}$ reflect the fact that increases in mean grade average over time at a college tend to result in systematic underprediction. The magnitudes of the coefficients for $\Delta\overline{Y}$ and DPINDX suggest that on a standard deviation basis, a given change in mean grade average typically results in more change in prediction bias than does a comparable change in differential prediction. These two different kinds of prediction bias do, of course, have different effects on individual students; therefore, the practical significance of these results will depend on particular characteristics of a college's admissions and counseling procedures. The BASEN regression coefficient for older students suggests that underprediction of their grades is greatest at large institutions.

A change in the mean freshman grade average at a college need not by itself cause prediction bias if there were a corresponding change in ACT scores and high school grades. The magnitudes of the regression coefficients for $\Delta\overline{Y}$ in Table 15 suggest, though, that changes in mean grade average are not linked to changes in the predictor variable means. Various explanations could be made of the causes of $\Delta\overline{Y}$ in this context. Two plausible interpretations are that $\Delta\overline{Y}$ represents a change in institutional grading standards, or that $\Delta\overline{Y}$ is a result of changes in the freshman curriculum. Different interpretations would likely be applicable at different institutions.

TABLE 14

Distribution of Institutional Characteristics, by Subgroup

Subgroup	Statistic	Institutional characteristic				Sample size
		BASEN	VINDX	DPINDX	$\Delta\overline{Y}$	
Female	Min.	25	.09	-.25	-.81	172
	Q_1	71	.29	-.08	-.17	
	Med.	120	.37	-.05	-.06	
	Q_3	236	.48	-.03	.04	
	Max.	1691	1.20	.16	.33	
	Mean	229	.41	-.05	-.07	
	SD	288	.18	.06	.18	
Male	Min.	27	.07	-.13	-.46	170
	Q_1	67	.34	.03	-.10	
	Med.	109	.45	.06	-.00	
	Q_3	213	.58	.09	.09	
	Max.	1430	2.05	.27	.61	
	Mean	204	.50	.06	.00	
	SD	245	.25	.06	.18	
Minority	Min.	26	.16	-.18	-.73	89
	Q_1	45	.47	-.00	-.17	
	Med.	63	.56	.09	-.01	
	Q_3	119	.69	.14	.16	
	Max.	496	1.41	.48	.70	
	Mean	97	.60	.08	-.03	
	SD	85	.21	.12	.28	
White	Min.	37	.22	-.17	-.40	99
	Q_1	195	.37	-.03	-.10	
	Med.	466	.45	-.01	-.03	
	Q_3	891	.55	.00	.09	
	Max.	2806	1.16	.12	.41	
	Mean	633	.49	-.02	-.01	
	SD	598	.17	.04	.17	
Age 17-19	Min.	38	.23	-.03	-.40	83
	Q_1	245	.41	.01	-.12	
	Med.	458	.49	.02	-.04	
	Q_3	762	.60	.06	.09	
	Max.	3079	1.10	.33	.39	
	Mean	682	.54	.05	-.02	
	SD	681	.20	.06	.12	
Older	Min.	25	.22	-.52	-.68	70
	Q_1	48	.64	-.26	-.15	
	Med.	84	.79	-.18	.03	
	Q_3	127	1.06	-.08	.13	
	Max.	320	2.02	.34	.95	
	Mean	105	.84	-.16	.03	
	SD	71	.33	.14	.28	

TABLE 15

Regression Coefficients (and *p*-values) Associated With Multiple Regression of BIAS(8V-TG) on DPINDX and $\Delta\overline{Y}$, by Subgroup

Subgroup	Institutional characteristic			Multiple R	Sample size
	BASEN	DPINDX	$\Delta\overline{Y}$		
Female	—	.30 (<.0001)	-.89 (<.0001)	.88	166
Male	—	.37 (<.0001)	-.84 (<.0001)	.81	167
Minority	—	.51 (<.0001)	-.82 (<.0001)	.80	86
White	—	.18 (<.0001)	-.91 (<.0001)	.90	96
Age 17-19	—	.23 (.002)	-.89 (<.0001)	.83	78
Older	-.22 (.007)	.40 (<.0001)	-.81 (<.0001)	.85	64

Note. These coefficients pertain to models with standardized variables (z-scores).

