Prediction of Recorded Accidents and Violations Using Non-Driving Predictors

EDWARD LEVONIAN, HARRY W. CASE, and RAYMOND GREGORY, University of California, Los Angeles

• TRAFFIC ACCIDENTS and violations are studied in relation to many variables, such as personality, attitudinal, sensory-motor, and physiological variables. Many studies involve several of these variables, and in such studies a common approach is to analyze each of these variables in turn in relation to accidents or violations.

In many such studies each analysis is made for the sample as a whole, whereas in other studies each analysis is made for a subsample homogeneous with respect to the other variables. When the entire sample is employed, and when each analysis involves a simple correlation or variance ratio, then the result is not controlled for other variables. When a homogeneous subsample is employed, the result, although controlled for the other variables, is less reliable.

Multivariate analysis not only provides for statistical control but also retains that reliability associated with the entire sample. An important consequence is the likelihood that the prediction of accidents and violations will be enhanced when several predictors are used. An even more important consequence of multivariate analysis is the clarity with which results can be interpreted; this last consequence accrues from the fact that multivariate results can be presented in an integrated fashion, rather than piecemeal.

Prediction is one criterion of knowledge. Unless one can predict accidents, in some sense, above the chance level, one cannot meaningfully claim to understand the processes involved in accidents.

The present study is an attempt to predict recorded traffic accidents and violations using objective multivariate techniques. The predictor variables were limited to those for which data could be collected before licensing.

The design of the study was simple. The subjects were divided randomly into two groups: a predictor group and a predicted group. Data from the predictor group were used to compute an objective criterion for predicting whether each subject in the predicted group had, or did not have, an accident during the preceding three years. Such a prediction was made for each subject in the predicted group, and this prediction was tested against the subject's accident record. An identical procedure was followed for violations.

METHOD

Subjects

The sample consisted of 720 truck drivers working for companies with headquarters in California. The companies were of two types: commercial carriers and those with fleets. The 720 drivers involved in this study included all persons examined at the Driver Testing Center, a facility of California Trucking Associations, Inc., during a 6-week period in the spring of 1960 and for whom measures were available on all the variables listed in the next section. Most subjects had come to the Center under the periodic testing program of their employing companies; the remaining subjects were applicants for driver positions. The test procedure was the same for both types of subjects. All data were collected under the prevailing standardized conditions, and the subjects were not aware that a study was in progress.

Paper sponsored by Committee on Highway Safety Research.

Independent Variables

The study involved 27 independent variables. These variables are capitalized in this report whenever they are defined by the descriptions and operations given in this section. The variables are divided into six categories; the 22 variables subsumed under the first five categories were specified before analysis began, whereas the five variables under the last category were determined from initial analyses involving the predicted group.

Personal. —The four variables in this category were Marital Status (score of 0 for single, 1 for married), Age (in years at the time of testing), Weight (in pounds), and Height (in feet, with inches expressed in decimal equivalents).

Cognitive. —The three variables in this category were Knowledge of Regulations (20-item open-book test of the Interstate Commerce Commission regulations, no time limit, and with subject's score representing the number of items answered correctly), Form Identification Speed (L. L. Thurstone's Identical Forms Test, Research Form 1F47, distributed by Science Research Associates, involving 60 items in which subject is required to match a semi-abstract pattern against five alternatives, and with subject's score representing the number of items completed in $4\frac{1}{2}$ min), Form Identification Accuracy (same test, with subject's score representing the number answered correctly in $4\frac{1}{2}$ min).

Visual. —The 11 variables in this category were Vertical Imbalance (as measured according to the Keystone Telebinocular Driver Vision Series, with a score of 0 for below normal, 1 for normal), Lateral Imbalance (same), Right Visual Acuity (Keystone Telebinocular Driver Vision Series scoring), Left Visual Acuity (same), Right Side Vision (subject focuses both eyes on a fixed object straight ahead while a white object is moved from a point initially behind subject along a perimeter in the plane of subject's eyes, with subject's score representing the angular location, with respect to the line straight ahead, of the moving object at the instant the subject reports detection of the object), Left Side Vision (same, but white object moves in from left side), Depth Perception (Keystone Telebinocular Driver Vision Series, with a score of 0 for below normal, 1 for normal), Distance Judgment (subject controls movement of a black block against a white background and attempts to bring block parallel with a similar block fixed at 20 ft, with subject's score representing the alignment error in millimeters), Night Vision (Porto-Glare device which involves five letters, O D Q G C, similar in size to 20/40 letters on Snellen eye chart and a 10-step light source behind each letter, with subject's score representing the minimum intensity or step, starting from the lowest, at which subject first reports the letter correctly when viewed at a distance of 10 ft in a semi-darkened room), Glare Recovery (same device presents glaring lights for about 5 sec while subject looks slightly down and away from lights, with subject's score representing the time in seconds between the termination of the lights and the subject's correct reporting of a display letter), and Glasses During Test (score of 0 for no glasses, 1 for glasses).

