Utility of Policy Capturing as an Approach to Graduate Admissions Decision Making

Frank L. Schmidt
U. S. Civil Service Commission

Raymond H. Johnson
Ford Motor Company

John F. Gugel
U. S. Civil Service Commission

The present study examined and evaluated the application of linear policy-capturing models to the real-world decision task of graduate admissions. Major findings were that (1) effectiveness of policy capturing was moderated by psychology sub-areas, with the experimental and clinical subgroups showing the highest and lowest predictability, respectively; (2) utility of the policy-capturing models was great enough to be of practical significance; and (3) least squares weights showed no predictive advantage over equal weights.

In psychology one body of research results that has not been fully exploited for its applications potential is that concerned with human judgment and decision making (Slovic & Lichenstein, 1971). Over 20 years of research have resulted in a number of verified principles of potentially great practical and social value. For example, the evidence is overwhelming that decision tasks requiring the integration and combination of information items to produce an overall judgment or prediction are better performed actuarially (using regression weights derived by analysis against the criterion measure) than by human decision makers (Meehl, 1954; Sawyer, 1966). It is also known that virtually all kinds of human decision makers (e.g., clinical psychologists, medical diagnosticians, stockbrokers) can be successfully simulated by linear models and that these models are as successful as, and sometimes more successful than, more complex non-linear and configural models in "capturing the policies" of human decision makers (Slovic & Lichenstein, 1971; Goldberg, 1971). Another well-verified finding is that the linear model of the human judge is often a better predictor of the actual outcome in question than is the judge from which the model was derived (Wiggins & Kohen, 1971; Goldberg, 1970). Dawes (1971) has termed this phenomenon "the bootstrapping effect."

The last of these findings appears to be the most remarkable and perhaps merits further comment. How can a model based on an individual's behavior do a better job than can the individual himself? Briefly stated, the answer is that while such essentially random factors as fatigue, boredom, and headaches cause the human judge to be inconsistent in the application of his/her (at least somewhat valid) judgmental principles, the linear model is always perfectly consistent in applying the judge's policy.

Essentially, it is the unreliability of human judges that makes them inferior to their own models. Empirical confirmation comes from a study by Dudycha and Naylor (1966), who found that subjects in a multiple-cue learning task applied appropriate relative weights to the various

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 3 Summer 1978 pp. 345–357

cues but made them so inconsistently as to make their judgments quite inaccurate. These researchers concluded that although humans may be used to generate decision strategies, they should usually not be used to apply them; they should be replaced by their own strategies in equation form.

Use of a model of human judges to replace the judges themselves has been termed "policy capturing." This procedure uses linear regression equations to predict, not actual outcome variables, but rather the judgment or decision of a judge or the joint decision of a group of judges. When policy capturing is successful, the linear model produces the same judgment or decision as the human decision maker, thus allowing substitution of the model for the judge. Hoffman (1960) called such linear models "paramorphic representations" of the judges' policies in order to avoid implying that the actual psychological process involved in making judgments was that of weighting variables. The linear model is, rather, a simulation of the judgmental process—a simulation which, because of the bootstrapping effect, usually produces more accurate "judgments" than judges themselves.

Goldberg's (1970) study is a classic, but typical, example of the bootstrapping effect. Twenty-nine clinical psychologists predicted the psychiatric diagnosis of 861 mental patients using MMPI profiles. A linear model was built to capture the policy of each clinician; of these models, 86% were more accurate predictors of the actual criterion diagnosis than the clinicians from whom the models were derived. There was no instance of persons being greatly superior to their own models.

The purpose of the present study was to demonstrate and evaluate the application of linear policy-capturing models to a real-world decision-making task—that of graduate admissions in a large psychology department. Because of the bootstrapping effect, such models of human decision makers are, in expectation, superior to the human judges themselves. In addition, because predictions by the models are more valid, they are by definition fairer to individual applicants. And finally, use of the models can be expected to lead to practical savings in faculty time, since in many cases the accept-reject decision can be made on the basis of model-produced predictions alone, thus freeing faculty members for more productive activities.