MSE

Regression models were also computed with MSE as the dependent variable:

$$\begin{aligned} MSE(i) = a &+ b\cdot BASEN(i) + c\cdot VINDX(i) \\ &+ d\cdot DPINDX(i) + e\cdot[DPINDX(i)]^2 \\ &+ f\cdot\Delta\overline{Y}(i) + g\cdot[\Delta\overline{Y}(i)]^2 \end{aligned}$$

As in the analysis of BIAS, all variables were standardized to have mean 0 and variance 1. Both linear and quadratic terms for DPINDX and $\Delta\overline{Y}$ were used because preliminary analyses revealed that doing so considerably improved the fit of the models. The quadratic terms for $\Delta\overline{Y}$ and DPINDX are the squares of the respective standardized variables.

The resulting coefficients and their associated significance levels are shown in Table 16. Among all subgroups $\Delta\overline{Y}$ and VINDX were the two strongest predictors of MSE; this is consistent with the result noted earlier that $\Delta\overline{Y}$ was the most important predictor of BIAS, but that error variance accounted for a larger proportion of MSE than predictor bias. Using the regression coefficients in Table 16 to plot MSE against $\Delta\overline{Y}$ shows that larger than average MSEs were associated with decreases in mean freshman grade average and that slightly smaller than average MSEs were associated with increases in mean freshman grade average.

Predicting Gains in Prediction Accuracy

The final stage of the analysis involved determining the statistical characteristics of institutions at which the alternative prediction methods led to gains in prediction accuracy. The four institutional characteristics BASEN, VINDX, DPINDX, and $\Delta\overline{Y}$ were used as predictors of the differences in squared BIAS shown in Tables 3, 7, and 11 ($\Delta BIAS^2$) and of the differences in MSE shown in Tables 5, 9, and 13 (ΔMSE).

$\Delta BIAS^2$

At a college where ACT score and high school grade means are stable over time, the expected value of $\Delta BIAS^2$ is equal to $E[DPINDX^2]-2E[DPINDX\cdot\Delta\overline{Y}]$. We therefore developed regression models for $\Delta BIAS^2$ with linear terms for BASEN and VINDX, linear and quadratic terms for DPINDX and $\Delta\overline{Y}$, and the cross-product term $DPINDX\cdot\Delta\overline{Y}$. As in the analyses of BIAS and MSE, outlier observations were deleted whenever their Cook D statistic values were associated with a confidence contour of .20 or higher. To make the samples for the alternative prediction methods identical, outlier observations deleted from the analysis for one prediction method were deleted from the analyses for all the other prediction methods.

TABLE 16

Regression Coefficients (and *p*-values) Associated With Multiple Regression of MSE(8V-TG) on Four Institutional Characteristics

	Institutional characteristic									
Subgroup	BASEN (Z)	VINDX (Z)	DPINDX (Z)	DPINDX (Z^2)	$\Delta\overline{Y}$ (Z)	$\Delta\overline{Y}$ (Z^2)	**Multiple R**	**Sample size**
Female	.01	.54	-.07	-.01	-.41	.09	.71	172
	(.86)	(<.0001)	(.34)		(<.0001)			
Male	.01	.66	.13	-.10	-.32	.09	.66	168
	(.84)	(<.0001)	(.001)		(<.0001)			
Minority	.16	.51	.15	-.00	-.35	.09	.61	87
	(.08)	(<.0001)	(.29)		(.0008)			
White	-.01	.57	-.06	.02	-.49	.13	.73	98
	(.85)	(<.0001)	(.47)		(<.0001)			
Age 17-19	.04	.55	.04	.07	-.43	.12	.83	82
	(.70)	(<.0001)	(.03)		(<.0001)			
Older	.18	.53	-.21	.09	-.22	.03	.71	68
	(.10)	(<.0001)	(.05)		(.11)			

Note. These coefficients pertain to models with standardized variables (z-scores).