Motor. —The two variables in this category were Foot Reflex Time (left foot rests initially on the brake pedal of a simulated driver compartment, with subject's score representing the time in seconds between the presentation of a flash of red light and the depression of the pedal), and Foot Braking Time (right foot rests initially on the accelerator pedal, with subject's score representing the time in seconds between the presentation of a flash of red light and the depression of the brake pedal).

Medical.— The two variables in this category were Diastolic Blood Pressure (measured by a physician using the conventional procedure involving a sphygmomanometer), and Systolic Blood Pressure (same).

Personality. —The five variables in this category were determined in the course of this study, and were based on responses to a questionnaire prepared by the Institute of Transportation and Traffic Engineering. This 25-item questionnaire presented the respondent with driving situations involving personal interaction. From past experience it was found that other items of the same format were as readily answered by non-drivers as by drivers (1). It was assumed that responses to these driving items were indicative of personality, and that these responses could be obtained before licensing. The subject's scores on these five variables are discussed later.

TABLE 1

RECORDED ACCIDENTS AND VIOLATIONS OF PREDICTOR GROUP IN RELATION TO PREDICTOR VARIABLES

Variable	Mean	Standard Deviation	Recorded Accidents				Recorded Violations			
			Order of Emergence	Corr. with Rec. Acc.	Proportion of Variance	Proportion of Variance	Order of Emergence	Corr. with Rec. Viol.	Proportion of Variance	Proportion of Variance
Personal:										
Marital Status	0.886	0.318	16	-0.0463	0.0016		9	-0.0793	0.0046	
Age	36.514	9.281	7	-0.0482	0.0032	0.0035	1	-0.1991	0.0336	0.0344
Weight	176.461	25.343	27	0.0111	0.0000		8	0.0520	0.0042	
Height	5.858	0.201	4	0.0966	0.0060	0.0066	4	0.0901	0.0035	0.0077
Cognitive:										
Knowledge of Regulations	17.286	2.067	22	-0.0285	0.0004		2	-0.1408	0.0171	0.0177
Form Identification Speed	41.794	10.153	19	0.0530	-0.0041		3	-0.0409	0.0066	0.0038
Form Identification Accuracy	39.328	10.253	11	0.0637	0.0079		19	-0.0237	-0.0015	
Visual:										
Vertical Imbalance	0.986	0.117	13	-0.0298	0.0014		22	0.0390	0.0005	
Lateral Imbalance	0.844	0.363	5	0.0799	0.0062	0.0064	16	-0.0416	0.0014	
Right Visual Acuity	19.369	6.722	26	-0.0160	0.0000		13	0.0178	0.0009	
Left Visual Acuity	19.706	7.921	17	0.0139	0.0004		14	-0.0279	0.0015	
Right Side Vision	89.567	2.423	21	0.0071	0.0001		5	0.0954	0.0112	0.0073
Left Side Vision	91.044	1.850	25	-0.0203	0.0001		7	-0.0152	0.0014	
Depth Perception	0.958	0.200	14	0.0250	0.0015		15	0.0271	0.0014	
Distance Judgment	2.764	1.617	12	-0.0175	0.0010		18	0.0176	0.0006	
Night Vision	5.258	0.481	9	0.0791	0.0056		25	-0.0297	0.0003	
Glare Recovery	2.404	0.418	10	0.0386	0.0024		27	-0.0028	0.0000	
Glasses During Test	0.161	0.368	3	0.0912	0.0076	0.0077	26	-0.0395	0.0001	
Motor:										
Foot Reflex Time	0.215	0.019	2	-0.0485	0.0059	0.0055	10	-0.0174	0.0007	
Foot Braking Time	0.411	0.030	1	0.1028	0.0135	0.0150	17	-0.0057	0.0002	
Medical:										
Diastolic Blood Pressure	80.294	6.903	23	-0.0294	-0.0004		12	-0.1031	0.0044	
Systolic Blood Pressure	124.594	12.406	15	-0.0592	0.0024		6	-0.1043	0.0066	0.0061
Personality:										
Factor 1	2.847	1.746	6	-0.0730	0.0058	0.0058	24	-0.0058	0.0001	
Factor 2	4.156	1.500	24	0.0375	0.0002		21	0.0270	0.0005	
Factor 3	3.569	0.874	8	0.0710	0.0046	0.0049	11	0.0280	0.0017	
Factor 4	1.906	1.383	18	0.0007	0.0000		20	0.0366	0.0006	
Factor 5	3.372	0.988	20	-0.0126	0.0002		23	-0.0335	0.0004	
					0.0735	0.0554			0.1026	0.0770

The first two columns of Table 1 give the means and standard deviations of these 27 independent variables. The means of the five dichotomously scored variables (Marital Status, Vertical Imbalance, Lateral Imbalance, Depth Perception, and Glasses During Test) indicate the proportion of the subject's receiving a score of 1.

