ED 063 332 TM 001 350

AUTHOR	Forsyth, Robert A.
TITLE	Considerations Related to the Usefulness of the Performance Indicators in Dyer's Student Change Model of an Educational System.
PUB DATE	Apr 72
NOTE	21p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Educational Accountability; *Evaluation Techniques; Input Output Analysis; *Mathematical Models; *Program Evaluation; Reliability; Research Methodology; *Statistical Analysis; Summative Evaluation
IDENTIFIERS	*Performance Indicators

ABSTRACT

In Dyer's Student Change Model, performance indicators (PIs) are utilized as a measure of the effectiveness of educational programs. These PIs are examined and analyzed in some detail in this paper. In particular, the reliability of the residuals, the sensitivity of the measure to sample size, the random half reliability, and the stability of the PIs for consecutive classes are examined. Empirical results of a study of PI stability are presented. (DG)

Considerations Related to the Usefulness of the Performance Indicators in Dyer's Student Change Model of an Educational System

Robert A. Forsyth
University of Iowa

A paper presented at the annual convention of the American Educational Research Association, Chicago, Illinois, April 4-7, 1972

Introduction

The evaluation of the effectiveness of educational programs is receiving wide attention by both professional educators and the public. The notion of educational accountability, to cite an obvious example, is largely concerned with evaluation of programs and those responsible for program implementation. Much of this discussion has dwelled on models for the evaluation. Simple as the problem may appear to be, the effectiveness of a given school's educational program is difficult to measure. One of the most intriguing proposals is that of Dyer, Linn and Patton (1967, 1969) and Dyer (1970a, 1970b). Dyer has called this proposal the "student change model of an educational system" (1970a, P. 98). Basically, the model conceptualizes four groups of variables -- input variables, educational process variables, surrounding condition variables (further subdivided into hard-to-change and easy-to-change), and output variables. Based on these groups of variables, performance indicators (to be described later and hereafter labeled PIs) are computed for each educational system. Dyer (1970a, Pp. 100-101) states:

> The first step in the student-change model to produce educational performance indicators is to put two groups of variables on the shelf for future reference. These are all the educational process variables...and all the easy-to-change surrounding conditions...we use the remaining variables -- input, output, and hard-to-change conditions -- to figure the performance indicators.

Performance indicators are to be computed for a number of different areas of student development and at several levels of the school system.

For example, Dyer (1970a, P. 102) gives the following matrix of performance indicators for a hypothetical school. In this example, "high" performance is implied by a value of five, "low" performance by a value of one.

Performance of School Ssytem X, 1976-1979

Areas of Student Development

	Self-understanding and self-acceptance	Academic development	Social behavior	Vocational development	Physical well-being
Grades 10-12	3	5	2	2	5
Grades 7-9	4	5	2	4	5
Grades 4-6	2	3	3	2	5
Grades 1-3	1	5	4	2	5

Hypothetical Matrix of
Educational Performance Indicators
Figure 1

As this example indicates, each school would have a large number of PIs. They would not be restricted to academic proficiencies alone.

Each PI is obtained by first computing a predicted output score using the input and hard-to-change variables as predictors in a multiple linear regression equation. Symbolically, (using the notation of Dyer, Linn, and Patton, 1967, P. 58):

$$\hat{O}_{cgs} = \sum_{c=1}^{m} b_{I_c} I_{cg's} + \sum_{j=1}^{p} b_{s_j} S_{j(g'-g)s} + a$$

where

$\hat{O}_{cgs}$ is the predicted output mean for category c at grade level g and school s.

$I_{icg's}$ is the mean input for school s on variable i at grade level g' for category c.

$S_{j(g'-g)s}$ is the mean value of hard-to-change surrounding condition j obtaining between years g'-g at school s.

b_{Ici} and b_{sj} are the regression coefficients and a is a constant for all schools.

m is the number of input variables and p is the number of hard-to-change variables operating in the situation.

The actual PIs are derived from the quantity

$$\frac{O_{cgs} - \hat{O}_{cgs}}{SEM_{cg}} = \text{Index}$$

where

$SEM_{cg} = \overline{SD}_{cg} / \sqrt{\overline{n}}$. That is, SEM_{cg} is the average within school standard deviation for the output measure in category c at grade level g divided by the square root of the average number of students per school in grade g.

