DOCUMENT RESUME

ED 106 368 TM 004 486

AUTHOR Convey, John J.
TITLE A Validation of Three Models for Producing School Effectiveness Indices.
PUB DATE [Apr 75]
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Comparative Analysis; Correlation; *Models; Predictor Variables; *Program Effectiveness; Sampling; *Schools; Simulation; Socioeconomic Status; *Statistical Analysis; Test Results; Tests of Significance; *Validity

ABSTRACT

The capability was studied of each of three models for producing indices that will reproduce school effectiveness rankings established a priori through simulation. The models used were a within-group regression technique, a regression model using individual scores, and a regression model using means. Data for 54 hypothetical schools on input, SES, school, and output variables were randomly generated from a multivariate normal distribution using parameters from previous studies. The results indicated that each type of model was capable of producing indices which were rather accurate reflections of the effectiveness ranks of schools. (Author)

ED106368

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Validation of Three Models for Producing School Effectiveness Indices[1]

John J. Convey

Catholic University of America

An aspect of the school effectiveness question which has received some attention recently is the determination of the relative effectiveness of schools from indices produced by statistical models. Marco (1974) examined five such models which are based on longitudinal data (see also, Convey, 1973). Typically, the models have been used with available data and conclusions have been made as to the relative effectiveness of the schools involved without the presentation of any evidence of the validity of the models (Burke, 1973; Dyer, Linn, & Patton, 1969; Forsyth, 1973; Marco, 1974).

The purpose of this study was to examine the capabilities of three types of statistical models to produce indices that would reproduce school effectiveness rankings established a priori. Simulated data were used in the study so that the parameters could be manipulated in order to make some schools more effective than others according to an established criterion. The use of simulated data then avoided the problem of relying on experts or consensus opinion to determine the relative ranking of the schools.

---

[1]This paper is based on the author's Ph.D. dissertation submitted to the faculty of the Florida State University.

A paper presented at the annual meeting of the American Educational Research Association, Washington, March 30-April 3, 1975.

TM 004 486

Method

Variables

The variables selected for the study were similar to those used in previous school effectiveness studies and were classified as output, input, socioeconomic (SES), and school variable The output and input variables were given the characteristics of the total math score on the Comprehensive Tests of Basic Skills (CTBS) for eighth graders (Level 3, Form R) and sixth graders (Level 3, Form Q), respectively. The expanded standard score scale developed for all levels and forms of the CTBS was used for the output and input scores (California Test Bureau/McGraw-Hill, 1970). The SES variable was given the characteristics of the index used in the Talent Project as reported by Cooley and Lohnes (1971). The characteristics of the scores on the Verbal Ability Test for Teachers, as used in the Coleman Report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966), were given to the measure of the school variable.

Models

Three types of models were used to produce indices:

1. Within-School Regression. For each school, a prediction equation was obtained from the regression of the individual student output scores on individual student input and SES scores. The equation is:

$$O' = a + b_1 I + b_2 SES \qquad (1)$$

where, O' is the predicted output for an individual; b_1 and b_2 are the least squares estimates of the coefficients of input (I) and SES within each school; and a is the least squares estimate of the constant for the school. This model produces a unique

regression plane for each school. In general, these planes will not be parallel. Hence this model allows schools to be tested for differential effectiveness at various values of I and SES. An effectiveness index is defined for each school at a specific combination of predictor values (I_0,SES_0) as the predicted value at that point. Normally, several such ordered pairs will be of interest. In addition to the two-predictor model, a model using input only as predictor was examined.

2. Individual Regression Residuals. For the total group, a prediction equation was obtained from the regression of the individual output scores on the predictors I, SES, and SV(school variable which was assumed constant for all individuals in a given school). The equation is:

$$O' = p + q_1 I + q_2 SES + q_3 SV \qquad (2)$$

where, O' is the predicted output for an individual; q_1,q_2, q_3 are the least squares estimates of the coefficients of the predictors based on the total group; and p is the least squares estimate of the constant based on the total group. In addition to the above three-predictor model (IRR), a two-predictor model (IR2) using I and SES, and a one-predictor model (IR1) using I were examined. In each model, the residuals for individuals were obtained. The effectiveness index for each school was calculated by averaging the residuals within each school.