Dependent Variables

Recorded Accidents constituted the first dependent variable, with the subject's score representing the number of accidents involving (a) injury or (b) property damage in excess of $100 which were reported to the California Department of Motor Vehicles for the three-year period preceding the time data were collected on the 27 independent variables.

Recorded Violations constituted the second dependent variable, with the subject's score representing the number of violations reported to the DMV during the same three-year period. A distinction is made here between conviction and violation. A conviction is associated with a citing instance, whereas a violation is associated with each section of the California Vehicle Code which was cited for that instance. If a subject was stopped once but cited for speeding and failure to signal, he would receive one conviction but two violations.

PROCEDURE AND RESULTS

The procedure involved six steps: (a) division of the sample into two equal-size groups, (b) analysis of the questionnaire based on the data of the predicted group, (c) generation of two multiple regression equations based on the data of the predictor group, (d) use of the second equation to compute an accident score for each subject in the predicted group, (e) prediction for each subject in the predicted group of whether he had a recorded accident during the past three years, and (f) comparison of this prediction against the subject's accident record. Steps 3 through 6 were repeated for violations.

Step 1: Sample Division

Step 1 involved the division of the 720 subjects into two groups of 360 each. The division was performed by ordering the subjects from 1 to 720, then placing all the even-numbered subjects in the predictor group, and all the odd-numbered subjects in the predicted group.

The subjects were ordered according to their California chauffeur license number. No account was taken of a driver (non-chauffeur) license that the subject might have had, because both chauffeur and driver license numbers would have been identical, and, in fact, only one type of license has been issued by California since 1958. The subjects tended to be ordered by age, because California has issued permanent license numbers since 1944, and many subjects would have had their license number assigned when they first became of eligible age.

Step 2: Questionnaire Analysis

Step 2 involved the analysis of the 25-item questionnaire based on the responses of the predicted group. At the beginning of the study the number of questionnaire variables was not known. Because these questionnaire variables were to be included as independent variables in the prediction equation, it was necessary to specify these questionnaire variables before further analyses could be undertaken.

The analysis of the questionnaire resulted in the grouping of items according to statistical considerations: each group consisted of items which tended to correlate with each other, and the groups tended to be independent. The procedure involved a series of factor analyses. Each factor analysis started with Pearson correlations and involved the insertion of the square of the multiple correlation (of each item on the remaining items) in the principal diagonal of the correlation matrix. The limit on the number of factors was equal to the number of eigenvalues which were greater than zero (2). However, only those factors were rotated which had at least one loading of 0.20 or greater. These orthogonal factors were rotated by the Kaiser Varimax method (3), an analytic procedure that maximizes for the table as a whole the variance in the factor loadings. The variance in the loadings for each factor is computed, then summed over factors, and it is this sum which is maximized. Rotational iterations were continued until the difference between four successive sums failed to exceed 10^{-7}.

At the end of each factor analysis, an item was eliminated if it (a) had a communality less than 0.10, or (b) failed to have a loading greater than 0.20 in a factor that had at least one other loading greater than 0.20, or (c) had loadings greater than 0.20 in at least two factors. The three criteria, respectively, were intended to eliminate items that (a) shared little in common with the remaining items, (b) were associated only with

TABLE 2

QUESTIONNAIRE ANALYSIS: FINAL ROTATED FACTOR MATRIX

Item	Factor					
	1	2	3	4	5	h^2
3	-0.39	0.12	-0.09	0.08	0.07	0.18
5	0.46	0.06	-0.03	0.09	0.04	0.23
25	0.50	0.15	-0.09	-0.03	-0.03	0.28
11	0.00	-0.35	-0.09	0.06	-0.04	0.13
21	-0.02	0.28	0.06	-0.16	0.16	0.13
24	0.08	0.31	-0.08	0.03	-0.05	0.11
1	-0.01	0.04	-0.38	0.02	-0.01	0.14
10	0.02	-0.08	-0.35	0.05	-0.08	0.14
7	-0.05	0.01	0.00	0.32	-0.03	0.11
20	-0.06	0.14	0.10	-0.33	0.02	0.14
9	-0.19	0.01	-0.03	0.17	-0.26	0.14
19	0.08	-0.03	-0.06	0.00	-0.05	0.14

