DOCUMENT RESUME

ED 044 441 TM 000 173

AUTHOR	Carroll, John B.
TITLE	Note on the Scoring of Foreign Language Speaking and Writing Fluency Tests.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.
REPORT NO	RB-70-52
PUB DATE	Sep 70
NOTE	25p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.35
DESCRIPTORS	Correlation, *French, *Language Fluency, Language Proficiency, Language Skills, *Language Tests, Performance Tests, Predictor Variables, *Scoring, *Second Languages, Test Reliability, Test Validity, Written Language
IDENTIFIERS	*England

ABSTRACT

The problem of determining relative weights for quantity and quality in scoring foreign language speaking and writing fluency tests is studied. French speaking and writing fluency tests were administered to students of French in several schools in England. Data from these tests was analyzed to support the suggestion that scoring formulas should reflect two components of performance: (1) quantity of correct response, and (2) relative quality of response. Two quantity and five quality variables were identified and correlated. Using a priori reasoning and the correlations, several scoring formulas were tried. The study and the cross-validation study indicate that nonlinear combinations of raw scores, probably ratios and products, may be needed. (IR)

RB-70-

NOTE ON THE SCORING OF FOREIGN LANGUAGE SPEAKING AND WRITING FLUENCY TESTS

John B. Carroll

Educational Testing Service
Princeton, New Jersey
September 1970

NOTE ON THE SCORING OF FOREIGN LANGUAGE SPEAKING AND WRITING FLUENCY TESTS

John B. Carroll

Educational Testing Service

ABSTRACT

Data from French speaking and writing fluency tests are analyzed to suppo the suggestion that scoring formulas should reflect two components of performa (1) quantity of correct response, and (2) relative quality of response. This may require nonlinear combinations of raw scores, usually, formulas involving ratios or products.

NOTE ON THE SCORING OF FOREIGN LANGUAGE SPEAKING AND WRITING FLUENCY TESTS[1]

John B. Carroll

Educational Testing Service

In the scoring of foreign language speaking and writing fluency tests, a perennial problem has been that of the relative weights to be given to quantity and quality of response. If quantity of response is small, the scores for quality tend to be unreliable; on the other hand, if quantity of response is large, the scorer is likely to be either overimpressed with it or negatively influenced by it, and scores based on the quantity of correct responses may consequently be either inflated or unfairly decreased.

An opportunity to study this problem was presented in connection with the author's work in developing a set of speaking and writing fluency tests in French as a foreign language for the International Study of Educational Attainment, familiarly known as I.E.A. (Husén, 1969).

The Tests

The speaking fluency test consists of pictures of situations which the respondent is asked to describe in French. There are two pictures in the test designed for a population of 10-year-old learners (called Population I in the I.E.A. study), and the child chooses one; in the test for older learners (14-year-olds and pre-university populations, i.e., Populations II and IV, respectively, as defined in the I.E.A. study) there are three pictures (not the same as those for Population I) from which the pupil has to choose two to respond to. This test is not timed; the child is simply told to describe the picture in French--"to say anything he likes about the picture."

In a preliminary scoring of the speaking fluency test responses, scores were assigned to each total response to each picture by a team of native French speakers, as follows:

X_1 = number of "propositions" (French for clauses) in the response

X_2 = number of different grammatical structures represented in the response

X_3 = number of propositions with correct structures

X_4 = number of propositions with correct morphology

X_5 = number of propositions with correct vocabulary

X_6 = number of propositions with correct pronunciation

X_7 = number of propositions exhibiting one or more hesitations

In addition, a global rating on a 5-point scale from 0 to 4 (high), here identified as Y, was assigned to the total response by this same team of scorers. For the Population I cases, this was assigned on the basis of the response to one picture; for the Population II and IV cases, it was based on the responses to two pictures. It will be noted that X_1 is a measure of sheer quantity; X_2, X_3, X_4, X_5, and X_6 are measures of both quantity and quality; X_7 is indicative of quantity but, presumably, negative quality. The problem posed by these data was to determine a suitable system for combining the X values into a single score that would well predict the global rating, which was regarded as a criterion score.