The Council of Graduate Schools estimated that there were over 7 million applications to U.S. graduate programs in 1970 (Dawes, 1971). Assuming review by four faculty spending an average of 10 minutes per application and an average 12-month faculty salary of $14,000 (which are rather conservative assumptions), Dawes (1971) estimated that approximately 4.67 million faculty hours with a salary value of 32.2 million dollars were spent reviewing these applications. In 1978 dollars, this would be approximately 60 million dollars. Many applicants, especially in the case of the more selective departments, have virtually no chance of being accepted. If an inexpensive and accurate psychometric procedure could be devised to identify and reject these applicants prior to faculty review, substantial savings in faculty time could be effected. For example, if an average of 40% of applicants to U.S. graduate programs could be rejected in this manner, the annual national savings would be approximately 1.87 million faculty hours or 46,750 faculty weeks. The salary value of this time is approximately 24 million dollars in 1978 dollars.

In a decision task of this sort, would not a straight actuarial approach (in which regression weights are derived against an external criterion measure) be superior to policy capturing? Even given the bootstrapping effect, Goldberg (1970) and Wiggins and Kohen (1971) found actuarial prediction to be superior to even the best of the linear models of the individual judges. Both studies recommend policy capturing (with its bootstrapping effect) as a method for improving decision accuracy only in situations in which criterion information is not available (eliminating the possibility of actuarial analysis).

While criterion information appears to be potentially available in the present case, unfortunately, restriction in range on both predictors

and criteria usually produces disappointingly low estimates of predictive validity for such variables as undergraduate GPA and GRE scores (Lannholm, 1968; Lannholm et al., 1968; Platz et al., 1959; Hyman, 1957). This is to be expected on purely logical grounds, since only selected applicants who are relatively similar to each other on indices considered can be studied. Formulae for estimating unrestricted validity, given the restricted validity, are available (Thorndike, 1949); but the large element of unreliability in the restricted validity estimate is carried over to the unrestricted estimates, often making such estimates quite unreliable *individually.*[1]

Without reliable estimates of the validity of individual predictors, regression weights cannot be accurately determined. In addition, because of differential faculty emphasis on different predictor indices, restriction in range in the selected group may be much greater on some predictors than others; the effect would be a distortion not only of validities, but also of regression weights (Whitla, 1968). It appears, then, that for this decision task, technical and measurement problems preclude meaningful use of the actuarial approach to decision making.

In a previous study, Dawes (1971) applied policy-capturing procedures to the graduate admissions process in the psychology department at the University of Oregon and found that the linear policy-capturing model predicted actual admissions committee ratings of applicants with a correlation of .78. This relationship allowed rejection of 55% of all applicants prior to faculty review with no false rejections. The present study is an attempt not only to replicate Dawes' (1971) findings, but also to extend them in two respects. First, the policy-capturing model was applied separately to individual areas (e.g., experimental, clinical) within a large psychology department to determine whether such subgrouping would produce increases in the model's efficiency. Here the hypothesis was that different programs within a large graduate department may, consciously or unconsciously, apply different admissions "policies" (i.e., have different predictor weight vectors) and that combining applicants from all programs into one large sample, as Dawes (1971) did, might result in decreased ability to model admissions decisions. Second, the present study attempted to assess empirically the number of faculty hours saved and the value of those hours in salary dollars when the linear policy-capturing model was used with a number of different decision strategies. This approach produced a more precise indication of the practical utility of the policy-capturing approach in the individual institution and related this utility to the specific decision strategies adopted.

Procedure

The subjects were 3,808 non-minority applicants to the psychology graduate program at Michigan State University for the years 1967–1971, inclusive. Number of applicants by year and interest group (IG) is shown in Table 1. Indices used as predictors of faculty admissions decisions were the three GRE Scores (Verbal, Quantitative, and Advanced) and undergraduate grade point average for the junior and senior years (UG-GPA).

The correlational procedure used was the linear discriminant function (*DF*); when applied to the predictors, it maximizes variance between groups while minimizing within-group variances. It thus uses the predictors to maximally separate the acceptee and rejectee groups. In this simple case in which there were only two

[1]Averages or medians of such estimates are more reliable. Schmidt (unpublished data) has corrected eight GRE validity coefficients reported in seven studies of success in graduate psychology programs for restriction in range. Mean of the corrected coefficients (using Fisher's transformation) was .45, indicating a quite useful level of validity. (Note that since such predictors are used, in practice, for selection from the unrestricted applicant sample, it is the unrestricted validity that is the index of the predictor's value.) These data directly contradict charges made by Marston (1970) and other critics to the effect that the GRE lacks validity.