3. School Regression Residuals. For the total group, a prediction equation was obtained from the regression of the mean output $\overline{O}$ for each school on the mean of each predictor, $\overline{I}$, $\overline{SES}$, and $\overline{SV}$,

for each school. The equation is:

$$\bar{O}' = r + s_1 \bar{I} + s_2 \overline{SES} + s_3 \overline{SV} \qquad (3)$$

where, $\bar{O}'$ is the predicted mean output for the school; s_1, s_2, and s_3 are the least squares estimates of the coefficients of the predictors when means are used; and r is the least squares estimate of the constant based on the means for each school. In addition to the above three-predictor model (SRR), a two-predictor model (SR2) using $\bar{I}$ and $\overline{SES}$, and a one-predictor model (SR1) using $\bar{I}$ were examined. In each model, the effectiveness index for each school is the residual obtained by subtracting the observed mean output $\bar{O}$ from the predicted mean output $\bar{O}'$

Sample

Data for 54 hypothetical schools were randomly generated for the four variables on the CDC 6500 computer at Florida State University using subroutine MSCORE. MSCORE generates random data from a multivariate normal distribution. The user specifies the means and standard deviations of, and the intercorrelations among, the variables.

Insert Table 1 about here

First, 54 ordered sets representing scores for output, input, SES, and school variable were generated according to the specifications shown in Table 1. These specifications were consistent with results of previous studies. The 54 scores for each variable were assumed to constitute the population means for the variables in each group. There is considerable empirical evidence to

indicate that, for achievement tests, the standard deviation of the distribution of school means is from .3 to .6 of the standard deviation of scores in the total population, regardless of group size (Lindquist, 1930, 1966; Lord, 1959). Lord (1959) suggested that .4 would be a good approximation. This approximation was used in the specifications for the generation of the input, output, and SES variables. The individual score standard deviation was used in generating the 54 school variable scores, since this variable was assumed to be constant within each group and maximum variability across groups was desired.

Second, the 54 groups were ordered from high to low on the input score. The top 18 were designated as high, the next 18 as medium, and the lowest 18 as low. Within each category, six groups were randomly designated as effective, six as average, and six as less effective. Effectiveness was defined in terms of the gain from the input mean to the output mean on the standard score scale. The input and output means were paired so as to satisfy the effectiveness criteria of effective (gain greater than 68), average (gain between 46 and 68), and less effective (gain less than 46). In this study, 68 units represented approximately one standard deviation on the input distribution. These criteria then appear to be consistent with previous studies (see, Coleman et al., 1966; Guthrie, 1970; Shaycoft, 1967). Table 2 shows the characteristics of the ordered sets after the pairing. These values were considered to be reasonably close to the generating values in Table 1.

Insert Table 2 about here

The input-output pairings were made so that within each effectiveness classification, groups with high, medium, and low inputs were equally represented. This procedure attempted to control for any bias that might be introduced by an overbalance of certain levels of inputs in any one classification. For example, it is well known that gain scores usually have a negative bias with respect to the initial scores (O'Connor, 1972). Thus, low inputs would tend to show larger gains, and the effective category might have contained a disproportionate number of low inputs had not the above procedure been adopted.

Insert Table 3 about here

Prior to the generation of individual scores within each school, group size was varied according to the plan shown in Table 3. This distribution is consistent with field results (Florida Ninth-Grade Testing Program, 1968). A table of random numbers was used to implement this plan. Group size was uniformly distributed over the different effectiveness classifications. The total group consisted of 9087 individuals and group size ranged from 20 to 399. The resulting distribution is given in Table 4.

Insert Table 4 about here

Next, individual student scores were randomly generated within each group using MSCORE with the intercorrelations shown in Table 1, the means in Table 2, and the standard deviations of 85.90 for output, 68.30 for input, and 9.42 for SES as parameters.

The score for the school variable assigned to each individual in each group was the mean for their respective group.

After each group was formed, the sample mean for each group was calculated. Some reranking of the groups occurred as a result of these sample values. It should be noted that these sample values would be directly observable when examining real data. The a priori data and the a priori ranks would not be observable.

Characteristics of Generated Data

The general characteristics of the generated data are given in Table 5. These values are reasonably close to the desired parameters. The negative correlations involving the school variable caused some concern. Because of the large number of cases, each of these correlations calculated from individual scores would be statistically different from the desired values at any reasonable level of significance. The school variable being constant within each group could have been a contributing factor to the negative correlations. However, it was concluded that the discrepancies noted were slight enough to continue with the applications of the models to the generated data.