All the regression models had high to very high levels of fit. For all subgroups except whites and older students, the simple model

$$\Delta BIAS^2 = a + b \cdot DPINDX^2 + c \cdot DPINDX \cdot \Delta\overline{Y} + error$$

fit nearly as well as the full model described in the preceding paragraph; therefore, the results are discussed in the context of simple models. For whites, the coefficient for $DPINDX^2$ was statistically insignificantly different from 0 ($p > .20$) in the full data set; but deleting outliers led to negative values of the coefficient. Therefore, we used only the cross-product term to predict $\Delta BIAS^2$ for white students. For older students, BASEN was statistically significant ($p < .004$) and was therefore included in the model. The regression statistics are displayed in Table 17.

The primary purpose of these analyses was to determine conditions under which $\Delta BIAS^2$ could be expected to be strongly positive, strongly negative, or near 0. Therefore, the regression coefficients in Table 17 pertain to nonstandardized (raw) scores, rather than to standardized (z-) scores.

The large magnitudes and negative signs for coefficients of the cross-product term imply that the alternative prediction methods were most successful in reducing squared BIAS when DPINDX and $\Delta\overline{Y}$ had opposite signs. A further implication is that even when DPINDX is large in magnitude, the alternative prediction methods can be ineffective or even counterproductive in reducing squared BIAS and DPINDX have the same sign. This would occur when mean grade averages shift in the opposite direction from the adjustment implied by an alternative prediction method. In predicting the grade averages of males, for example, DPINDX is typically positive, and the alternative prediction methods result in lower predicted grade averages than a total group equation; if the mean grade averages of males increase over time, though, one would have been better off using the total group equation.

Figure 1 is a contour plot for predicted values of Δ $BIAS^2$(8V-DV) for older students, given values of DPINDX and $\Delta\overline{Y}$. The various colors correspond to ranges in the predicted values of $\Delta BIAS^2$. For example, the dark green regions correspond to values of DPINDX and $\Delta\overline{Y}$ in which the predicted value of $\Delta BIAS^2$ is greatest; the light green regions correspond to predicted values of $\Delta BIAS^2$ that are small, but positive; and the red regions correspond to negative predicted values of $\Delta BIAS^2$. Note that in the green regions, DPINDX and $\Delta\overline{Y}$ tend to have opposite signs, but in the red regions, they have the same sign; this reflects the importance of the cross-product term in predicting $\Delta BIAS^2$.

TABLE 17

Regression Coefficients Associated With Multiple Regression of $\Delta BIAS^2$ on Institutional Characteristics

Subgroup	Prediction method	Intercept	$DPINDX^2$	$DPINDX \cdot \Delta\overline{Y}$	BASEN	SEE	Multiple R	Sample size
Female	8V-DV	.00	1.47	-1.35	—	.015	.79	162
	8V-SG	.00	1.47	-1.44	—	.017	.76	
	2V-DV	.00	.70	-1.04	—	.018	.75	
	2V-SG	.00	.73	-1.07	—	.018	.76	
Male	8V-DV	-.00	1.04	-1.66	—	.017	.88	161
	8V-SG	-.00	1.02	-1.75	—	.022	.82	
	2V-DV	-.00	.74	-1.04	—	.020	.82	
	2V-SG	-.00	.74	-1.12	—	.022	.80	
Minority	8V-DV	.00	1.04	-1.69	—	.019	.89	81
	8V-SG	.00	.88	-1.55	—	.020	.86	
	2V-DV	.00	.91	-1.52	—	.018	.90	
	2V-SG	.00	.87	-1.57	—	.020	.88	
White	8V-DV	-.00	—	-1.37	—	.003	.86	82
	8V-SG	-.00	—	-1.45	—	.003	.85	
	2V-DV	-.00	—	-1.26	—	.004	.76	
	2V-SG	-.00	—	-1.32	—	.004	.76	
Age 17-19	8V-DV	.00	.94	-1.17	—	.007	.75	68
	8V-SG	.00	.99	-1.07	—	.007	.73	
	2V-DV	.00	.86	-1.19	—	.007	.75	
	2V-SG	.00	.85	-1.15	—	.007	.77	
Older	8V-DV	-.02	.79	-1.23	.00020	.046	.88	62
	8V-SG	-.03	.90	-1.39	.00027	.061	.85	
	2V-DV	-.02	.71	-1.21	.00022	.052	.85	
	2V-SG	-.03	.84	-1.28	.00025	.056	.86	