Table 1
Number of Applicants and Proportion Accepted By Interest Group, 1967-71
(Total N=3,808)

Year	Clin.	Exper.	Soc.-Pers.	Indus.	Devel.	Quant.	Ecol.	Total Dept.[a]
1967	278	70	46	55	27	15		577
	(.40)	(.66)	(.59)	(.36)	(.52)	(.67)		(.46)
1968	326	101	89	59	35	9		629
	(.39)	(.60)	(.56)	(.47)	(.66)	(.67)		(.47)
1969	339	83	109	56	41	7		644
	(.34)	(.48)	(.58)	(.50)	(.66)	(.71)		(.44)
1970	357	77	71	45	42	13	8	622
	(.18)	(.21)	(.24)	(.47)	(.33)	(.50)	(.50)	(.23)
1971	411	61	66	43	74	8	29	683
	(.08)	(.44)	(.08)	(.28)	(.13)	(.53)	(.37)	(.15)
$\bar{X}$	384	69	69	44	58	11	11	653
70-71	(.13)	(.31)	(.16)	(.38)	(.22)	(.52)	(.40)	(.19[b])

[a]Total Dept. is greater than sum of interest groups because some applicants are not identified by IG.

[b]National Average is approximately .46 for APA approved doctoral programs, .53 for non-approved programs (Cates, 1972).

groups, the *DF* is mathematically identical to a multiple regression equation. The multiple correlation in question is a multiple point-biserial. In this study, however, because group membership was a matter of position on an underlying continuum and not a truly dichotomous variable, all point-biserial correlations were converted to biserial correlations to provide the best estimate of the Pearson correlation with continuous variables. This conversion also had the advantage of eliminating the effect of differential acceptance rates on the correlation.

Validation and Cross-Validation Procedures

Principles derived from earlier research (Schmidt, 1971) indicated that after applicants with incomplete data were removed, sample sizes were large enough to justify computation of separate *DF* weights only in the clinical, social-personality, industrial, and experimental interest groups (IGs). For each of these groups and for the total department, order of applicants was randomized with respect to year of application. Each group was then divided into a validation and cross-validation sample in approximately the proportions 60:40, respectively. Validation and cross-validation groups were checked for equivalence on sex, year of expected entry, proportion of acceptees, the three GRE scores, and UG-GPA. In the few instances in which significant differences were found, applicants were resorted randomly to eliminate these differences. Total final sample sizes in validation and cross-validation groups are shown in Table 2.

DF weights were derived on validation samples and then applied to the independent cross-validation samples to provide unbiased estimates of their effectiveness in general use. (Estimates of the separation in standard deviation units between acceptees' and rejectees' means are biased in the validation samples because the

Table 2
Sample Sizes Used in Validation and Cross-Validation Groups in Four Interest Groups and in Total Department

Group	Clin.	Exper.	Soc.-Pers.	Indus.	Total Dept.
Validation	849	193	183	132	1532
Cross-Validation	562	131	129	84	1014

weights derived have been fit not only to the real differences between the two groups, but also to error peculiar to the specific sample.)

In each of the five *DFs* computed, the F-max test was used to test for significance of differences in variances between acceptee and rejectee distributions. Since none of these tests reached significance, the mean of the acceptee and rejectee standard deviations on each *DF* was taken as the common standard deviation. A check of skewness and kurtosis indices indicated that the cross-validated *DF* distributions in rejectee groups were essentially normal for all five *DFs*. The skewness index ranged from .53 to −.51 (.00 indicates no skewness), and kurtosis varied from 4.14 to 3.31 (3.00 indicates perfect normality). This finding justified the use of normal curve tables in connection with these distributions. The acceptee *DF* distributions, however, were more skewed (median skewness = .92) and somewhat peaked in form (median kurtosis = 4.36). Consequently, the standard normal curve was not employed as the model for these distributions; empirical frequencies (in cross-validation samples) were used instead to determine proportions above and below various cutoff scores.