TABLE 3

QUESTIONNAIRE ANALYSIS: ITEMS IN FINAL FACTOR MATRIX

Factor 1	
3.	Do you think if you ever got in a serious accident it would more likely be your fault or the other person's fault? (my fault = 0, uncertain = 1, other person = 2)
5.	Do you feel that you are able to park a little better than, or about as well as most drivers? (as well as = 0, uncertain = 1, better than = 2)
25.	Would you say your driving is better than average or average? (average = 0, uncertain = 1, better than average = 2)
Factor 2	
11.	An old car is stalled ahead on the highway. The driver is waving, but you are not sure what he wants. Do you usually drive by, or do you stop to see what he wants? (drive by = 0, uncertain = 1, stop = 2)
21.	When you reach an intersection at the same time as a car approaching from the side street, do you usually wait for it to cross first, or do you try to cross first? (cross first = 0, uncertain = 1, wait for it = 2)
24.	Suppose while you're waiting at a signal, the car ahead of you rolls back and hits your bumper. Would you get out to see if your car was damaged? (no = 0, uncertain = 1, yes = 2)
Factor 3	
1.	Would you double park to let a passenger out even though it meant that the driver behind you would have to wait? (yes = 0, uncertain = 1, no = 2)
10.	Do you feel that you can exceed most speed limits without endangering yourself or others? (yes = 0, uncertain = 1, no = 2)
Factor 4	
7.	Suppose you are prepared to enter a parking space and another driver grabs the space. Do you sometimes tell him off? (no = 0, uncertain = 1, yes = 2)
20.	Suppose you have stopped in the street waiting for a driver to pull out of a parking space so that you can enter. A car behind you honks to get by. Do you move on and try to find another space, or do you stay where you are? (move on = 0, uncertain = 1, stay put = 2)
Factor 5	
9.	Suppose you are stopped in bumper-to-bumper traffic and the car ahead of you moves forward, but before you have a chance to move up yourself, the driver on your left cuts in front of you. Do you occasionally honk at him? (yes = 0, uncertain = 1, no = 2)
19.	Suppose you are waiting in the front row of a stop signal. After a long time you begin to feel that the signals must be stuck, but see that the other drivers are not moving. Would you cross the intersection against the signal? (yes = 0, uncertain = 1, no = 2)

item-specific factors, and (c) tended to have their variance contributed by more than one factor. Only one item was eliminated after each factor analysis, and analyses were continued until no item could be eliminated according to the three criteria.

It was necessary to perform 14 factor analyses before the remaining items passed all three criteria. At this time only five factors and 12 items remained. This final factor matrix is given in Table 2, with the factors and items rearranged for clarity. Table 3 presents the 12 items grouped by factor.

Having determined that five questionnaire variables would be included in the study, it was necessary to compute a score on each variable for each subject. A subject's score for a variable was determined by his responses to the items grouped under the associated factor. Each item had three choices, which were orignially assigned the values 0, 1, and 2, with the 1 representing the uncertain choice. This assignment was arbitrary, but some numerical assignment was necessary in order to perform the factor analyses. However, after the factor analyses were finished, the numerical assignment was reversed for each item whose major factor loading was negative (Table 2). These final numerical assignments are given in Table 3 in parentheses after each item. A subject's score was based on these final values. Scores ranged from 0-6 for the first two questionnaire variables, and 0-4 for the last three. A higher score indicates more of that characteristic which is measured by the variable.

Now, with five questionnaire variables established, it was possible to add them to the 22 other independent variables and to proceed with the generation of the multiple regression equations.

Step 3: Regression Analyses

Step 3 involved the generation of two multiple regression equations, each with Recorded Accidents as the dependent variable and based on the data of the predictor group.

The first multiple regression equation involved all 27 independent variables, and the results are given in Table 1.

The four columns under "Recorded Accidents" indicate (a) the order in which each independent variable emerged, (b) its correlation with Recorded Accidents, (c) its contribution to the variance in Recorded Accidents for the first multiple regression analysis, and (d) its contribution in the second multiple regression analysis.

The Order of Emergence column was determined step-wise, one independent variable at a time (4). At the first step that independent variable was selected among the 27 that accounted for the greatest proportion of variance in Recorded Accidents. At the second step that independent variable was selected among the remaining 26 that accounted for the greatest proportion of variance in Recorded Accidents not already accounted for by the variable selected in the first step. This procedure was repeated for 27 steps, and the numbers in the third column of Table 1 indicate the step at which the independent variable made its appearance.

The second multiple regression equation involved a subset of the 27 independent variables chosen on the basis of their contribution to the variance in the dependent variable: Recorded Accidents. The independent variables were added step-wise, in the order shown in the third column of Table 1, starting with Foot Braking Time. The 2nd, 3rd, ... kth independent variable was added successively as long as it increased the significance level of the F ratio of the contributed variance to the error variance. This criterion resulted in the inclusion of eight independent variables. The data from the predictor group on these eight variables were used to compute the second multiple regression equation, and it was this second equation which was used to compute an accident score for each subject in the predicted group.