Note. All coefficients for $DPINDX^2$ and $DPINDX \cdot \Delta\overline{Y}$ are statistically significant ($p < .0001$). The coefficients for BASEN are statistically significant ($p < .004$).

Each dot in Figure 1 represents the ordered pair (DPINDX, $\Delta\overline{Y}$) for a college in the sample. The distribution of dots in Figure 1 shows that at most colleges, DPINDX for older students was negative, but that this fact did not guarantee that separate subgroup predictions would reduce prediction bias: for, at some colleges (in the red regions) large negative values of $\Delta\overline{Y}$ resulted in negative values of $\Delta BIAS^2$. On balance, though, more colleges were in the green regions than in the red regions, and therefore, at most colleges, separate subgroup prediction equations led to reductions in prediction bias for older students.

Figure 2 is a similar plot for minority students. Note that at most colleges DPINDX for minority students was positive, but that positive values of $\Delta\overline{Y}$ resulted in negative values of $\Delta BIAS^2$. On the whole, a much larger proportion of colleges lie in the red regions of Figure 2 than in the red regions of Figure 1. This corresponds to the poorer performance of separate subgroup predictions for minority students than for older students.

ΔMSE

Regression models of the form

$$\Delta MSE = a + b \cdot BASEN + c \cdot VINDX + d \cdot DPINDX + e \cdot (DPINDX)^2 + f \cdot \Delta\overline{Y} + g \cdot (\Delta\overline{Y})^2 + h \cdot DPINDX \cdot \Delta\overline{Y} + error$$

were fit to the cross-validation statistics for the colleges in the different samples. As in the regression analyses of $\Delta BIAS^2$, outlier observations were deleted, and regression models for the four alternative prediction methods were developed from identical samples.