The *DF* functions were next corrected for unreliability in the original faculty accept-reject decisions. These decisions contained error variance resulting from disagreement between faculty; and this unreliability in the criterion acted to reduce the apparent effectiveness of the *DF* weights in separating rejectees from acceptees, thus causing an underestimation of *DF* effectiveness. The question is, how well can the *DF* predict faculty accept-reject decisions that are free of random error? In effect, this is what it will be attempting to predict in practice. Thus, in practice it will not be used to predict the partly unreliable faculty judgments, but rather each applicant's "true" faculty evaluation score. This "true score" can be conceptualized as the applicant's mean evaluation over an infinite number of faculty evaluators.

Some interest groups evaluated applicants in committee sessions, and thus independent evaluations were not available. In other groups applicants were prescreened by one faculty member, and only those judged promising were evaluated by the rest of the interest group faculty. It was feasible to obtain fully independent evaluations of an unrestricted sample of applicants only in the industrial IG. The reliability (Cronbach's Alpha) of ratings by 6 industrial faculty of 29 applicants was .86; this coefficient was used in all corrections for unreliability. It should be noted that this is a rather conservative strategy. The larger the coefficient used, the smaller the correction. Use of this relatively large coefficient insures against overcorrections (which would result in overestimates of *DF* effectiveness).

Determination of Cut Scores, Selection Ratios, and Computation of Decision Tables

It is obviously possible to use any subset of the infinite number of cutoff scores in assessing the effectiveness of each *DF*. In the interest of standardization and economy of time, it was decided to employ four cutoff scores which essentially covered the range in which such scores could reasonably fall and to keep them constant

for the five *DFs*. For each *DF*, Cut Scores 1, 2, 3, and 4 were set at 2.00, 1.50, 1.00, and .50 *SD* units, respectively, below the mean of the acceptee distribution. Because of differences between IGs in shape of acceptee *DF* distributions (some were more skewed and peaked than others), these cutoff scores resulted in slightly different rates of false rejection in different IGs. The standard score in the rejectee distribution corresponding to each cutoff was computed for each of the five *DFs* and converted to percents accepted and rejected using an ordinary normal curve table.

Table 1 shows the proportions of the total applicant pool accepted in each IG and the total department for each year in the period 1967–71. Because most IGs appeared to show systematic changes (decreases) in selection ratio over these six years, the selection ratios employed in computations in this study (except in converting point-biserial to biserial correlations) were averages of the 1970–1971 and 1971–1972 admission years. These figures refer to the proportion admitted without reference to support offered or acceptance-rejection by the applicant.

These selection ratio estimates were used to compute the percent of the total applicant pool that would be rejected in each interest group using each of the four cutoff scores. For each IG and the total department, total number of applicants for each year of the 1967–1971 period was determined (Table 1); and as with the selection ratio computation, the average of the 1970–1971 and 1971–1972 admission years was taken as the best estimate of current average number of applicants per year in each interest group. Number rejected was then simply the product of percent rejected and average number of applicants.

Estimation of Faculty Hour and Dollar Costs of Present Admission Procedures

A questionnaire was sent to all psychology faculty to obtain for each IG estimates of (1) average time spent per faculty member at each rank per applicant folder in individual review, (2) average time spent per applicant in group admissions meetings (if held), and (3) average proportion of IG members attending group admissions meetings (if held). Respondents were asked for information on (2) and (3) for their secondary, as well as primary, IG. Rank of respondents was also obtained. In IGs which employed a prescreening process whereby the least promising applicants were rejected after review by only a few faculty, the average percentage of applicants rejected and the number and rank of reviewing faculty was determined. There were 27 respondents, 21 of whom provided information on secondary, as well as primary, IGs. Respondents were well distributed across IGs and ranks.

Using a published salary list, the mean salary per hour in the department for each of the three ranks was computed. For IGs with no prescreening process, the following formulae were used to compute yearly faculty hours and the value of these hours in salary dollars:

> Faculty Hours (FH) at rank i = [(No. in rank) × (mean no. applicants) × (mean minutes/applicant, individual review) / 60] + [(no. in rank) × (mean proportion in attendance at group meetings) × (mean minutes per applicant, group review) × (mean no. applicants) / 60].
>
> Value in salary (VS) at rank i = (Rate/hr. at rank i) × (FH at rank i).