PIs are assigned via the following scheme (Dyer, et al., 1967, P. 59):

$\text{Index} < -1.5$: $PI_{cgs} = 1$

$-1.5 \leq \text{Index} < -.5$: $PI_{cgs} = 2$

$-.5 \leq \text{Index} \leq .5$: $PI_{cgs} = 3$

$.5 \leq \text{Index} \leq 1.5$: $PI_{cgs} = 4$

$\text{Index} > 1.5$: $PI_{cgs} = 5$.

The PIs are utilized to identify schools that seem to be performing either above expectations or below expectations with respect to a particular class of educational outcomes. [Note: There is an implicit assumption in the model that within the group of schools used in developing the frame of reference, a range of quality is represented -- some schools are really better than others. Otherwise, the PIs will reflect random differences, since by the very nature of the mathematics some schools will have positive deviations and other negative deviations. In current jargon, one might say that PIs are wholly norm referenced.] Once the exceptional schools are identified, it is presumed that they can be studied in terms of the educational process variables and the easy-to-change input variables. The obvious purpose would be to identify possible causes of the discrepancies. (See Dyer 1970, P. 5 for a specific example.)

Of course, before the PIs are utilized to identify exceptional schools, it is reasonable to demand that differences in PIs represent more than just error variation. Dyer, et al. (1967, Pp. 45-46) write:

> The meaningfulness of the departures of actual means from predicted means would be investigated in three ways. First, the variance of the simple differences between school output means and the corresponding input means would be computed and the hypothesis that this variation is due only to errors of measurement would be tested by means of an F-test. Similarly, an F-test would be used to test the hypothesis that the standard error of estimate is only measurement error...Following these analyses, the

> performance indices would be computed for each school system, using [random] half-samples [of each grade]. The regression weights developed in the first sample would then be applied to the hold-out sample within each system and the performance indices computed for the hold-out samples. The indices for the two samples for each school would then be computed in order to get an indication of the stability of the indices....

The first test suggested by Dyer et al. is not of much interest. It makes sense only when the output mean is compared to a similar input mean. For example, it would be reasonable to compare 6th grade reading achievement to 3rd grade reading achievement. However, it is obvious that even in this instance no one expects the mean of the output measure to be equal to the mean of the input measure. Also, a variety of input variables will usually be used for each PI, each measured in its own scale of units. Therefore, it isn't reasonable to test the equality of the output mean against each input mean.

The second and third procedures implied in the above quote do, however, seem relevant. The second F-test relates to the reliability of the residuals $(O-\hat{O})$.* In essence we must first establish that the residuals have a reliability greater than zero. The type of reliability of concern is related to the use of parallel measures of imput and output. That is, would the residuals obtained under one set of measures be consistent with those obtained for a parallel set of measures.

One of the major purposes of this paper is to consider the conditions under which the residuals would be "reliable." Specifically, a

*Throughout the remaining part of this paper, subscripts for O and $\hat{O}$ will be omitted.

rationale will be given to support the conclusion that for most, if not all, applications of this procedure the residuals would be judged reliable by the proposed F-test. However, in some applications the practical significance of residuals would certainly be questioned.

The final criterion proposed by Dyer et al. related to the stability of the PIs across random samples of students from the same grade level of the local system. (The previous reliability question was related to stability across forms of the measuring instruments.) To assess this, Dyer et al. (1967) propose that the available group be randomly subdivided, and the procedures be applied independently to the two subsamples. This method of taking random halves of classes seems to be a legitimate procedure for considering the stability of the indices, since the PIs are measures of school performance and not individual class performance. It is likely that the students in these two subsamples will not have had exactly the same educational program. However, the procedure randomizes the various factors that have contributed to differential educational programs. Essentially, this type of reliability is concerned with pupils as a source of error.