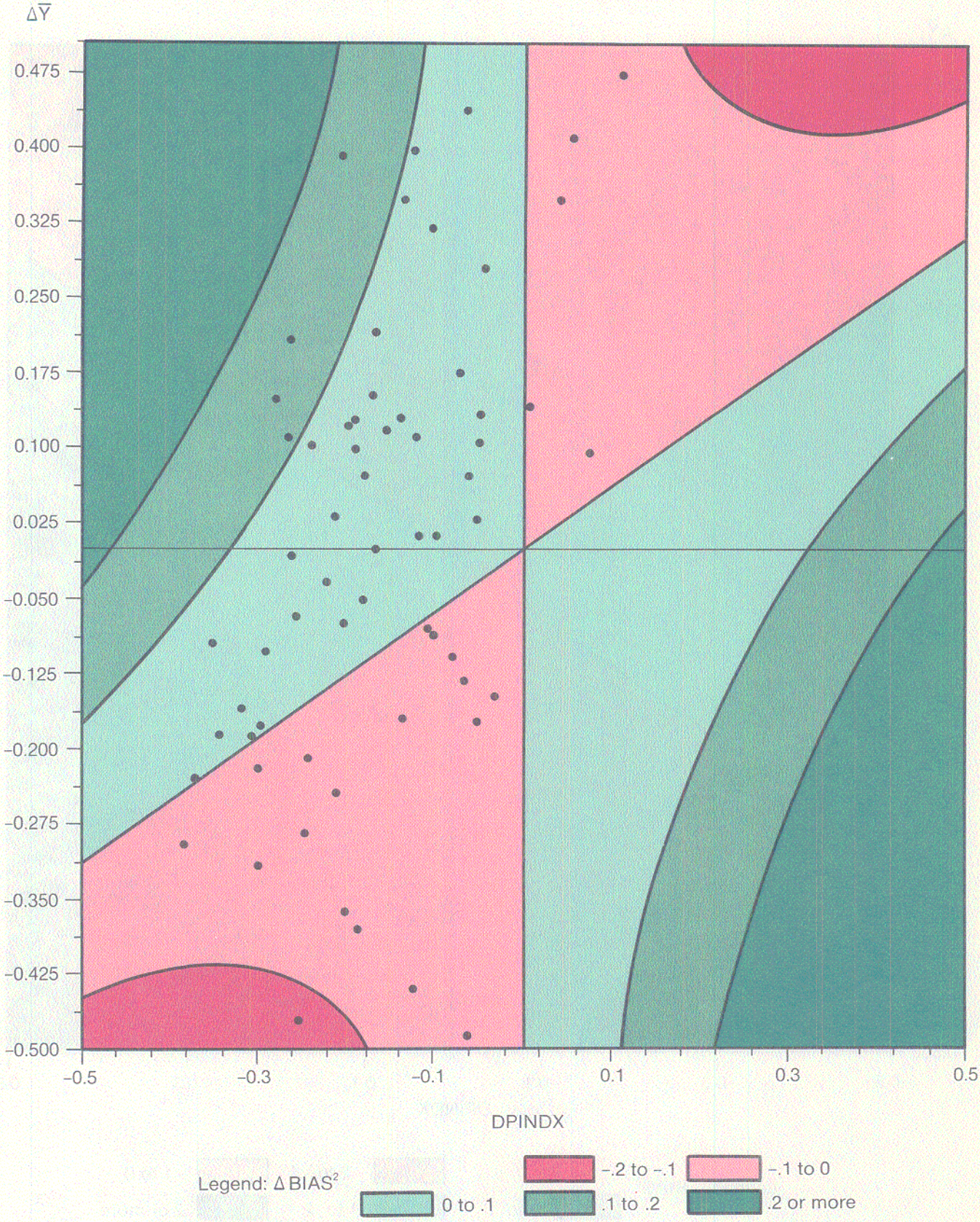

Figure 1. Predicted values of $\Delta BIAS^2$(8V-DV), given DPINDX and $\Delta\overline{Y}$ for older students.