The same procedure was followed for Recorded Violations as the dependent variable, and the results are given in the last four columns of Table 1. Using the criterion just given, the first six independent variables, according to their order of emergence, were included in the second multiple regression equation.

Step 4: Accident Score

Step 4 involved the computation of an accident score for each subject in the predicted group. A given subject's accident score was computed by multiplying his score on one of the eight independent variables by the regression coefficient for that variable, summing eight such products, and adding the constant. An identical procedure was followed for computing the subject's violation score, but in this case, of course, there were only six products to sum.

Step 5: Accident Prediction

Step 5 involved the prediction of whether a subject in the predicted group had a recorded accident during the previous three years. This prediction was based on whether his accident score, computed in Step 4, exceeded a critical accident score.

The critical accident score was computed from the predictor group data as follows: (a) an accident score was computed for each subject in the predictor group according to the procedure outlined in Step 4; (b) the predictor group was dichotomized on Recorded Accidents as close to the median as possible, resulting in a dichotomy point between no accidents and one or more accidents; (c) the accident scores of these dichotomized groups were used to compute a critical accident score by use of an equation derived by Guilford and Michael (5, 6).

The critical accident score turned out to be 0.758. Thus, a prediction of no accident was made for each subject in the predicted group whose accident score was less than 0.758, and a prediction of at least one accident was made for each subject whose accident score was greater than 0.758.

An identical procedure was followed for violations. The dichotomy point fell between 3 and 4 Recorded Violations over the previous three years. Thus, subjects in the predictor group with 0-3 violations fell below the dichotomy point, whereas those with 4 or more were assigned to the upper category. The critical violation score was 4.926.

Step 6: Accident Prediction Validation

Step 6 involved a test of significance of the relation between Recorded Accidents and predicted accidents of the predicted group. A fourfold table was formed with Recorded Accidents (none vs one or more) on one axis and predicted accidents (none vs one or more) on the other axis. Because the distribution was so extreme, the probability of such a distribution (and all other distributions more extreme) was computed directly (7), rather than determined indirectly through a Yates-corrected χ^2. The exact probability was 0.0298.

An identical procedure was followed for violations, but in this case a χ^2 could be computed legitimately. The prediction of violations was in the right direction, but the χ^2 of 1.31 was not statistically significant. The details of the accident and violation predictions are shown in Figure 1.

		Predicted Accidents: none	Predicted Accidents: 1 or more	
Recorded Accidents	1 or more	130	5	135
	none	224	1	225
		354	6	

$p < .03$

		Predicted Violations: 3 or less	Predicted Violations: 4 or more	
Recorded Violations	4 or more	109	69	178
	3 or less	122	60	182
		231	129	

$p < .25$

Figure 1. Recorded accidents (and violations) in relation to predicted accidents (and violations) for the predicted group.

ANALYSIS

Before presenting the interpretation of the substantive results, a brief discussion is presented concerning (a) the validity of the statistical procedure employed, and (b) the conditions of this study which made prediction more difficult.

Statistical Procedure

Two aspects of the procedure are likely to be of interest. First, product-moment correlations and multiple regression analyses were employed despite the fact that many of the variables were not normally distributed. Under this condition, the results of multiple regression analyses cannot be, and were not, tested for significance. The values given in Table 1 should be taken only as approximations of the values that would have resulted from normally distributed variables. Despite the procedural liberties taken with the analysis of the data of the predictor group, the probability of the relation between recorded and predicted accidents (and analogously for violations) for the predicted group is valid for the conditions under which it was determined--the probability was computed under the nonparametric conditions that prevailed.

Second, the predicted group was used for the analysis of the questionnaire, and the resulting five variables were used in the prediction equation, which was then applied to each subject in the predicted group. In general, a sample that is used preliminarily in such a way as to influence scores computed subsequently cannot, thereafter, be used in a statistical test involving these scores. In this study, however, the predicted group determined only the number of variables in the questionnaire, not the weights (coefficients in the prediction equation) to be attached to these variables. The alternative of using the predictor group to determine both the number of variables as well as their weights would most certainly have created a bias. Such a bias would occur because the factor analyses lead to an independence of the variables, an independence optimized with respect to the predictor group data. This independence in turn would result in a positive bias in the weights attached to these variables by the multiple regression analysis.

Factors Influencing Prediction

The results indicate that it is possible to predict at a statistically significant level whether each of 360 truck drivers had or did not have at least one recorded accident during a three-year period. This result emerged despite the composition of the sample and despite the predictor variables employed.