The writing tests were slightly different for Populations II and IV; there was no writing test for Population I. The Population II test directed the pupil to write, within 10 minutes, a six-exchange dialogue between two persons (Louis and Paul), including in the dialogue, in the order given, nine designated words or phrases (with any appropriate grammatical changes necessary). Each exchange was required to have at least three words, but could include more if necessary "to tell the story clearly." The Population IV test directed the pupil to write,

within 10 minutes, a short "free" composition comparing the merits of living in the country and in a big city. Certain "themes" were suggested, to be used in the order given (e.g., advantages of country life--peace and quiet, scenery, good food, health). For both Populations II and IV, the compositions were scored by native French speakers with respect to three 5-point scales (0 to 4): fluency or completeness (amount written), grammatical accuracy, and style. Again, the problem was how to combine these scores into a single index. For the compositions, however, there was no direct criterion.

Subjects

The tests were given to pupils in several schools in England where they were being taught French. For the speaking tests, there were 17 pupils in Population I, 13 in Population II, and 33 in Population IV. For the composition tests, data were available for 28 pupils in Population II and 180 pupils in Population IV.

Analysis of Results

Speaking test scores. The first step was to compute and examine the Pearsonian correlations among the raw, untransformed, uncombined scores for all three populations pooled (N = 63) for the responses to the first picture chosen. (In the case of Population I, these were the only data available.) The plan was to develop a scoring procedure for the first response and cross-validate it on the second response (available only for Populations II and IV pupils). The correlations thus obtained are shown in Table 1. Several initial conclusions were drawn from this table:

Insert Table 1 about here

---

(1) Sheer quantity of response (X_1) had little correlation with the global rating yet it would be a mistake to omit it from the scoring scheme since it showed

appreciable correlations with other variables; X_1 could be a suppressor variable. (2) The highest correlations with the global rating were yielded by variables X_4, X_3, X_5, and X_6 in that order; all these variables appeared to form a rather tight cluster. Variable X_2 also showed an appreciable correlation with the criterion but smaller correlations with variables X_3 through X_6. (3) Variable 7, the number of clauses with hesitations, showed a negligible correlation with the criterion; its high correlation with the number of clauses indicated that it was primarily another measure of quantity.

At this point the standard method of procedure would dictate computing a regression equation for the predictor variables. Before such a procedure was followed, however, it was decided to investigate methods of transforming or non-linearly combining the measures of quantity and quality. Variables X_4 and X_1 were selected for special study in view of the former's high correlation with the criterion. By making various three-dimensional scatterplots for transformations or combinations of these variables (the criterion variable being entered as numbers to represent the third dimension) it appeared that the best procedure for combining the variables would be to establish a new variable, $X_4' = X_4/X_1$, and then to compute the optimal weights for X_4' and X_4 for predicting the criterion. The resulting multiple correlation was .8088, with $\beta_{X_4} = .5271$ and $\beta_{X_4'} = .3260$. This multiple correlation was in fact slightly superior to $R_{Y.14} = .8035$, with $\beta_{X_1} = -.2148$ and $\beta_{X_4} = .8745$. It was decided, also, that this way of combining variables made psychological sense, in that it represented a postulated process whereby the scorer takes into account not the sheer quantity of response but, rather, two perceptible aspects of the response: (1) the quantity of correct response, and (2) the proportion of the total response that is correct. Such a judgmental process seems intuitively more reasonable than one whereby the scorer takes into account the quantity of correct response and then "subtracts"

points for the quantity of total response. If, for example, a respondent produced a large quantity of response that was all correct, there would be no reason (and it would be unfair) to penalize him for producing a lengthy response--a procedure that would be implied by the straightforward linear combination of the raw scores, with its negative beta-weight for X_1 .

This matter was later checked by comparing the multiple correlations and beta weights for the two procedures as applied to all variables X_2 through X_6 . The results are shown in Table 2. It will be there observed that, actually, the non-linear combination procedure produces higher multiple correlations for only two

---
Insert Table 2 about here
---

of the variables. Nevertheless, the a priori line of reasoning developed above suggests that the nonlinear combination procedure makes for more sensible and fairer results. It was concluded that the final scoring formula should be based on the nonlinear combination procedure.