Total faculty hours (TFH) was then the sum of faculty hours across the three ranks, and total value in salary (TVS) was the sum of salary values across the ranks.

For each IG reporting a prescreening procedure, the formulae for FH was modified as follows:

> FH at rank i = [(No. prescreening at rank i) × (mean minutes/applicant, ind. rev.) × (mean no. applicants) / 60] + [(no. in rank i) × (mean proportion in attendance at group meetings) × (no. applicants not prescreened out) × (mean minutes per applicant, group review) / 60].

Other formulae remained constant.

TFH and TVS figures for each IG and the department as a whole are given in Table 3. Departmental figures are sums across all IGs. All figures are in 1972 dollars and would be over 50% greater in 1978 dollars.

Table 3
Estimates of Cost in Total Faculty Hours and Total Value in Salary by Interest Group of Reviewing Applicants

Group	Total Faculty Hours	Total Value in Salary
Clinical	744.48	8807.64
Experimental	359.09	3872.20
Soc.-Pers.	120.32	1367.89
Industrial	114.80	1355.74
Developmental	34.74	350.25
Quantitative	30.80	304.39
Ecological	271.98[a]	2789.01[a]
Dept. Total	1676.22	18347.12

[a]Reported time spent per applicant at admissions meeting was unusually long.

Results and Discussion

Table 4 shows the biserial *r*'s and the degree of separation between acceptee and rejectee groups in *SD* units in Validation and Cross-Validation samples for all interest groups and for the total department. The expected shrinkage is evident for all groups except the experimental group. There is apparently no ready explanation—other than an inexplicable sampling fluctuation—for the results shown by the experimental group in Table 4. The *DF* weights actually worked considerably better in the cross-validation than in the validation sample. Chance presence of a large number of highly unpredictable individuals (or unreliable faculty judgments) in the validation sample could have caused these anomalous results.

It is obvious that the *DF* was most successful in the experimental IG, where separation of 1.42 *SD* units was found between the two groups, (r_b = .71) and least successful in the clinical IG, where the between-groups separation was only .61 *SD* units (r_b = .34). These results indicate a possible continuum from experimental to industrial to social-personality to clinical in the extent to which admissions decisions were based on the four indices of applicant suitability used here. However, in terms of absolute levels, the relationships found in all IGs and for the total department were large enough to be of potential practical value.

Table 5 shows the percent of acceptees and rejectees that the *DF* would accept and reject at each cutoff score in each IG and the total department. For example, if Cutoff Score 3 was adopted by the industrial interest group, 58.7% of the actual rejectees would be rejected, while 41.3% would fail to be rejected. This would be achieved at a cost of falsely rejecting only 9.7% of acceptees. Because of the way in which the cutoff scores were determined, the percentage of acceptees that was falsely rejected is similar in all IGs. For example, at Cut-Score 2, this rate varied from 2.7% in the clinical IG to 5.0% in the experimental IG, a range of 2.3 percentage points.

The percent of true rejectees rejected by the *DF*, however, varied more widely from group to group. In the social-personality IG, for example, 52.0% of applicants who would be rejected by the faculty could be eliminated at a cost of rejecting 14.3% of acceptees. Virtually the same false rejection cost in the clinical IG (14.4%) allowed rejection of only 34.8% of eventual rejectees, a difference of 17 percentage points. Given the willingness to accept false rejection of from 9.7% to 14.4% of acceptees, the *DF* could be used (Cut-Score 3) to reject 34.8%, 52.0%, 58.7%, 65.9%, and 41.7% of rejectees in the clinical, social-personality, industrial, experimental, and total department groups, respectively. In terms of the total applicant pool, this would mean rejection of 32.2%, 46.0%, 40.1%, 49.6%, and 36.4% of all applicants in the clinical, social-personality, industrial, experi-

Table 4
DF Effectiveness Estimates in Validation and Cross-Validation Samples[a]