Insert Table 5 about here

Finally, a characteristic of the generated data which may have some influence on the results should be noted. The range of correlations between input and output was .654 to .808, between input and SES, .209 to .589, and between output and SES, .263 to .621. These values appear to be within the bounds of sampling

variation. However, it may be that in some actual settings, correlations outside of this range may systematically occur for some groups. This would influence the behavior of the regression equation for each group and may effect the rankings produced by the Within-School Regression models for certain choices of predictor values. This limitation should be kept in mind when examining the results.

Procedures

The models were applied to the 54 groups and effectiveness indices were calculated. For the two-predictor Within-School Regression model, the following combinations of predictor values were used:

1) input mean (474) and SES mean (98.54);
2) 1 σ above the input mean (542.3) and 1 σ above the SES mean (107.96);
3) 1 σ below the input mean (405.7) and 1 σ below the SES mean (89.12).

Each application was designated at WSR1, WSR2, WSR3, respectively. Effectiveness indices were also calculated using the one-predictor Within-School Regression model with values at the mean, and one standard deviation above and below the mean, of the input. These models were designated as WR1, WR2, and WR3, respectively.

The ability of the models to produce accurate rankings was examined by obtaining correlations between the a priori ranks and the ranked indices produced by each model. In addition, selected pairwise comparisons between the correlations

associated with each model and the a priori ranks were obtained and tested for significance in order to determine if any one model was superior to the others in reproducing the a priori ranks. Simultaneous inference procedures were employed to prevent the compounding of a Type I error beyond a specified value.

Next, the ability of the models to discriminate between the 18 effective groups and the 18 less effective groups was examined. For the Within-School Regression models, one-tailed Bonferroni confidence intervals were constructed on the prediction surface representing each of the 18 effective groups and 18 less effective groups (see Miller, 1966). Each interval was constructed at $\alpha = .001$, thus maintaining an overall experimental α less than .324 for each model. The length of each interval was given by:

$$t_{\alpha,\nu} \, s \, [\underset{\sim}{a}^t \, (\underset{\sim}{x}^t\underset{\sim}{x})^{-1} \, \underset{\sim}{a}]^{1/2} \qquad (4)$$

where, $t_{\alpha,\nu}$ is the one-tailed Student t value; s is the estimated standard error for each surface; $\underset{\sim}{a}$ is $(1, I_o, SES_o)$ for the two-predictor models and $(1, I_o)$ for the one-predictor models; and $(\underset{\sim}{x}^t\underset{\sim}{x})^{-1}$ is the value from the appropriate normal equations. For each of the six Within-School Regression models, each effective group was compared with each less effective group. A comparison was declared different if the respective confidence intervals did not overlap.

The ability of the residuals models to discriminate was investigated using "Performance Indices" (PIs) as suggested by Dyer, Linn, and Patton (1967). For each model, the following ratio was calculated for each group:

$$R = \frac{\text{residual}}{\frac{\overline{SD}}{(\overline{n})^{1/2}}} \qquad (5)$$

where, $\overline{SD}$ is the average within-group standard deviation on the output for all groups, and $\overline{n}$ is the average group size. PIs were then calculated by the following rule:

$$\begin{aligned} R &< -1.5, && PI = 1; \\ -1.5 \leq R &< -.5, && PI = 2; \\ -.5 \leq R &< .5, && PI = 3; \\ .5 \leq R &\leq 1.5, && PI = 4; \\ 1.5 &< R, && PI = 5. \end{aligned} \qquad (6)$$

Different decision rules to determine when two PIs are different were used. Forsyth (1973) suggested that the criterion be at least 2 units. In addition, rules requiring at least 3 units and at least 4 units were examined.

Results and Discussion

The ranks of the effectiveness indices produced by each model are shown in Table 6 along with the a priori and the sample ranks. The effects of sampling can be seen by comparing the first two columns of Table 6. For the most part, the sample ranks were reasonably close to the a priori ranks. Only 11 of the 54 groups showed a discrepancy of more than 5 ranks and, of these, only groups 13 and 32 have discrepancies of more than 10 units. These discrepancies influenced the behavior of the models since the models will reproduce the sample ranks more accurately than the a priori ranks.

Insert Table 6 about here

Examination of Table 6 revealed that the ranks assigned by each model were rather consistent for most groups. A striking consistency occurred among the top 10 groups and the last 14 or 15 groups. This seemed to indicate that each model would be rather accurate for at least gross discriminations of effective from less effective groups.