It was desired that the final scoring formula be as simple as possible to apply. It was decided, therefore, to determine optimal weights for two summational variables:

$$X_8 = X_2 + X_3 + X_4 + X_5 + X_6 ;$$

$$X_9 = (X_2 + X_3 + X_4 + X_5 + X_6)/ X_1 .$$

The results are shown in Table 3. Of interest is the fact that the correlation

---
Insert Table 3 about here
---

between X_8 and X_9 is far from unity, also the fact that the beta-weights for the two variables are approximately equal, indicating that they make approximately

equal independent contributions to the prediction. The multiple correlation with the criterion is very appreciably higher than any of the zero-order criterion correlations in Table 1, and also higher than any of the multiple correlations shown in Table 2. In order to simplify the scoring formula still further, it was noted that the ratio of the b-weights was approximately 10. Therefore, the final scoring formula was defined as follows:

$$X_{10} = X_8 + 10X_9 .$$

The correlation of X_{10} with Y is very nearly .8537.

The scoring formula represented by X_{10}, developed on the basis of the data available for the first speaking test, was "cross-validated" by applying it to the data for the second speaking test. It will be recalled that data were available for the second speaking test only for the 46 cases in Populations II and IV, a subset of the cases used in developing the scoring formula. Strictly speaking, this was not cross-validation in the usual sense of applying a formula to a completely different set of cases. The "cross-validation" was in truth a matter of applying a scoring formula to a different set of data (an "alternate form" of the test, so to speak) from the same set of cases, or actually a subset. For the 46 cases in Populations II and IV, correlations were obtained among variables X_8, X_9, and X_{10} for both the first and second speaking tests, as well as the correlations of these variables with the global rating. The results are shown in Table 4. The scoring formula produced a validity coefficient of

---

Insert Table 4 about here

.80 in the case of the first speaking test (a figure analogous to the value of .85 yielded for the complete set of 63 cases), but the validity shrank to .67 when

the scoring formula was applied to the second test. In this same sample, the correlation between the final scores of the first and second speaking test was .73, a value that indicates the reliability of the scoring formula. (By the Spearman-Brown formula, the reliability of scores combined from both tests for Populations II and IV would be estimated as .84.)

It was of interest to investigate the reasons for the shrinkage in the validity of the scoring formula. Correlations were computed among the raw variables of the second test for the restricted sample, as well as with the global ratings, with results shown in Table 5. It is evident from this table that the structure

---

Insert Table 5 about here

---

of the variables in the restricted sample is somewhat different from that observed in Table 1. Quantity of response (X_1) is much more highly correlated with the remainder of the predictor variables, as well as with the criterion variable. Even the presumably negatively oriented variable X_7 (number of clauses with hesitations) has an appreciable positive correlation with the criterion. If we had begun our investigation with the data of Table 5, it is possible that we would not have come up with the conclusion that we arrived at from the data of the first test. On the other hand, the second speaking test did not yield the high correlations of variables X_3 and X_4 that were observed with the first speaking test. In view of the larger and more varied sample that was available for arriving at the scoring formula, as well as the intuitively persuasive rationale for this formula, it was decided to accept it despite the appreciable shrinkage that occurred for the data of the second speaking test.

Another feature of the data that makes the interpretation of the "cross-validation" difficult is the fact that if we compare the means and standard deviations shown in Tables 1 and 5 for the seven raw scores on the first and second

speaking tests, the means for the second test (N = 46) are not in every case higher than the means on the first test for the complete sample (N = 63), as we might expect them to be in view of the fact that Populations II and IV are more advanced than Population I cases. Furthermore, the standard deviations for the second-test scores are in most cases larger than those for the first-test scores. These features may be artifacts of the data, due partly to the fact that the stimuli for the speaking tests were different between populations, or to possible practice effects occurring from the first to the second test.

It is interesting to notice in Table 5, for the cross-validation data, that for both the first and the second speaking tests the use of the ratio variable X_9 produces an increment in the validity of the final scoring formula, X_{10}, over the "number right" variable X_8.

Writing test scores. As noted previously, there was no appropriate criterion for evaluating the writing test scores. On the assumption that the Population IV responses should be on the average better than the Population II responses, a nominal criterion, here called Y, was assigned such that the Population II cases had $Y = 2$ and Population IV had $Y = 4$. It was recognized that the tests for the two populations differed in important respects, and any differences between the populations revealed by the tests would be attenuated by the fact that each test had been geared to a specific range of competence. Also, we must recognize that the number of cases in Population II was only 28 as compared to the 180 cases in Population IV. Nevertheless, in the absence of any better criterion, it was felt that statistical operations based on optimal weightings of scores to differentiate the samples from the two populations would suggest a scoring formula that would have some likelihood of holding up against a superior

criterion. (It is contemplated that better and more complete data will become available at a later time.)