Interest Group	Validation Sample Biserial r	Validation Sample Group Diff. in SD Units	Cross Validation Sample Biserial r	Cross Validation Sample Group Diff. in SD Units
Clinical	.42	.84	.34	.61
Experimental	.54	.96	.71	1.42
Soc.-Pers.	.56	1.25	.51	.94
Industrial	.95	1.54	.63	1.23
Total Dept.	.44	.85	.43	.79

[a]All correlations corrected for attenuation due to criterion unreliability

Table 5
Performance of the Discrimination Function in Four Interest Groups and in the Total Department

Actual Faculty Decision	Discriminant Function Decision Cut 1 Acc.	Cut 1 Rej.	Cut 2 Acc.	Cut 2 Rej.	Cut 3 Acc.	Cut 3 Rej.	Cut 4 Acc.	Cut 4 Rej.
Clinical								
Accept	100.0	.0	97.3	2.7	85.6	14.4	65.8	34.2
Reject	91.7	8.2	81.3	18.7	65.2	34.8	46.0	54.0
Experimental								
Accept	100.0	.0	95.0	5.0	86.7	13.3	68.4	31.6
Reject	71.9	28.1	53.2	46.8	34.1	65.9	18.1	81.9
Social-Pers.								
Accept	100.0	.0	95.9	4.1	85.7	14.3	67.4	32.6
Reject	85.3	14.7	70.8	29.1	48.0	52.0	33.0	67.0
Industrial								
Accept	100.0	.0	96.8	3.2	90.3	9.7	64.5	35.5
Reject	77.9	22.1	60.9	39.4	41.3	58.7	23.6	76.4
Total Dept.								
Accept	99.7	.3	95.4	4.6	86.3	13.7	67.0	33.0
Reject	88.5	11.5	76.1	23.9	58.3	41.7	38.9	61.0

mental, and total department groups, respectively.

The savings value of the *DF* in TFH saved and VS for each IG and cutoff score is shown in Table 6. This table makes clear that when these practical indices were considered, the rank order of *DF* effectiveness in the various IGs was quite different from that indicated in Table 4. For example, if Cutoff Score 3 was employed, 40.1% of industrial, but only 32.2% of clinical, applicants were rejected by the *DF*. However, this means rejection of 124 clinical, but only 18 industrial, applicants. The savings in FH and VS was 239.7 and $2,836.06, respectively, in the clinical IG but only 46.0 and $543.65, respectively, in the industrial IG.

It is clear from Table 6 that practically significant savings would result from use of the *DF* in most IGs and in the total department, at least at the upper three cutoff scores. Use of Cut Score 2, for example, by the department as a whole would produce a yearly savings of approximately 340 FH, or about 8.5 faculty weeks. Value in salary of this savings would be almost $4,000. Department-wide use of Cut Score 3 could be expected to save about 610 FH annually, or approximately 15.3 faculty weeks; value of savings in salary would approach $7,000. Savings resulting from use of Cut Score 4 were, of course, even higher. In 1978 dollars, all savings figures would be more than 50% greater.

Table 4 appears to indicate that predictors of faculty admissions decisions by interest group was more effective than department-wide prediction in all but the clinical interest group. In the original uncorrected point-biserial correlations, however, the subgroup-total department difference was significant ($p < .001$) only for the experimental group. In addition, the clinical group was significantly less predictable than both the experimental and industrial groups ($p < .01$). Although it was not as strong as one might have supposed (probably because of less than optimal statistical power), there is evidence that predictability was moderated by subgroup; that is, there is evidence that the faculty in different interest groups differed in the extent to which they relied on the four indices of applicant quality examined in this study. Failure to examine the data by interest group would have masked these differences.

The standardized weights shown in Table 7 seem to indicate that the different interest groups did in fact follow different admissions policies. In no two groups was the rank order of the *DF* weights the same, although all groups (but not the departmental equation) agreed in assigning the lowest weight to GRE-Quantitative. The decision policy of the industrial group appears particularly different from the others; this group apparently assigned almost no weight to the two aptitude sections of the GRE and a very large weight to UG-GPA. Upon casual examination, the evidence seems to indicate that separate prediction of admission decisions by psychology subgroup can improve the effectiveness of policy-capturing models.