Reproduction of A Priori Ranks

The intercorrelations among the a priori ranks, the sample ranks, and each of the model ranks are shown in Table 7. Each correlation is substantial and significant at $\alpha = .001$. Since 91 hypotheses were being considered simultaneously, the use of a Bonferroni strategy guaranteed that the overall α was not greater than .091.

Insert Table 7 about here

The correlations of the model ranks with the a priori ranks were highest for the one-predictor application of each model type. When SES was added as a predictor, each of the correlations decreased. Very little change occurred when the SV predictor was added. This same trend was present in the correlations of the model ranks with the sample ranks.

The phenomenon of decreasing correlations between the model ranks and the a priori ranks was initially unexpected. Perhaps this phenomenon was due to the increased influence of random error on the ranks as the number of predictors is increased. When a predictor is added, there is less error variation in the system. However, a larger proportion of that variation may be

due to random error. When the indices were ranked and correlated with the a priori ranks, these correlations may have diminished as the number of predictors increased because of this increased role of random error. A similar phenomenon was also illustrated in data reported by Castright (1974). Gastright attributed this to over-fitting the model to the data. Perhaps, but the increased influence of random error on the residuals as the number of predictors increased seems more plausible.

Insert Table 8 about here

The results of the significance tests of selected differences between each of the correlations of the model ranks with the a priori ranks are shown in Table 8. The most drastic reduction in the ability to reproduce the a priori ranks occurred in the School Regression Residuals model. The differences between the correlations with the one-predictor SR1 ranks were significant at $\alpha = .005$. A less stringent individual comparison level of .05 would have resulted in most of the other one-predictor models being declared different from their two- and three-predictor counterparts.

The above evidence seems to indicate that basically different results are possible when models are used with varying number of predictors. If the groups are to be ranked on a given criterion which is influenced by all variables, the observed mean gain seems to be the best measure of effectiveness. The correlation between the sample ranks and the a priori ranks was .9587 in this study. If a ranking is desired where input is controlled, a model using

input as a predictor would seem to be appropriate. The similarity of the rankings from these one-predictor models to the a priori ranks will depend upon the relationship between the criterion predicted by input and the criterion used to establish the a priori ranks. If groups are to be ranked controlling for the influence of SES and input, then a model using both of these as predictors should be employed. Thus, it appears that the issue being illustrated here is one of proper model specification.

Once the number of predictors is decided upon, the question still remains as to which type of model is superior. The correlations involving the different types of one-predictor models were very similar. These results are consistent with previous research using nonhypothetical data (see, Dyer et al., 1969; Marco, 1974). The results among the two-predictor models were somewhat less consistent. For example, the correlation of the SR2 ranks with the a priori ranks was significantly different from the correlation of the IR2 model ranks with the a priori ranks. However, the intercorrelations among the ranks of these two-predictor models ranged from .9685 to .9855. These latter correlations are probably more representative of the agreement among the models than are the correlations of the model ranks with the a priori ranks, since they are comparing ranks based on similar criteria. Thus, the significant difference noted may have resulted from the high relationship between the models and may not indicate any superiority on the part of the IR2 model. The same was noted in contrasting the results of the three-predictor models. Hence, no conclusive evidence appears to be

available to indicate the superiority of any model type.

Another trend evident from the examination of Table 7 was that, for the Within-School Regression models, higher correlations with the a priori ranks occurred in the models where prediction was made about the means rather than at one standard deviation above or below the means. However, none of the comparisons was significant at $\alpha = .005$ within either the one-predictor or two-predictor models. One significant difference was noted across models when WR1 and WR2 were compared.

It should be recalled that each group received one a priori effectiveness rank which was based on all the individuals within the group. No attempt was made to manipulate the parameters so that some groups would be made more effective for individuals who had high or low predictor values. In nonhypothetical situations, some schools may be differentially effective for students at different levels of achievement or SES. The Within-School Regression models seem to be ideal for this type of situation. However, it was not intended to create this type of differential effectiveness in this study. Any variations in correlations between the model ranks and the a priori ranks were products of the models and sampling error, and not differential effectiveness.

Discrimination Ability

Within-School Models. The confidence intervals calculated for each effective and less effective group for each of the Within-School Regression models varied considerably in length. The length of the intervals is a function of group size, the

standard error for each equation, and the point about which the interval was constructed.