It was decided to explore the possible generality of the rationale developed for the speaking test formula. Recall that there were three scores assigned for the writing test, all on a scale from 0 to 4: X_1, a measure of the "length" or completeness of the response; X_2, an assessment of the relative grammatical accuracy of the response; and X_3, an assessment of the quality of "style" of the response. The rationale developed for the speaking test scoring suggested that quantity of correct response and relative correctness of the total response should be the two factors considered in a scoring formula. Applying this rationa to the writing test scores, we would conclude that X_2 and possibly also X_3 are measures of relative correctness as they stand. To obtain measures of the quantity of correct response, however, we should use some function of the product of X_1 times X_2 and/or X_3. To gain insight into the relationship among the raw scores and such functions, a matrix of correlations was computed among the raw variables, several functions of them, and the nominal criterion Y. The functions of the raw scores investigated were: X_1X_2, X_1X_3, $X_2 + X_3$, and $X_1(X_2 + X_3)$. The correlation matrix is shown in Table 6. Also, multiple

---

Insert Table 6 about here

---

regression systems were computed for several combinations of the variables, as shown in Table 7. From the results in Table 6, it will be immediately noticed

---

Insert Table 7 about here

---

that none of the variables correlates highly with the nominal criterion; only cor relations equal or greater than .0895 are significantly positive at the 5% level, or .1434 at the 1% level (by a one-tailed test, considered legitimate here becaus

we should expect the correlations to be positive). We have already mentioned the limitations of the data that are likely to have resulted in such low correlations Nevertheless, the correlations of X_2 and X_1X_2 with the criterion are significan at the 1% level.

In Table 7, the computations for Combinations 1, 2, 3, and 4 permit us to examine whether the nonlinear combinations are superior to the linear combination In the case of X_1 and X_2, the nonlinear combination is slightly superior, supporting the rationale for such a combination. This is not the case for variables X_1 and X_3, however; in fact, variable X_3 receives a negative weight in the nonlinear combination. We take as a principle the proposition that a score should not receive a negative weight in a scoring formula. It is noteworthy, however, that X_1X_3 receives a positive weight, and in fact its zero-order correlation has a higher correlation with the criterion than does X_3 in its origin form. This suggests that the assessment of style should enter the scoring formul in the form X_1X_3. Adding this fact to the fact that the multiple regression for X_1 and X_1X_2 yields approximately equal beta-weights for these variables, we conclude that the final scoring formula should possibly be a linear function of X_2, X_1X_2, and X_1X_3.

First, however, let us examine the multiple regressions for Combinations 5 and 6; these are, respectively, for the linear combination of X_1, X_2, and X_3, and for the variables X_2, X_3, X_1X_2, and X_1X_3. The nonlinear combination of variables yields a slightly higher multiple R than does the linear combination. However, because of the negative weights for variables X_3 and X_1X_2 in Combination 6 it is not reasonable to use it as a basis for a scoring formula.

Combination 7 shows the multiple regression system for variables X_2, X_1X_2, and X_1X_3. Unfortunately, variable X_1X_3 again receives a negative weight of appreciable size. Although the multiple correlation is still nearly as high as

would be obtained from Combinations 5 or 6, we must reject this multiple regressior system as a basis for a scoring formula.

At this point we might decide to eliminate variable X_3 completely from the scoring formula, but this seems a possibly unfortunate thing to do because it loses information. It is possible that the negative weight of X_3 arises from some sort of sampling error. Under the circumstances, it seems advisable to include X_3 in some fashion. Considering that variable X_1X_3 has been shown to have a reasonably "high" correlation with the criterion, we decide to combine it with X_1X_2 and make the scoring formula a linear combination of X_2 and $X_1(X_2 + X$ The multiple regression for such a combination is shown in Table 7 under the headi Combination 8. Now the variable $X_1(X_2 + X_3)$ receives a relatively small weight, but at least it remains positive. The ratio of the b-weight for the first of these variables to the second is about 36. As a quite arbitrary matter, let us prescribe the final scoring formula as, for the sake of simplicity,

$$\text{Score} = 10X_2 + X_1(X_2 + X_3) .$$

The coefficient of 10 is used for X_2 rather than 36 in order to give relatively more weight to the second term in the formula than would be assigned by the multi regression weights. Whereas the multiple correlation of the variables with the criterion is .1517, the scoring formula with the coefficient of 10 yields a correlation nearly as high, namely .1492. The standard deviations of the two terms in the formula are 8.0250 and 5.5352, respectively. Figure 1 depicts the final scores that will be obtained for various combinations of scores on X_1 , X_2 , and X_3 . It can be seen that the score on X_2 is the principal determiner of