Another type of reliability that might be considered important is the stability of the PIs for consecutive classes. This type of reliability is concerned with pupils and factors which vary over time as sources of error. It is true that it might be unreasonable to consider two consecutive classes as random samples of the same population. However, if the PIs are to be useful as measures of school performance they should remain relatively stable from class to class, assuming no "out-of-the-ordinary" changes in the educational program have been made. Thus, it seems reasonable to investigate the stability of the PIs assigned to schools for two distinct classes going through "similar" educational programs. If the PIs prove to be stable under these conditions, then they would seem to be useful indicators of school performance. They would reflect presumably changeable aspects of the school program. However, if the PIs are not stable under these conditions, then they may be more a function of the idiosyncrasies of a particular class rather than the school. (As noted earlier, this assumes that no systematic changes were made in the school programs for one of the classes and/or that environmental conditions didn't change drastically.)

Dyer, et al. (1969) have provided some evidence related to the stability of PIs [between random halves of school classes] for some cells of the matrix of PIs. Such indices represent estimated upper bounds for the stability indices for consecutive classes. However, no evidence has been presented related to the stability of the PIs for two consecutive classes.

Another major purpose of this paper is to present some empirical results related to this type of stability. These results are related only to a part of the matrix of educational performance indicators. Specifically, PIs for grades 9-12 in the academic development area are examined.

Reliability of Residuals

Stanley (1971) and Thorndike (1963) present a formula for the reliability (parallel form assumptions) of the difference between actual and predicted output. This formula may be adapted to the present context of school means and multiple predictors.

Let $R_{(O-\hat{O})}$ = population reliability of residuals.

R_O = population reliability of output measure.

$R_{\hat{O}}$ = population reliability of the predicted output measure (a linear composite of input measures).

$R^2_{O\hat{O}}$ = population squared multiple correlation coefficient.

$$\text{Then, } R_{(O-\hat{O})} = \frac{R_O - 2R^2_{O\hat{O}} + R^2_{O\hat{O}} R_{\hat{O}}}{1 - R^2_{O\hat{O}}} \qquad (1)$$

Before considering the factors which influence the magnitude of the $R_{(O-\hat{O})}$, it would be beneficial to estimate what sample value of $R_{(O-\hat{O})}$ $[\hat{R}(O-\hat{O})]$ is needed before the variability of the residuals may be considered more than measurement error. The F-test proposed by Dyer et al. would probably be of the following form:*

$$F_{(N-k-1),\ (N-1)} = \frac{1 + \hat{R}_{(O-\hat{O})}}{1 - \hat{R}_{(O-\hat{O})}} \qquad (2)$$

Where N = number of schools
k = number of predictor variables

*F-test is a modified form of Equation (13) in Kristof (1969).

If, for example, we assume $N = 50$ and $k = 3$ (such values would seem to be lower bounds for computing PIs), then an $\hat{R}_{(O-\hat{O})}$ of approximately .25 would be required for significance at the .05 level.

Under what conditions will values of $\hat{R}_{(O-\hat{O})}$ above .25 be obtained? To consider this question rewrite Equation (1) as follows:

$$R_{(O-\hat{O})} = \frac{R_O - R^2_{O\hat{O}}\,(2 - R_{\hat{O}})}{1 - R^2_{O\hat{O}}} \qquad (3)$$

From Equation (3) two inferences may be drawn. First, the maximum value of $R_{(O-\hat{O})}$ is R_O. That is, the reliability of the residuals cannot be greater than the reliability of the criterion measure. Furthermore, and this is not obvious from Equation (3), the reliability of the output measure is much more crucial in determining $R_{(O-\hat{O})}$ than the reliability of the predicted measure. This is true because the variability of O is $R_{O\hat{O}}$ times the variability of O. Given these conditions, it is useful to consider the expected magnitude of R_O. Of course, no definite answer can be given to this question for all output measures. However, it should be noted that both R_O and $R_{\hat{O}}$ will probably be very high in most applications of this PI procedure. Since both input and output measures are means of school systems, it would be very likely that these measures would have high reliabilities even if the reliability of individual measures were low. Thus, unless the schools are unusually homogeneous, the reliability of the output measure would be of such a magnitude that the maximum possible value of $R_{(O-\hat{O})}$ would be much greater than .25.*

Of course, the reliabilities of the input and output measures are not the only factors influencing the magnitude of $R_{(O-\hat{O})}$. The second fairly obvious condition indicated by Equation (3) is that the maximum value of $R_{(O-\hat{O})}$ (i.e., $R_{(O-\hat{O})} = R_O$) occurs when $R^2_{O\hat{O}} = 0.0$. That is, the residuals have maximum reliability when no differential prediction is possible. Thus, if R_O and $R_{\hat{O}}$ are constant, $R_{(O-\hat{O})}$ decreases as $R^2_{O\hat{O}}$ increases. (Equations (1) and (3) are not defined when $R^2_{O\hat{O}} = 1$.)