Insert Table 9 about here

Table 9 summarizes the number of significant comparisons between the effective and less effective groups for each Within-School Regression model. As expected, the models using means as comparison points (WSR1 and WR1) had a higher number of significant comparisons than did the other models in the family. Two of the 36 groups in question exhibited atypical behavior. One showed only four significant results in the 108 possible comparisons across the six models, the other showed only eleven. In both cases, sampling errors resulted in classifying the groups as average, thus making them more proximate to the comparison groups against which they were being contrasted. In addition, one group consisted of only 24 individuals. If these groups would be removed from the comparisons, the resulting discrimination accuracy of the WSR1 model would increase to 83.7% and the WR1 model to 88.9%, with a corresponding reduction in the probability of a Type I error for all comparisons considered simultaneously from .324 to .289.

The ability of the Within-School Regression models to discriminate between effective and less effective groups when prediction is made at the means was quite good despite the rather stringent α of .001. The one-predictor models showed slightly better discrimination than the two-predictor models. Group size and the location of the comparison points influenced

the ability of the models to discriminate to a greater extent than did the standard errors of the groups.

Residuals Models. Table 10 shows the distribution of PIs over the a priori classifications for each model. Only the SR1 model assigned an index of 5 exclusively to effective groups. Also, this same model assigned an index of 1 to 17 of the less effective groups. The distributions in the other models were rather similar. From the evidence presented in Table 10, it does not appear than one model is superior to the others in discriminating between effective and less effective groups using PIs.

Insert Table 10 about here

Table 11 shows the effects of different rules used to decide if two groups should be considered different on the basis of their PIs. For the data in this study, a rule of at least a a difference of 2 units correctly identified almost all of the comparisons between the effective and the less effective groups. The models using individual scores were slightly more accurate than the models using mean scores. Both one-predictor models correctly identified all but one comparison. However, a number of incorrect decisions would have been made using this rule, both within categories and between categories. For example, one effective group would have been misclassified if the Individual Residuals models were used, and three effective groups, if the School Residuals models were used. Also, most of the groups in the average category would have been declared

more effective than almost all of the groups in the less effective category, despite the fact that the real differences between some of these groups may not be large enough to warrant this discrimination.

Insert Table 11 about here

A more stringent decision rule requiring a difference of at least 3 units resulted in a percentage of the real differences being lost, however the number of misclassifications was likewise reduced. The most stringent rule of a difference of at least 4 units resulted in at least 74% correct classifications on each model with almost no misclassifications. This 74% compares favorably with the percentage of significant comparisons found with the WSR1 and WR1 models when statistical procedures were employed at a rather low significance level of .001.

From the above, it appears that PIs are useful in discriminating between groups which have rather large differences in effectiveness. Attempts to make fine discriminations would appear to be unwarranted. If it is desired to be able to discriminate almost all of the effective groups from the relatively ineffective ones in order to examine more closely why they may be effective or ineffective, a decision rule requiring at least a difference of 2 PI units would seem to be most useful. If misclassifications are of concern, more stringent decision rules would be more appropriate. A general strategy might be to use the most stringent rule initially to determine gross differences, and then apply the less strict rules in turn with increasing

caution. In this way, a rather good profile of the relative effectiveness of the schools involved should be obtained.

Summary

The results indicated that each type of model was capable of producing indices which were rather accurate reflections of the effectiveness ranks and classifications established a priori. No conclusive evidence was present to indicate that any one model was superior to the others in accomplishing this task. Since the School Regression Residuals models are easier to apply than the other models and the data for them are usually readily available, they could be considered superior in a cost-effectiveness sense. The use of PIs in conjunction with the School Regression Residuals models will enable appropriate discriminations to be made between most of the schools possessing differences in effectiveness. Different decision rules can be employed in accordance to their relative strictness in order to identify almost all of the schools possessing a certain degree of differential effectiveness.

The Within-School Regression model seems to be most useful in a situation where it is suspected that the schools may be differentially effective for students possessing different characteristics on the predictors used in the model. This model will generally produce a different set of ranks for each combination of predictor values. These ranks depend to a great extent on the sizes of the schools used and the location of the predictor values relative to their respective means. If schools

are not differentially effective for certain kinds of students, this model will yield results very similar to those produced by the other models. This model is difficult to apply since a regression equation must be obtained for each school and individual student data are required.