But is the conclusion warranted? Is prediction better in the experimental group, for example, because the weights used are those peculiar to that group—that is, because the specific policy of that group has been "captured"? Or is prediction better in this group because faculty in this group *generally* rely more heavily on GREs and UG-GPAs, with the specific set of weights relatively unimportant? It has repeatedly been found (Dawes & Corrigan, 1974) that in most psychological data, when predictors are positively correlated, different sets of positive weights tend to yield very similar correlations. It has further been found that equal weights (that is, a summing of standardized scores) often produces correlations in new samples or in the population equal to, or even greater than, those resulting from regression weights (Dawes & Corrigan, 1974; Schmidt, 1971).

Schmidt's (1971) results indicate that when the total sample size on which the weights are derived is in the neighborhood of 25 times as great as the number of predictors, least squares weights will, on the average, be at least slightly superior to equal weights (given the absence of

Table 6

Percent and Number of Total Applicants That Would Be Rejected By the Discriminant Function and Resultant Savings in Four Interest Groups and Total Department for Four Cut-Off Scores

	Cut Score 1				Cut Score 2			
Interest Group	Percent Rejected	Number[a] Rejected	Faculty Hours Saved	Dollar Value in Salary	Percent Rejected	Number[a] Rejected	Faculty Hours Saved	Dollar Value in Salary
Clinical	7.1	27	52.9	625.34	16.6	64	123.6	1462.00
Experimental	19.4	13	69.7	751.21	33.8	23	121.4	1308.80
Social-Personality	12.3	9	14.8	168.25	25.1	17	30.2	343.34
Industrial	13.7	6	15.7	185.74	25.6	11	29.4	347.07
Total Department[b]	9.4	61	157.6	1771.63	20.2	132	338.6	3807.12
	Cut Score 3				Cut Score 4			
Interest Group	Percent Rejected	Number[a] Rejected	Faculty Hours Saved	Dollar Value in Salary	Percent Rejected	Number[a] Rejected	Faculty Hours Saved	Dollar Value in Salary
Clinical	32.2	124	239.7	2836.06	51.4	197	382.7	4527.13
Experimental	49.6	34	178.1	1920.61	66.3	46	338.1	2567.27
Social-Personality	46.0	32	55.4	629.23	61.5	42	74.0	841.25
Industrial	40.1	18	46.0	543.65	60.9	27	69.9	825.65
Total Department[b]	36.4	238	610.1	6860.35	55.7	364	932.0	10497.85

[a]Based on average yearly number of applicants, 1970-1971.

[b]Faculty hour and cost totals used are sums across all interest groups, including the three not shown here.

Table 7
Standardized Discriminant Function Weights in Four Interest Groups and Total Department

	Standard Score Weights			
Interest Group	GRE-V	GRE-Q	GRE-A	U.G.-GPA
Clinical	.453	.338	.722	.399
Experimental	.493	.227	.482	.688
Social-Personality	.631	.178	.583	.480
Industrial	.080	.043	.279	.956
Total Department	.238	.578	.674	.394

suppressor variables, which was the case here). Below this ratio, equal weights will typically perform better. The industrial group had the smallest validation sample in this research, and in this group there were 33 times as many subjects (132) as predictors (4; see Table 2). Thus, it would be expected that least squares weights would be superior to equal weights in all groups and in the total department. Further, it would be expected that least squares weights derived on a given group would work better (in the cross-validation sample) than weights derived on other groups.

To examine these questions, each group's weights were applied to all other groups. In addition, all predictors were standardized, and the sum of standard scores was correlated with the criterion in each group; the results are shown in Table 8. Only one group—experimental—was best predicted by its own weights, and even here the difference was trivial. In all other groups, including the total department, weights derived on other groups performed as well or better than the group's own weights. From an overall point of view, all sets of weights performed about equally well, with the possible exception of the industrial weights (which were derived on the smallest validation sample).