First, the sample was undoubtedly restricted in the range of Recorded Accidents, leading to a more difficult prediction of accidents. This restriction is the result of the common employment practice of not retaining employees who are involved in too many accidents. All subjects were employed by companies that had the freedom to fire those drivers whose performance proved inadequate. Truck accidents tend to involve considerable financial loss to the employer, though accident data do not reveal the extent of this loss, accident culpability, or the number of accidents that may have been incurred while the employee was off-duty and driving his own car. Another factor that tended to limit the range in the two dependent variables is the California negligent operator program. When recorded accidents and convictions reach a critical value, State authorities, after considering accident culpability and driving exposure, may restrict the driving privilege for a period of time.

Second, the predictor variables used were those on which data could be collected before licensing. It is relatively easy to predict accidents above the chance level if driving variables (previous exposure rate, accident rate, violation rate, etc.) are used for prediction. If the objective is prediction only, then these driving variables should be employed, but if the objective is to gain some appreciation of pre-driving variables associated with subsequent accidents, then the use of driving variables is likely to compound further an already resistant problem.

Of course, in the present study the data were not collected before licensing. Such a refinement in research design will become more desirable as the delineation of variables becomes more clarified as the result of studies using a cruder design, such as the one employed in the present study. The following discussion assumes that the relations obtained in the present study are only a crude approximation to the relations that would have been obtained if the data on the predictor variables had been collected before licensing, or at least before the three-year period to which the Recorded Accidents and Violations apply. But it is also assumed that even these crude approximations are likely to give direction to subsequent studies.

Variables Contributing to Recorded Accidents

The prediction of Recorded Accidents demonstrated in this study justifies a discussion of the independent variables involved in the prediction. This discussion is based only on the results from the predictor group, as given in Table 1, unless otherwise specified.

Before discussing independent variables individually, two generalizations are made. A comparison of columns 5 and 6 indicates that Foot Braking Time makes essentially the same contribution in the second regression analysis as it did in the first, and this constancy holds also for the other seven variables in column 6. This implies that the 19 variables that do not appear in column 6 did not act in the first analysis as suppressor variables with respect to the 8 that do appear.

Another generalization can be made on the basis of the totals of columns 5 and 6, as shown at the bottom of Table 1. Each total represents (a) R^2, the square of the multiple correlation of Recorded Accidents with the independent variables, and (b) the proportion of variance in Recorded Accidents accounted for by the independent variables. The eight variables account for much of the variance accounted for by all 27 variables. However, in each analysis, the independent variables account for only a small proportion of the total variance in Recorded Accidents.

The independent variables are discussed according to the six categories given in Table 1, beginning with the personal variables. Age correlated with Recorded Accidents in the expected direction—younger men tended to have more accidents--but age was not a dramatic predictor of accidents. This divergence from what might be expected probably lies in part on the employment selection process already discussed, and in part on the statistical control of other variables used in this study. The relation of age to recorded accidents is not clear, though the relation appears to be rather complex. Because more efficient functioning of the body tends to be associated with younger drivers, one might expect younger drivers to have fewer accidents. As everyone knows, quite the opposite is true. Obviously there are other factors associated with youth which tend to counteract the more efficient functioning of the body. Until the contribution of age to the variance in accidents can be completely eliminated, any claim to the understanding of the factors in the age-accident relationship should be considered conjectural.

The cognitive variables contributed essentially nothing to the variance in Recorded Accidents. The contribution of Form Identification Accuracy in the 27-variable analysis is due primarily to the suppressor effect of Form Identification Speed, with which it is highly correlated ($r = 0.96$). If Form Identification Accuracy or Form Identification Speed, but not the other, had been included in the second regression equation, its entry in column 6 would have been less than that which appears in column 5.

Of the visual variables, Lateral Imbalance and Glasses During Test contributed moderately to the variance in Recorded Accidents, as did Night Vision, although the latter was not one of the eight predictor variables. The positive correlation of Lateral Imbalance with Recorded Accidents indicates that subjects with normal lateral imbalance tend to be involved in more accidents. Also, the positive correlation of Glasses During Test with Recorded Accidents indicates that subjects who wore glasses during the visual tests (and presumably while driving) tend to be involved in more accidents.

Both motor variables were retained as predictor variables. Foot Braking Time emerged first on predictive order because it was the independent variable with the highest correlation with Recorded Accidents.

Neither of the two medical variables contributed to the variance in the dependent variable.

The last five variables have been designated as personality variables, though admittedly there is no critical evidence to uphold this interpretation. But even if the questionnaire factors do represent personality variables, the sparseness of the items makes difficult the identification of these personality variables. Personality Factors 1 and 3 made small contributions to the variance in Recorded Accidents. A high score on the three items of Factor 1 (Table 3) seems to imply self-confidence, and the negative correlation in Table 1 indicates that drivers with accidents tend to have less self-confidence. Factor 3 may be a social desirability factor—a tendency to respond so as to appear in favorable light. In any case, drivers with accidents tend to have higher scores on this factor.