The results also indicated that, as additional predictors were added to the models, the correlations between the ranked indices and the a priori ranks decreased. This could be due to random error playing an increased role in establishing the effectiveness indices as the residual variation in the models decreased. This probably is related to the restriction in range phenomenon. Therefore, results from models using a different number of predictors may not be directly comparable. As a result, proper model specification, either through theory or the results of previous research or personal insight, is deemed essential in attempting to determine the relative effectiveness of schools through the use of indices from statistical models.

References

Burke, H. R. A study in public school accountability through the application of multiple regression through selected variables. Unpublished doctoral dissertation, Indiana University, 1972.

California Test Bureau/McGraw-Hill. Technical report: Comprehensive Tests of Basic Skills. Monterey:McGraw-Hill, 1970.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. Equality of educational opportunity. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1966.

Convey, J. J. Some methodological considerations for school effectiveness studies. Tallahassee: Florida State University, 1973. (ERIC Document Reproduction Service No. ED 092 597)

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Dyer, H., Linn, R., & Patton, M. Feasibility study of educational performance indicators: Final report to New York State Education Department. Princeton: Educational Testing Service, 1967.

Dyer, H., Linn, R., & Patton, M. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 4, 591-605.

Florida Ninth-Grade Testing Program. Technical report: 6-68. Tallahassee: Florida State University, 1968.

Forsyth, R. Some empirical results related to the stability of performance indicators in Dyer's Student Change Model of an educational system. Journal of Educational Measurement, 1973, 10, 7-12.

Gastright, J. F. Some empirical evidence on the stability of discrepancy measures based on observed and predicted school means on achievement tests. A paper presented to the annual meeting of the American Educational Research Association, Chicago, April, 1974.

Guthrie, J. W. A survey of school effectiveness studies. In U.S. Department of Health, Education, and Welfare, Do teachers make a difference? Washington: U.S. Government Printing Office, 1970.

Lindquist, E. F. Factors determining reliability of test norms. Journal of Educational Psychology, 1930, 21, 512-520.

Lindquist, E. F. Norms of achievement by schools. In A. Anastasi (Ed.) Testing problems in perspective. Washington: American Council on Education, 1966.

Lord, F. M. Test norms and sampling theory. Journal of Experimental Education, 1959, 27, 247-263.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 1974, 11, 225-234.

Miller, R. G. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

O'Connor, E. F. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Shaycoft, M. F. The high school years: Growth in cognitive skills. Pittsburgh: American Institute of Research, 1967.

Table 1

Specifications for Population Group Means

Variable	Mean	S.D.	Correlations		
			Input	SES	School
Output	539.00	34.360	.73	.45	.02
Input	474.00	27.320		.45	.02
SES	98.54	3.770			.03
School	23.14	1.635			

Table 2

Characteristics of Population Group Means

Variable	Mean	S.D.	Correlations		
			Input	SES	School
Output	535.37	40.06	.7772	.5329	-.0237
Input	477.70	31.91		.4135	-.0200
SES	98.21	4.20			-.1855
School	23.14	1.54			

Table 1

Specifications for Population Group Means

Variable	Mean	S.D.	Correlations		
			Input	SES	School
Output	539.00	34.360	.73	.45	.02
Input	474.00	27.320		.45	.02
SES	98.54	3.770			.03
School	23.14	1.635			

Table 2

Characteristics of Population Group Means

Variable	Mean	S.D.	Correlations		
			Input	SES	School
Output	535.37	40.06	.7772	.5329	-.0237
Input	477.70	31.91		.4135	-.0200
SES	98.21	4.20			-.1855
School	23.14	1.54			

Table 3

Plan for Distribution of Group Sizes

Number per Group	Number of Groups
20 - 99	18
100 - 199	18
200 - 299	9
300 - 399	9

Table 4

Group Sizes

Group	N	Group	N	Group	N	Group	N	Group	N	Group	N
1	166	10	185	19	293	28	179	37	188	46	156
2	94	11	223	20	143	29	89	38	349	47	205
3	213	12	330	21	270	30	145	39	24	48	150
4	62	13	104	22	55	31	326	40	174	49	110
5	104	14	368	23	399	32	143	41	259	50	337
6	69	15	54	24	23	33	127	42	289	51	188
7	71	16	375	25	241	34	48	43	323	52	20
8	102	17	296	26	70	35	333	44	47	53	53
9	42	18	103	27	97	36	129	45	69	54	75