The order of predictability of the interest groups was almost the same regardless of which set of weights was used. The experimental interest group was always the most predictable, and the clinical group was always the least predictable, followed by the total department (with one exception). Equal weights were slightly superior

Table 8
Effectiveness of Discriminant Function Weights Derived in One Group and Applied to Other Groups[a, b]

Group Weights Applied To	Group Weights Derived On					
	Clinical	Exper.	Soc.-Pers.	Indus.	Tot. Dept.	Unit
Clinical	.34	.35	.35	.29	.32	.33
Exper.	.67	.71	.69	.66	.65	.68
Soc.-Pers.	.55	.51	.51	.35	.59	.57
Indus.	.64	.70	.68	.63	.63	.68
Tot. Dept.	.43	.43	.43	.35	.43	.44
$\bar{r}_b$	.53	.54	.53	.46	.53	.54

[a]All correlations corrected for attenuation due to criterion unreliability.
[b]All data used is from cross-validation samples.

to own weights in the social-personality and industrial groups and trivially inferior to own weights in the clinical and experimental groups. In the total department group, the group with the largest validation sample (1,532), unit weights were superior to least squares weights by .01. (It is not suggested here that these small differences are in any sense "real.")

These results indicate that although the groups may differ in predictability—that is, in the extent to which faculty rely on GRE scores and UG-GPAs in admissions decisions—nothing is to be gained by attempting to capture the decision policies of the individual groups. Further, nothing is to be gained by use of total departmental regression weights; use of equal weights can be expected to be at least as effective. These findings obviously indicate that the data in Tables 5 and 6 on decision errors and dollar savings, respectively, would remain virtually unchanged if unities were substituted for the weights in Table 7.

In most graduate departments, the large sample sizes used in computing the discriminant function weights in this study will not be available. Weights based on smaller samples can be expected to be less effective relative to unit weights than those in this study. This consideration, combined with the findings in Table 8, lead to the recommendation that graduate departments considering use of a similar technique employ equal weights only. This study has demonstrated that the information contained in GRE scores and UG-GPAs can be used in admissions decisions to save significant amounts of money and faculty time. The results in Table 8 show that it is not even necessary to compute and/or use least squares weights to realize these savings; that is, the technique is even easier to set up and use than had originally been thought.

References

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, 1971, *26*, 180–186.

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin*, 1974, *81*, 95–106.

Dudycha, L. W., & Naylor, J. C. Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance*, 1966, *1*, 110–128.

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin*, 1970, *73*, 422–432.

Goldberg, L. R. Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance*, 1971, *6*, 458–479.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin*, 1960, *57*, 116–131.

Hyman, S. R. The Miller Analogies Test and University of Pittsburgh Ph.D.'s in psychology. *American Psychologist*, 1957, *12*, 35–36.

Lannholm, G. V. *Review of studies employing GRE scores in predicting success in graduate school, 1952-1967.* (GRE Special Report No. 68–1.) Princeton, NJ: Educational Testing Service, 1968.

Lannholm, G. V., Marco, G. L., & Schroder, W. B. *Cooperative studies of predicting graduate school success.* (GRE Special Report No. 68–3.) Princeton, NJ: Educational Testing Service, 1968.

Marston, A. R. It is time to reconsider the Graduate Record Examination. *American Psychologist*, 1970, *26*, 653–655.

Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and review of the literature.* Minneapolis, MN: University of Minnesota Press, 1954.

Platz, A., McClintock, C., & Katz, D. Undergraduate grades and the Miller Analogies Test as predictors of graduate success. *American Psychologist*, 1959, *14*, 285–289.

Sawyer, J. Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 1966, *66*, 178–200.

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 1971, *31*, 699–714.

Slovic, P. & Lichenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, *6*, 649–744.

Thorndike, R. L. *Personnel selection.* New York: Wiley, 1949.

Whitla, D. K. Evaluation of decision making: A study of college admissions. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences.* Reading, MA: Addison-Wesley, 1968.

Wiggins, N., and Kohen, E. S. Man vs. model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology,* 1971, *19,* 100–106.

Acknowledgments

The opinions contained herein are those of the authors and do not necessarily represent the official policy of the U.S. Civil Service Commission. Earlier work on this research was carried out while the first author was at Michigan State University.

Author's Address

Frank L. Schmidt, Personnel Research and Development Center, U. S. Civil Service Commission, 1900 E. Street, N. W., Washington, DC 20415.