It would be enlightening to examine the magnitude of $R_{(O-\hat{O})}$ when $R^2_{O\hat{O}}$ is relatively great. Assume that $R_O = R_{\hat{O}} = .90$ (This assumption does not seem unreasonable in view of our earlier discussion.). Also, assume $R^2_{O\hat{O}} = .64$ (i.e., $R_{O\hat{O}} = .8$). In this instance the reliability of $(O-\hat{O})$ is approximately .54. If $R^2_{O\hat{O}} = .7225$ ($R_{O\hat{O}} = .85$), $R_{(O-\hat{O})}$ is still approximately .38.

In summary, the above discussion has been intended to support the following statements:

1) The reliability of the residuals is primarily influenced by the reliability of the criterion measure.

2) In most, if not all, applications of this technique R_O will be very high. Only if the schools under study are extremely homogeneous, would this not be true.

3) As $R^2_{O\hat{O}}$ increases, $R_{(O-\hat{O})}$ decreases (assuming constant reliabilities for O and $\hat{O}$).

*The question could be raised, "What is unusually homogeneous?" Perhaps, an example would give some indication. Assume an average school size of 50. Furthermore assume that the variance of school means is equal to 5% of the pupil score variance. If the reliability of pupil scores is .50, then the estimated reliability of the means is still .80.

4) Since R_O and $R_{\hat{O}}$ will usually be high relative to $R^2_{O\hat{O}}$, $R_{(O-\hat{O})}$ would be judged significantly greater than zero.

Therefore, given these "facts", low reliability (as judged by a significance test) of the residuals would usually not be a factor restricting the use of PIs. However, if $R^2_{O\hat{O}}$ is too high the practical significance of the residuals is open to question and should be considered. Furthermore, if $R^2_{O\hat{O}}$ is very high, output is being determined by input and hard-to-change surrounding variables regardless of easy-to-change variables or educational process variables.

Sample Size Considerations

It is not just the reliability of residuals that must be considered when the utility of PIs is examined. It may be remembered that Dyer's PIs were derived from the following quantity:

$$\frac{O-\hat{O}}{SEM} = \frac{O-\hat{O}}{\overline{SD}/\sqrt{\overline{n}}}$$

An examination of the reliability of the residuals merely concerns the numerator of this quantity. It is also important to examine the denominator and the relationship of the denominator to the numerator. Let's assume first that the residuals are highly reliable. The numerator and denominator of the index are, for all practical purposes, independent. For any given $R^2_{O\hat{O}}$, there exists a wide range of possible values for $\overline{SD}/\sqrt{\overline{n}}$. Since $\overline{SD}$ would be a relatively constant value for a given outcome regardless of the size of the schools, the value of the denominator is basically a function of $\overline{n}$. As $\overline{n}$ increases, the denominator decreases. If $\overline{n}$ is very large, a large number of PIs with values of 5 and 1 would be obtained. A specific example will help clarify the above comments. Assume $R_{O\hat{O}}$ is .68, $\overline{SD} = 6.7$, and $\overline{n} = 66$.

If the residuals are assumed to be normally distributed with mean = 0 and standard deviation = 1.06 (these represent actual results from analyses to be presented later), then the estimated percent of PIs with values of 5 or 1 is 24%. (For the actual data this percent was 26.) However, if $\overline{n} = 400$, then for approximately 64% of the schools the PIs would be either a 5 or a 1. In the second situation approximately 12% would have PIs of 3. Thus, with a very large $\overline{n}$ the PIs have little utility, since relatively small residuals could yield PIs of 5 or 1. This situation is analogous to the statistical vs. practical significance problem confronted in inferential statistics when hypothesis testing procedures are based on large sample sizes.