Table 5

Characteristics of Generated Data

Variable	Mean	S.D.	Correlations		
	Based on the 54 Sample Group Means				
			Input	SES	School
Output	535.82	41.99	.7695	.4975	.0154
Input	477.43	32.86		.3960	.0577
SES	98.06	4.17			-.1013
School	23.14	1.54			
	Based on the 9087 Individual Scores				
			Input	SES	School
Output	536.36	94.35	.7429	.4577	-.0130
Input	478.95	74.59		.4401	-.0302
SES	98.38	10.09			-.0359
School	23.24	1.38			

TABLE 6

A Priori, Sample, And Model Ranks

[a]Group	Samp	WSR1	WSR2	WSR3	WR1	WR2	WR3	IRR	IR2	IR1	SRR	SR2	SR1
1	2	2	3	2	2	2	2	2	2	2	2	2	2
2	1	1	2	1	1	1	1	1	1	1	1	1	1
3	3	4	8	3	3	3	3	4	4	3	6	6	3
4	4	6	5	6	5	8	4	5	6	6	5	5	4
5	6	3	1	4	6	6	7	3	3	5	3	3	6
6	9	10	10	11	9	7	11	11	11	9	14	14	9
7	8	7	4	10	8	9	9	7	7	7	8	8	8
8	10	9	12	7	11	10	8	9	8	10	9	9	10
9	5	5	7	5	4	5	5	6	5	4	4	4	5
10	7	8	6	14	7	4	10	8	9	8	10	10	7
11	11	13	21	9	10	11	6	16	16	11	23	23	11
12	14	19	22	19	14	15	14	20	18	14	18	18	14
13	27	29	34	22	29	28	28	32	32	29	35	35	28
14	17	18	16	20	17	17	19	18	19	18	17	17	16
15	12	11	11	8	12	12	15	10	10	13	7	7	12
16	21	23	25	23	19	20	23	23	23	19	24	24	20
17	15	12	15	13	15	21	17	12	12	16	11	11	15
18	13	14	19	12	13	14	12	13	14	12	16	16	13
19	18	15	14	18	18	18	20	15	15	17	15	15	19
20	16	31	32	27	27	27	22	28	29	22	28	28	17
21	23	26	26	25	26	25	24	24	25	25	22	22	23
22	30	33	30	34	30	31	31	33	33	31	32	32	31
23	26	24	31	17	24	29	18	25	24	27	25	25	26
24	19	25	9	45	32	13	49	14	13	15	13	13	18
25	28	17	20	16	16	24	16	22	21	24	26	26	27
26	31	21	27	15	21	32	13	30	27	30	29	29	30
27	33	43	43	42	31	34	29	38	39	32	44	44	32
28	25	30	23	36	28	23	30	26	26	26	30	30	25
29	22	16	13	21	22	19	27	17	17	23	12	12	22
30	24	22	17	30	23	22	26	21	22	21	21	21	24
31	29	27	28	24	25	26	21	27	28	28	34	34	29
32	20	20	18	26	20	16	25	19	20	20	19	19	21
33	35	32	29	32	33	33	34	28	30	33	27	27	34
34	36	38	33	43	34	30	39	38	38	36	40	40	36
35	34	35	42	29	35	41	32	34	34	35	33	33	35
36	37	37	39	33	38	38	37	37	37	37	36	36	37
37	42	34	38	28	39	44	35	35	35	41	31	31	41
38	45	47	47	47	46	47	48	48	48	46	46	46	45
39	32	28	24	31	36	43	33	31	31	34	20	20	33
40	43	36	35	37	43	42	40	40	40	42	38	38	43
41	40	40	37	39	41	36	42	36	36	39	37	37	39
42	44	44	41	44	44	40	46	45	44	44	45	45	44
43	38	39	45	35	37	39	36	41	41	38	42	42	38
44	52	52	52	51	52	50	52	44	45	52	43	43	52
45	39	46	40	50	40	35	45	46	46	40	48	48	40
46	48	49	50	49	50	46	50	52	52	51	52	52	49
47	46	42	44	38	45	48	41	43	42	45	41	41	46
48	41	41	36	40	42	37	43	42	43	43	39	39	42
49	51	50	53	48	49	51	44	50	51	49	53	53	51
50	50	51	46	52	51	45	51	51	50	50	50	50	50
51	49	45	48	41	48	49	47	47	47	47	47	47	48
52	47	48	49	46	47	53	38	49	49	48	49	49	47
53	54	54	54	53	54	54	53	54	54	54	51	51	54
54	53	53	51	54	53	52	54	53	53	53	54	54	53

[a]Group number and a priori rank are identical.