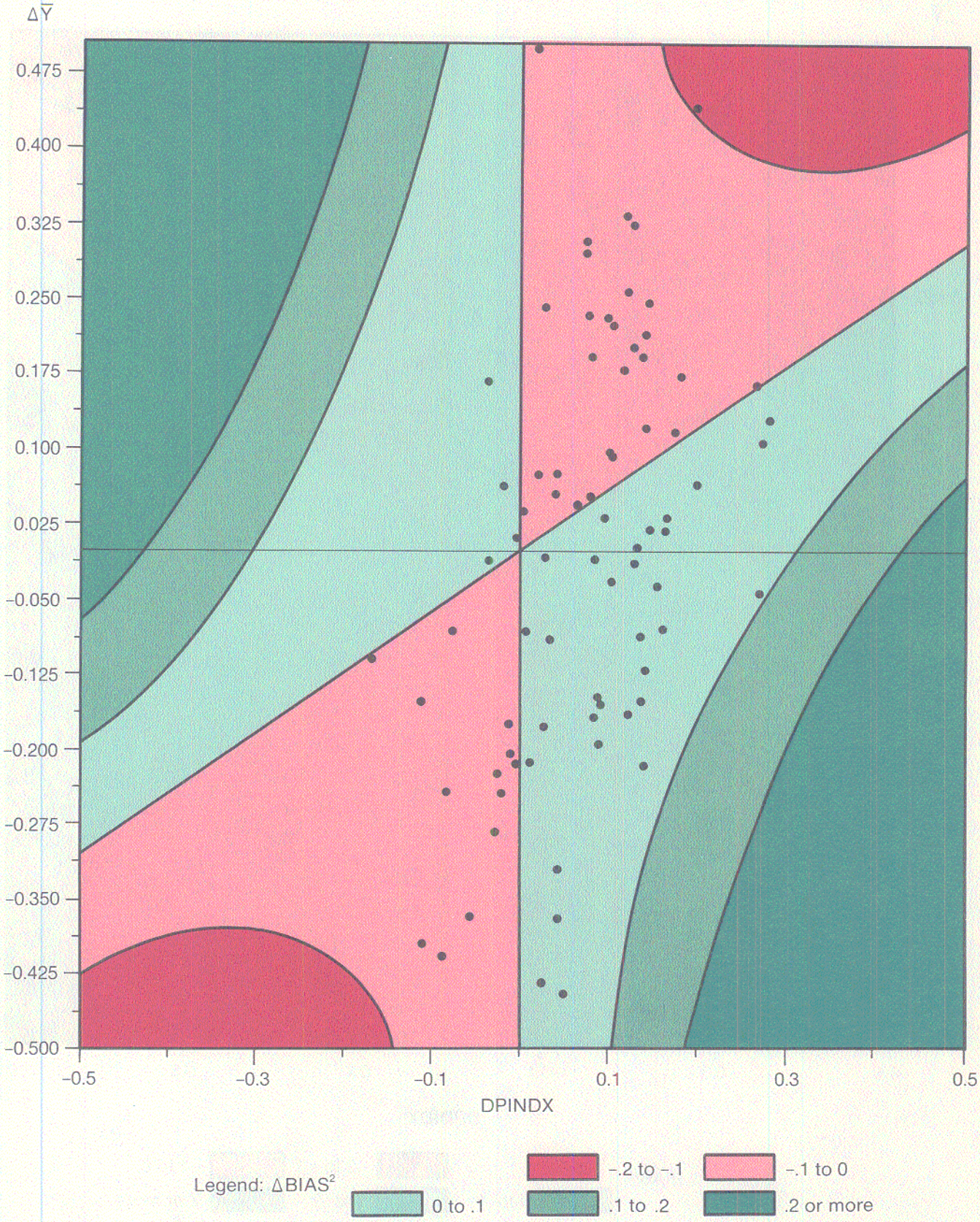

Figure 2. Predicted values of $\Delta BIAS^2(8V\text{-}DV)$, given DPINDX and $\Delta\overline{Y}$ for minority students.

The resulting regression coefficients are displayed in Table 18. For all subgroups and prediction methods, only a few regression coefficients in the full model were statistically significant ($p < .05$). The coefficients in Table 18 pertain to models in which only the variables with numerical coefficients are present; dashes for a regression coefficient indicate that the corresponding variable was omitted from the model.

For nearly every combination of subgroup and prediction method, the regression coefficient for the cross-product term was the largest and had the lowest p value. DPINDX or its square was usually the second largest and had the second lowest p value. VINDX was of moderate importance in a few instances for males and minority students, and BASEN was of moderate importance for older students. The multiple correlations associated with all these models were considerably smaller than the corresponding multiple correlation for predicting $\Delta BIAS^2$. These results suggest that the institutional characteristics considered in this study affect ΔMSE primarily through $\Delta BIAS^2$, and that there are other factors, not considered in this study, that are related to ΔMSE. It is not known what these other factors are.

Conclusions

In predicting college freshman grade average from ACT test scores and self-reported high school grades, prediction bias caused by differential prediction among student populations is dominated by prediction bias caused by changes over time in colleges' grading practices. Moreover, squared prediction bias, from whatever source, is typically much smaller than error variance in its contribution to mean squared error.

Dummy variable and separate subgroup equations based on age are typically effective in reducing both cross-validated prediction bias and mean squared error. Dummy variable and separate subgroup equations based on sex are marginally effective in reducing prediction bias and mean squared error. Using dummy variable and separate subgroup equations based on race is more often than not counter-productive in reducing bias and mean squared error.