To ascertain the accident results given in Table 1, a similar table was prepared, but with the results based on the combined data of the predictor and predicted groups. The five questionnaire variables were not included in the analysis, because these variables were defined originally on the basis of the predicted group. The results from the remaining 22 independent variables were similar to those already given. In a rough sense, accidents are more likely to be associated with the younger driver with glasses who is slower in braking.

Variables Contributing to Recorded Violations

It is recalled that the prediction of Recorded Violations was in the right direction but was not statistically significant. The failure to achieve a higher prediction was due in large part to the shrinkage that resulted from dichotomizing (so as to retain the same procedure for violations as was used for accidents) Recorded Violations, a variable that is essentially continuous with substantial frequencies in violation categories 0, 1, ... 10 for both the predictor and predicted groups. Evidence of this shrinkage is given by the product-moment correlation between Recorded Violations and violation scores (also continuous) for the predicted group. This correlation of 0.14 was significant at the 0.01 level, indicating that the basis for the computation of violations scores has validity. This result suggests that Recorded Violations could have been predicted at a sufficiently high level as to warrant a discussion of the variables that entered into the prediction equations. Again, this discussion is based only on the results from the predictor group as given in Table 1, unless otherwise noted.

First, Table 1 reveals that three of the six prediction variables show different contributions to the variance in Recorded Violations for the two multiple regression analyses (the 27-variable and the 6-variable analyses). There are two cases of a decrease in the 6-variable analysis, and one case of an increase. These three cases are discussed separately.

In considering Right Side Vision, which accounts for less variance in the 6-variable analysis, the enhanced contribution in the 27-variable analysis is probably due to the suppressor action of Left Side Vision. This action tends to inflate the apparent contribution of Right Side Vision.

Form Identification Speed also accounts for less variance in the 6-variable analysis. Again, the explanation probably lies in the suppressor action contributed by Form Identification Accuracy in the 27-variable analysis. Such a decrease was already anticipated in the preceding discussion on accidents. Despite this reduction, the selection of variables to be included in the prediction equation should be based on predictive order, rather than on the proportion of variance revealed in the 27-variable analysis. Had this latter criterion been used, Height would not have been included in the prediction equation, and its eventual contribution would have been lost.

For Height, the increase in contribution for the 6-variable analysis is probably due to the common variance between Weight and Height. The sum of the variance contributed by Weight and Height in the 27-variable analysis appears to be contributed by Height alone in the 6-variable analysis.

These cases highlight an interesting question: if two variables in the 27-variable analysis share a common variance, and if only one of these variables is retained in the 6-variable analysis, why does one retained variable show a decrease, whereas another retained variable show an increase? The answer lies in the extent to which the variance shared by the independent variables is also common to the variance that each independent variable shares with the dependent. Often the correlations among the independent variables and the dependent variable allow one to predict the effect of eliminating one of the independent variables.

Consider first the correlations between Right Side Vision and Left Side Vision (r = 0.47), between Right Side Vision and Recorded Violations (r = 0.10), and between Left Side Vision and Recorded Violations (r = 0.02). Although Left Side Vision has an essentially zero correlation with the dependent variable, it has a substantial correlation with a variable (Right Side Vision) which is correlated with the dependent variables. Under these conditions, Left Side Vision is likely to suppress that variance which is shared by Right Side Vision and Left Side Vision but which is not shared by Right Side Vision and Recorded Violations. Left Side Vision becomes a suppression variable, increasing the contribution of Right Side Vision. If Left Side Vision were to be removed, the contribution of Right Side Vision would be decreased, as in fact occurs in the 6-variable analysis. Thus, an independent variable that correlates near-zero with the dependent variable may still aid in predicting the dependent variable.

In considering next the correlations between Weight and Height (r = 0.53), between Weight and Recorded Violations (r = 0.35), and between Height and Recorded Violations (r = 0.09), the correlation between the independent variables is essentially the same as in the previous case, but unlike the first, the two correlations of the independent variables with the dependent are rather similar in magnitude. Under these conditions, the elements that Weight shares with Recorded Violations are also likely to exist in the Height variable. If so, when Weight and Height are included in the same analysis, they are likely to split the total variance they share with Recorded Violations. If one of the independent variables is removed, the retained variable will contribute most of the variance previously contributed by the two. Thus, when Weight is eliminated in the 6-variable analysis, Height accounts for as much of the variance in Recorded Violations as was previously accounted for by Weight and Height combined. Thus, for prediction purpose two independent variables are not necessarily better than one.