TABLE 7

Intercorrelation Matrix of Ranks[a]

	A Pr	Samp	WSR1	WSR2	WSR3	WR1	WR2	WR3	IRR	IR2	IR1	SRR	SR2	SR1
Samp	9578													
WSR1	9247	9626												
WSR2	8824	9443	9647											
WSR3	9038	9095	9611	8627										
WR1	9508	9756	9781	9319	9510									
WR2	9318	9748	9450	9531	8646	9607								
WR3	9158	9238	9405	8472	9703	9690	8770							
IRR	9275	9725	9830	9738	9193	9636	9616	9028						
IR2	9297	9717	9855	9740	9232	9655	9612	9059	9986					
IR1	9536	9944	9702	9550	9094	9800	9809	9213	9817	9821				
SRR	8902	9447	9685	9688	9005	9261	9197	8634	9834	9832	9497			
SR2	8902	9447	9685	9688	9005	9261	9197	8634	9834	9832	9497	1.00		
SR1	9597	9992	9648	9463	9114	9779	9764	9254	9756	9751	9963	9467	9467	

[a] all entries are significant at $\alpha = .001$.

TABLE 8

Z Values For Selected Pairwise Differences Between Correlations
Of The Models With The A Priori Rankings

	WSR1	WSR2	WSR3	WR1	WR2	WR3	IRR	IR2	IR1	SRR	SR2
WSR2	-2.44										
WSR3	-1.37	.77									
WR1	2.35	2.94*	2.63								
WR2	.46	2.53	1.21	-1.51							
WR3	-.53	1.20	.89	-2.53	-.76						
IRR	.30	nc	nc	-1.82	nc	nc					
IR2	.58	nc	nc	-1.72	nc	nc	.81				
IR1	2.31	nc	nc	.35	nc	nc	2.54	2.42			
SRR	-2.21	nc	nc	-2.72	nc	nc	-2.90*	-3.00*	-3.08*		
SR2	-2.21	nc	nc	-2.72	nc	nc	-2.90*	-3.00*	-3.08*	0.0	
SR1	2.57	nc	nc	1.08	nc	nc	2.72	2.60	1.67	3.23*	3.23*

* indicates significant at $\alpha = .005$.

nc indicates that no comparison was made.

Table 9

Summary of Confidence Interval Comparisons

Model	Significant Comparisons	Percent Significant
WSR1	246	75.9
WSR2	138	42.6
WSR3	142	43.8
WR1	262	80.9
WR2	196	60.5
WR3	185	57.1

Table 10

Relationship Between Performance Indices And A Priori Classifications For Each Residuals Model

SRR / A Priori	5	4	3	2	1
Effective	15	0	2	1	0
Average	3	3	6	3	3
Less Effective	0	1	1	0	16

IRR / A Priori	5	4	3	2	1
Effective	15	2	1	0	0
Average	3	5	5	2	3
Less Effective	0	0	1	1	16

SR2 / A Priori	5	4	3	2	1
Effective	15	0	2	1	0
Average	3	3	6	3	3
Less Effective	0	1	1	0	16

IR2 / A Priori	5	4	3	2	1
Effective	16	1	1	0	0
Average	3	5	5	2	3
Less Effective	0	0	1	1	16

SR1 / A Priori	5	4	3	2	1
Effective	15	2	1	0	0
Average	0	7	5	3	3
Less Effective	0	0	0	1	17

IR1 / A Priori	5	4	3	2	1
Effective	15	2	1	0	0
Average	2	8	2	3	3
Less Effective	0	0	0	1	17

Table 11

Frequency of Differences Under Three Decision Rules

Models	At Least 2 Units		At Least 3 Units		At Least 4 Units	
	Freq	Percent	Freq	Percent	Freq	Percent
SRR	302	93.2	240	74.1	240	74.1
SR2	302	93.2	240	74.1	240	74.1
SR1	323	99.7	304	93.8	255	78.7
IRR	320	98.8	287	88.6	240	74.1
IR2	321	99.1	288	88.8	256	79.0
IR1	323	99.7	304	93.8	255	78.7