The simpler methods for using demographic information in prediction (dummy variable instead of separate subgroup equations; two predictor variables instead of eight predictor variables) are usually more effective than the more complex methods. In particular, eight-variable separate subgroup equations often result in less accurate prediction than the other alternatives. This is not to say that separate subgroup analyses should never be done; on the contrary, they very often provide useful descriptive information about student populations. They are generally less effective in prediction, though, than dummy variable equations.

Following are recommendations for the student populations we investigated:

- Females: Using dummy variable predictions will more often than not reduce the underprediction for females and reduce mean squared error. Greater reduction in prediction bias occurs at colleges where mean grades are increasing or stable over time, and where the DPINDX statistic suggests more than the average amount of underprediction for females.
- Males: Using dummy variable predictions typically will slightly reduce overprediction for males. Greatest improvement occurs at colleges with larger than average DPINDX values and where mean grades are stable or decreasing over time. Separate subgroup prediction equations for males are not recommended.
- Minority students (blacks and Chicanos): None of the alternative methods is particularly successful in reducing prediction bias in colleges generally, and all the alternative methods tend to increase MSE. At colleges with stable mean grades over time and with DPINDX statistics that suggest very strong overprediction for minorities, the alternative prediction methods do tend to reduce prediction bias. Otherwise, none of the alternative methods is to be preferred over the total group predictions.
- Whites: Since white students are a large majority, bias in their predicted grade averages is very small to begin with. The alternative prediction methods are typically unsuccessful in reducing prediction bias and MSE.
- Students age 17-19: The alternative prediction methods are able to reduce the overprediction of the grade averages of traditional-age students, especially at colleges with stable or decreasing mean grades and with DPINDX statistics that suggest strong overprediction for this group. Since traditional-age students are a large majority, though, the amount of overprediction is small.
- Older students: The standard 8V-TG predictions typically underpredict the grades of older students. The alternative methods are usually successful in reducing the underprediction and in reducing mean squared error, particularly at colleges with large DPINDX values and stable or increasing grades. They are most successful when the base sample

TABLE 18

Regression Coefficients (and *p*-values) Associated With Multiple Regression of ΔMSE on Institutional Characteristics

Subgroup	Prediction method	Intercept	BASEN	VINDX	DPINDX	DPINDX^2	DPINDX$\cdot\ \Delta\overline{Y}$	SEE	Multiple R	Sample size
Female	8V-DV	.00	—	—	—	.82	−1.18	.017	.67	153
		(.08)				(<.0001)	(<.0001)			
	8V-SG	−.00	—	—	—	.41	−1.09	.034	.40	
		(.04)				(.14)	(<.0001)			
	2V-DV	.00	—	—	—	.38	− .74	.030	.43	
		(.01)				(.003)	(<.0001)			
	2V-SG	.00	—	—	—	.30	− .68	.003	.37	
		(.03)				(.03)	(<.0001)			
Male	8V-DV	−.01	—	.02	—	1.60	−1.99	.027	.79	153
		(.01)		(.06)		(<.0001)	(<.0001)			
	8V-SG	.02	—	−.08	—	.64	−2.11	.076	.53	
		(.18)		(.002)		(.33)	(<.0001)			
	2V-DV	−.02	—	.05	—	1.12	−1.46	.044	.64	
		(.008)		(.002)		(<.0001)	(<.0001)			
	2V-SG	−.02	—	.03	—	.64	−1.58	.056	.54	
		(.08)		(.07)		(.02)	(<.0001)			
Minority	8V-DV	.00	—	−.00	—	1.15	−2.01	.029	.83	77
		(.86)		(.85)		(<.0001)	(<.0001)			
	8V-SG	.06	—	−.15	—	.73	−1.36	.063	.51	
		(.02)		(.0003)		(.06)	(.0005)			
	2V-DV	.00	—	−.00	—	.95	−1.58	.034	.71	
		(.77)		(.88)		(<.0001)	(<.0001)			
	2V-SG	.02	—	−.03	—	.90	−1.57	.047	.58	
		(.34)		(.43)		(.001)	(<.0001)			
White	8V-DV	−.00	—	—	—	—	−1.45	.008	.63	83
		(.22)					(<.0001)			
	8V-SG	−.00	—	—	—	—	−1.42	.011	.49	
		(.29)					(<.0001)			
	2V-DV	−.00	—	—	—	—	−1.57	.014	.44	
		(.05)					(<.0001)			
	2V-SG	−.00	—	—	—	—	−1.88	.016	.48	
		(.10)					(<.0001)			
Age 17-19	8V-DV	−.00	—	—	.09	—	−1.07	.007	.74	69
		(.66)			(<.0001)		(<.0001)			
	8V-SG	.00	—	—	.08	—	− .97	.007	.71	
		(.95)			(<.0001)		(<.0001)			
	2V-DV	.00	—	—	.09	—	−1.16	.007	.77	
		(.90)			(<.0001)		(<.0001)			
	2V-SG	.00	—	—	.08	—	−1.10	.007	.78	
		(.60)			(<.0001)		(<.0001)			
Older	8V-DV	−.01	.00019	—	—	.79	−1.24	.046	.87	64
		(.08)	(.005)			(<.0001)	(<.0001)			
	8V-SG	−.04	.00031	—	—	.91	−1.35	.074	.80	
		(.005)	(.005)			(<.0001)	(<.0001)			
	2V-DV	−.01	.00021	—	—	.71	−1.22	.052	.85	
		(.16)	(.006)			(<.0001)	(<.0001)			
	2V-SG	−.02	.00024	—	—	.84	−1.29	.056	.85	
		(.02)	(.003)			(<.0001)	(<.0001)			