Form Identification Speed emerged third on predictive order, despite the fact that its correlation with Recorded Violations was exceeded by seven other variables which emerged later. This occurred because only Form Identification Speed contributed variance to Recorded Violations not already contributed by Age, which had already emerged. This is substantiated by the fact that only Form Identification Speed correlated significantly with Age in a direction opposite to what would be expected on the basis of the correlations between Recorded Violations and the independent variables being considered. Again, this points out the limitations of considering only the correlation with the dependent variable as the criterion for the selection of predictor variables.

The relative contribution of the independent variables to the variance in Recorded Violations is discussed according to the six categories shown in Table 1. The personal variables yield fully one-half the accounted variance in Recorded Violations. Age, the most important contributor, is correlated with Recorded Violations in the same direction as that noted in many other studies—younger drivers incur more violations. The second personal variable to be included in the violation prediction equation, Height, is correlated positively with Recorded Violations, but it is not clear why taller truck drivers should incur more violations.

Of the cognitive variables, Knowledge of Regulations is the only one that contributes appreciably to the variance in Recorded Violations. The direction is one that would be expected on a logical basis—the person with a lower Knowledge of Regulations score is likely to have more violations.

The only visual variable included in the final prediction equation was Right Side Vision, with a higher score being associated with more violations. Neither motor variable is included in the final prediction equation.

The medical variables are correlated negatively with Recorded Violations. This is as one would expect from his knowledge that blood pressure is correlated positively with age, which in turn is correlated negatively with violations. But it is surprising that Systolic Blood Pressure emerged as high as it did on predictive order. One might have expected that since Age, which emerged first, is correlated positively with Systolic

Blood Pressure (and to a lesser extent with Diastolic Blood Pressure), any variance that Systolic Blood Pressure shares with Recorded Violations would have been preempted by Age. Apparently Age alone does not account for the relation between Systolic Blood Pressure and Recorded Violations.

Finally, none of the personality variables was included in the violation prediction equation.

To ascertain the violation results given in Table 1, a similar table was prepared, but with the results based on the combined data of the predictor and predicted groups. As previously mentioned under the discussion on accidents, the five personality variables were not included in the analysis. The results from the remaining 22 independent variables were similar to those given previously for the predictor group alone. In a rough sense, violations are more likely to be associated with the younger, taller (or heavier) driver whose knowledge of regulations is low, whose right side vision (and possibly night vision) is better than average, and whose systolic blood pressure is lower than average.

SUMMARY

An attempt was made to predict whether each of 360 truck drivers had at least one accident recorded during the past three years with California's Department of Motor Vehicles. The prediction was based on 27 non-driving variables on which data could have been collected before licensing. The same prediction procedure was also used in an attempt to predict recorded violations for the same time period.

The prediction of recorded accidents, although not high in any absolute sense, was statistically significant. Thus, a justification existed for a discussion of the variables that contributed to that prediction. In a rough sense, accidents are more likely to be associated with the younger driver with glasses who is slow in braking.

The prediction of recorded violations was in the right direction, but it failed to reach a statistically significant level. Analysis seemed to indicate that this failure was due to the use of the same procedure for recorded accidents; had a procedure been used that would utilize the essentially continuous characteristic of recorded violations, the prediction of recorded violations would also have been significant. Thus, it was felt justified to discuss also the variables that contributed to that prediction. In a rough sense, violations are more likely to be associated with the younger, taller (or heavier) driver whose knowledge of driving regulations is low, whose side vision (and possibly night vision) is better than average, and whose systolic blood pressure is lower than average.

ACKNOWLEDGMENT

The authors are pleased to acknowledge the cooperation of Wade Sherrard, Managing Director of the California Trucking Associations Inc., and of Robert Hansen, Manager of the Associations' Driver Testing Center. All computations were performed at UCLA's Computing Facility.

REFERENCES

1. Levonian, E., and Case, H. W., "Behavioral and Demographic Correlates of Responses to a Driving Questionnaire." HRB Proc., 40:582-592 (1961).
2. "BIMD Computer Programs Manual." BIMD 17, Div. of Biostatistics, Dept. of Preventive Med. and Public Health, School of Med., UCLA (1961).
3. Kaiser, H. F., "The Varimax Criterion for Analytic Rotation in Factor Analysis." Psychometrika, 23:187-200 (1958).
4. "BIMD Computer Programs Manual." BIMD 29, Div. of Biostatistics, Dept. of Preventive Med. and Public Health, School of Med., UCLA (1961).
5. Guilford, J. P., and Michael, W. B., "The Prediction of Categories from Measurements." Sheridan Supply Co., Bevery Hills, Calif. (1949).
6. Guilford, J. P., "Fundamental Statistics in Psychology and Education." McGraw-Hill (1956).
7. Kendall, M. G., "The Advanced Theory of Statistics." 3rd ed., Vol. 1, Griffin, London (1951).