---

Insert Figure 1 about here

---

the score, but the pupil gets additional points for quantity of correct response in terms of grammar and style: the increment of score he gets depends upon X_2, his rating for grammatical correctness. Obviously, he cannot get a score of other than $X_2 = 0$ if he does not produce any response; for this reason, the scores shown for $X_1 = 0$ on the chart are spurious. Also, it happens that because of the correlations among the scores, a number of score combinations are extremely unlikely to occur. The small figures shown on the chart are the actual frequencies of the score combinations in the data employed for this analysis, and the distributions of scores in the two populations are shown at the right of the chart.

REFERENCE

Husén, T. International impact of evaluation. In Tyler, Ralph W. (Ed.), Educational evaluation: New roles, new means. Chicago: University of Chicago Press, 1969. (68th Yearbook of the National Society for the Study of Education, Part II.) Pp. 335-350.

FOOTNOTE

[1]I am grateful to John L. D. Clark and W. B. Schrader, both at Educational Testing Service, for critical comments on an early draft of this paper.

Table 1

Pearsonian Correlations, Raw Scores and Global Rating,

First Speaking Test, All Pupils in Populations

I, II, and IV (N = 63)

Variable		1	2	3	4	5	6	7	Y
No. clauses	1	1.00	.10	.48	.44	.64	.62	.91	.17
No. different structures	2	.10	1.00	.55	.59	.45	.38	-.03	.61
No. clauses w/correct structure	3	.48	.55	1.00	.85	.86	.71	.25	.74
No. clauses w/correct morphology	4	.44	.59	.85	1.00	.82	.72	.24	.78
No. clauses w/correct vocabulary	5	.64	.45	.86	.82	1.00	.77	.42	.66
No. clauses w/correct pronunciation	6	.62	.38	.71	.72	.77	1.00	.44	.63
No. clauses w/hesitations	7	.91	-.03	.25	.24	.42	.44	1.00	-.02
Global rating	Y	.17	.61	.74	.78	.66	.63	-.02	1.00
	Mean	7.38	2.57	4.48	3.49	4.22	2.76	5.27	1.29
	S.D.	4.26	1.46	2.84	2.62	2.86	2.89	3.80	1.09

Table 2

Comparison of Linear and Nonlinear Scoring Procedures, First Speaking Test (N = 63)

Variable	Linear Combination			Nonlinear Combination		
	β_{x_1}	β_{x_i}	$R_{Y \cdot 1i}$	β_{x_i}	$\beta\left(\frac{x_i}{x_1}\right)$	R
X_2 = No. different structures	.1101	.5990	.6197	.7482	-.2226	.6315
X_3 = No. clauses w/correct structure	-.2406	.8550	.7693	.5629	.2473	.7567
X_4 = No. clauses w/correct morphology	-.2148	.8745	.8035	.5271	.3260	.8088
X_5 = No. clauses w/correct vocabulary	-.4275	.9336	.7372	.4148	.3613	.7067
X_6 = No. clauses w/correct pronunciation	-.3583	.8522	.6899	.1931	.5035	.6766

Table 3

Correlations and Regression Analysis for Components of Final Scoring Formula, First Speaking Test (N = 63)

	Correlations			Mean	S.D.	β	b
	X_8	X_9	Y				
X_8	1.00	.69	.78	17.52	11.14	.4561	.0446
X_9	.69	1.00	.79	2.42	1.11	.4521	.4644
Y	.78	.79	1.00	1.29	1.09	R =	.8537

Table 4

Correlations Across First and Second Speaking Tests, Pupils in Population II and IV (N = 46)