size for older students is larger than average (say, BASEN > 100).

In applying these recommendations at an institution one should, as was stated earlier, make certain that the intended uses of the predictions are educationally and ethically appropriate. One should also attempt to determine, from DPINDX and from local trends in grades, whether using demographic information would result in any practical increase in prediction accuracy. Finally, one should take into account any special local circumstances that could make the above recommendations inapplicable.

REFERENCES

The American College Testing Program. (1973). *Assessing students on the way to college: Technical report for the ACT Assessment Program.* Iowa City, IA: Author.

The American College Testing Program. (1983a). *ACT research services.* Iowa City, IA: Author.

The American College Testing Program. (1983b). *College student profiles: Norms for the ACT Assessment.* Iowa City, IA: Author.

Breland, H. (1979). *Population validity and college entrance measures.* Princeton, NJ: College Board Publication Orders.

Browne, M. N. (1975). A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology, 28,* 112-120.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*(10), 1067-1077.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Techometrics, 19*(1), 15-18.

Gamache, L., & Novick, M. R. (in press). Choice of variables and gender-differentiated prediction within selected academic programs. *Journal of Educational Measurement.*

Levitz, R. (1980). Summary-Survey of admissions practices for nontraditional-age freshmen. (Available from The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52243.)

Levitz, R. (1982). The predictability of academic achievement for nontraditional- and traditional-age freshmen (Doctoral dissertation, University of Michigan, 1982). *Dissertation Abstracts International, 43*/06, 1852-A.

Linn, R. L. (1978). Single-group validity, differentiate validity, and differential prediction. *Journal of Applied Psychology, 63*(4), 507-512.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*(1), 33-47.

Maxey, E. J., & Sawyer, R. (1981). *Predictive validity of the ACT Assessment for Afro-American/Black, Mexican-American/Chicano, and Caucasian-American/White students* (ACT Research Bulletin 81-1). Iowa City, IA: The American College Testing Program.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. *Journal of Educational Statistics, 7*(2), 91-104.

Sawyer, R., & Maxey, E. J. (1979a). The validity over time of college freshman grade prediction equations (ACT Research Report No. 80). Iowa City, IA: The American College Testing Program.

Sawyer, R., & Maxey, E. J. (1979b). The validity of college grade prediction equations over time. *Journal of Educational Measurement, 16*(4), 279-283.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin, 76*(4), 295-310.