		1st Test			2nd Test			
		X_8	$10X_9$	X_{10}	X_8	$10X_9$	X_{10}	Y
1st Test	X_8	1.00	.67	.95	.70	.50	.69	.76
	$10X_9$	.67	1.00	.87	.62	.48	.63	.70
	X_{10}	.95	.87	1.00	.73	.54	.73	.80
2nd Test	X_8	.70	.62	.73	1.00	.60	.95	.65
	$10X_9$	.50	.48	.54	.60	1.00	.81	.52
	X_{10}	.69	.63	.73	.95	.81	1.00	.67
	Y	.76	.70	.80	.65	.52	.67	1.00
	Mean	20.46	29.75	50.21	22.20	30.12	52.32	1.67
	S.D.	10.52	6.85	15.95	15.71	8.20	21.66	1.00

Table 5

Correlations Among Original Variables, Second Speaking Test, Pupils in Population II and IV (N = 46)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
X_1	1.00	.86	.94	.89	.92	.83	.89	.53
X_2	.86	1.00	.79	.73	.83	.74	.74	.50
X_3	.94	.79	1.00	.92	.92	.85	.84	.57
X_4	.89	.73	.92	1.00	.91	.85	.77	.63
X_5	.92	.83	.92	.91	1.00	.87	.79	.68
X_6	.83	.74	.85	.85	.87	1.00	.70	.63
X_7	.89	.74	.84	.77	.79	.70	1.00	.48
Y	.53	.50	.57	.63	.68	.63	.48	1.00
Mean	6.91	3.30	5.61	4.46	5.04	3.78	4.72	1.67
S.D.	4.03	2.16	3.85	3.31	3.46	3.93	2.58	1.00

Table 6

Intercorrelations of Selected Functions of Writing Test Scores and the Nominal Criterion (Y)

(N = 208*)

		1	2	3	4	5	6	7	8
X_1	1	1.0000	.4409	.3980	.7614	.7395	.4689	.7974	.0921
X_2	2	.4409	1.0000	.6038	.8260	.5495	.9006	.7337	.1509
X_3	3	.3980	.6038	1.0000	.5307	.8152	.8903	.7117	.0424
X_1X_2	4	.7614	.8260	.5307	1.0000	.7722	.7613	.9438	.1474
X_1X_3	5	.7395	.5495	.8152	.7722	1.0000	.7585	.9388	.0816
$X_2 + X_3$	6	.4689	.9006	.8903	.7613	.7585	1.0000	.8073	.1093
$X_1(X_2 + X_3)$	7	.7974	.7337	.7117	.9438	.9388	.8073	1.0000	.1223
Y	8	.0921	.1509	.0424	.1474	.0816	.1093	.1223	1.0000
	Mean	2.2644	1.4856	1.4038	3.7644	3.5240	2.8894	7.2885	3.7308
	S.D.	1.1318	.8025	.7661	3.0012	2.8789	1.4048	5.5352	.6826

* $r_{.05} = .0895$, $r_{.01} = .1434$ (one-tailed test).

Table 7

Multiple Regression Systems for Several Combinations of Scores on French Writing Test

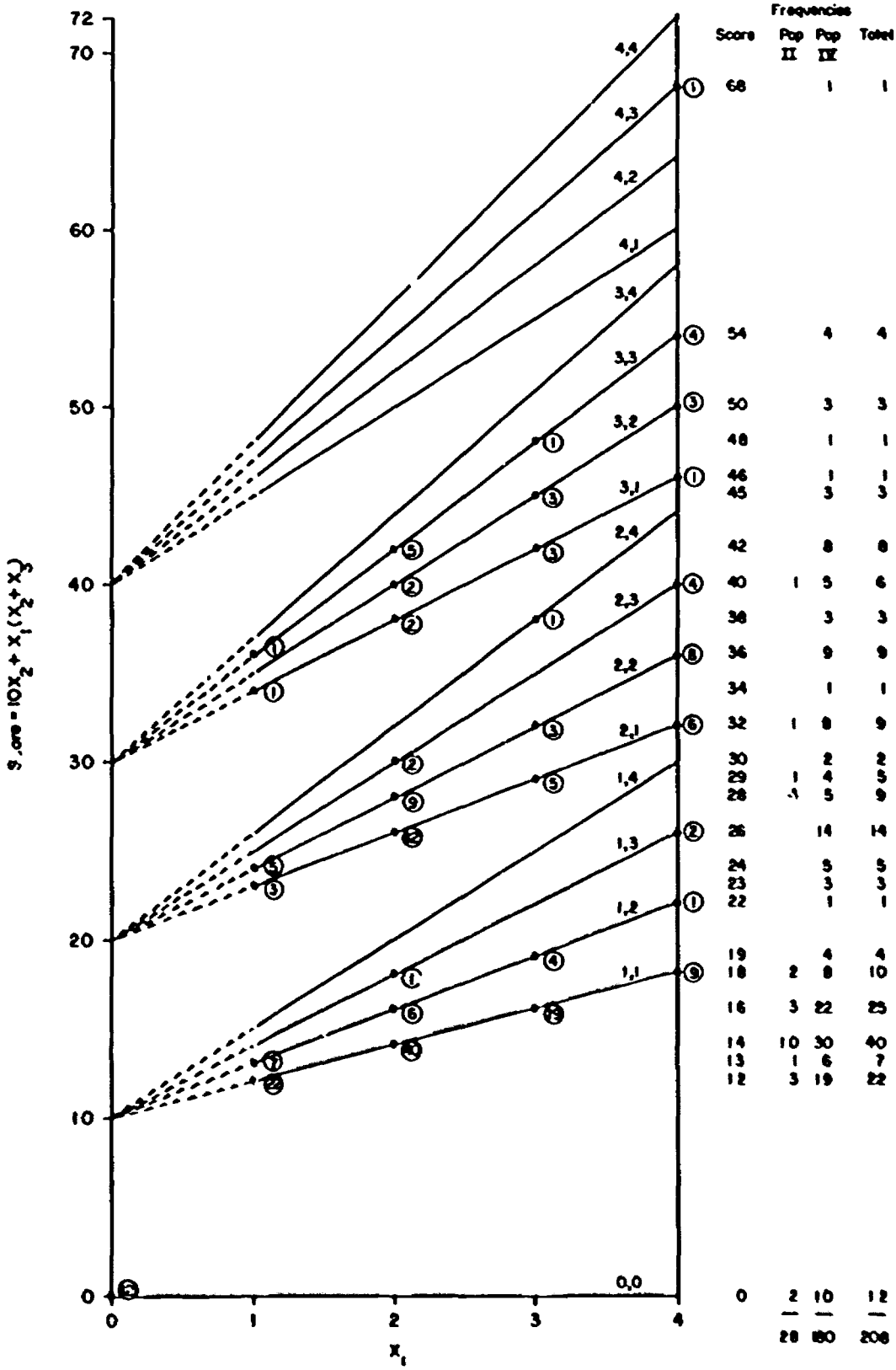
	Combination 1		2		3		4	
	β	b	β	b	β	b	β	b
X_1	.0318	.0191			.0894	.0539		
X_2	.1369	.1165	.0916	.0779				
X_3					.0068	.0061	-.0719	-.0641
X_1X_2			.0718	.0163				
X_1X_3							.1401	.0332
R	.1536		.1526		.0855		.0757	

	5		6		7		8	
	β	b	β	b	β	b	β	b
X_1	.0459	.0277						
X_2	.1826	.1554	.2735	.2327	.0740	.0629	.1323	.1126
X_3	-.0861	-.0767	-.2343	-.2088				
X_1X_2			-.1204	-.0274	.1353	.0308		
X_1X_3			.2153	.0510	-.0635	-.0150		
$X_1(X_2 + X_3)$							.0251	.0031
R	.1677		.1765		.1610		.1517	

Figure Caption

Figure 1. Nomograph for final scores, Score = 10 X_2 + $X_1(X_2 + X_3)$, on French writing test, where X_1 = fluency or completeness, X_2 = grammatical accuracy, and X_3 = style. Each line is labeled with an ordered pair of scores on (X_2, X_3). Numbers in small circles are frequencies of scores at the given points. At the right are found the frequency distributions of final scores for cases in Population II, Population IV, and the total.

Figure 1


Score = 10X2 + X1(X2+X3)
X1
4,4
4,3
4,2
4,1
3,4
3,3
3,2
3,1
2,4
2,3
2,2
2,1
1,4
1,3
1,2
1,1
0,0
Frequencies
Score
Pop II
Pop III
Total
68 1 1
54 4 4
50 3 3
48 1 1
46 1 1
45 3 3
42 8 8
40 1 5 6
38 3 3
36 9 9
34 1 1
32 1 8 9
30 2 2
29 1 4 5
28 4 5 9
26 14 14
24 5 5
23 3 3
22 1 1
19 4 4
18 2 8 10
16 3 22 25
14 10 30 40
13 1 6 7
12 3 19 22
0 2 10 12
28 180 208