In summary, it would seem that the utility of PIs is greatest when the average school size is not so great that small residuals become too important. This is not to say that attempts to identify "over- and under-performing" schools is not important when average school size is large. It merely implies that the PIs do not have great utility in this situation. If $\overline{n}$ is very large, perhaps some judgment should be attempted of the magnitude of an "important" deviation of O from $\hat{O}$, rather than use $\frac{O-\hat{O}}{SD/\sqrt{\overline{n}}}$ to determine PIs. Alternatively, one might arbitrarily undertake a study of the 10% or 20% of schools at the extremes of the $(O-\hat{O})$ distribution. The hope would be, that these extremes would help identify what the good schools do that is so much better than the poor schools.

Stability of PIs

Dyer, Linn, and Patton (1969) have provided some evidence regarding the stability of PIs for random halves of the same class. Their results pertain to only a part of a matrix of performance indicators. They identified a matched sample of students in 64 schools that were in grade 5 during the 1957-58 school year and in eighth grade during the 1960-61 school year. Eighth grade means of the 5 major subtests and the composite of the Iowa Tests of Basic Skills (Lindquist and Hieronymus, 1956) were used as output variables. Input variables included all 15 ITBS scores in the 5th grade. The students in the grades in the 64 schools were divided at random into two groups and independent regression analyses were conducted for each output measured. The residuals ($O-\hat{O}$) were computed for each half and the correlations between the residuals were calculated. The correlations ranged from .62 (Vocabulary) to .84 (Work Study Skills). The other correlations were .66, .67, .76, and .77. Although the statability of the PIs can only be inferred from these correlations, it would seem that there was relatively high stability.

The final part of this study presents similar empirical evidence related to the stability of PIs for consecutive classes. As stated above, these results are related only to one segment of the matrix of performance indicators, specifically, the area of academic development at the secondary level was examined.

Empirical Results Related to the Stability of PIs

Procedures

Each year a large number of Iowa high schools participate in a state-wide testing program by administering the Iowa Tests of Educational Development (Lindquist and Feldt, 1960) during the first few weeks of school. A total of 320 schools had administered these tests (Forms X-4 and Y-4) in 1965 and 1966 to their ninth grade students and also in 1968 and 1969 to the 12th grades. A random sample of 50 schools was selected from these schools. Matched longitudinal samples were formed for two time periods 1965-68 and 1966-69. The school means for the matched longitudinal samples of the ITED were utilized as input and output variables. That is, the 1965 means on the nine subtests and composite on the ITED were used as predictors of each of the ten ITED means for 1968. Likewise, the 1966 ITED means were used as predictors of the 1969 means.

A stepwise multiple linear regression technique was utilized with both groups of students. To be included in the prediction equation a variable had to increase the squared multiple correlation significantly ($\alpha = .05$). [This procedure is slightly different than the one used by Dyer, Linn and Patton (1969). They added input variables until the squared multiple correlation increased by less than .01. However, we are dealing with a sample of schools. Therefore, the use of a statistical test seems justified.] Deviations of actual means from predicted means were computed for both groups.

To compute the PIs it was necessary to find the SEM $= \overline{SD}/\sqrt{n}$ for each test for both the 1968 and 1969 groups. The average sample size for each year was 66. Rather than use the $\overline{SD}$ for these 50 schools, it was decided (primarily for simplicity) to estimate the within school standard deviation from the norms for the entire state of Iowa. This was done by using the relationship $\sigma^2_{TOT} = \overline{\sigma^2}_{\text{W Schools}} + \sigma^2_{\text{School Means}}$.

Thus, the SEM used in these analyses were computed from state norms for pupils and school averages. After the SEM was obtained the quantity, $\frac{O-\hat{O}}{SEM}$, was computed for each school on each of the ten scores for both the 1965-68 and 1966-69 data. The PIs for the two classes were then compared.

Results and Discussion

Table 1 presents the multiple correlations between output means and input means for the two classes: (1) 1965 scores (9th grade) to predict 1968 scores (12th grade); (2) 1966 scores (9th grade) to predict 1969 scores (12th grade). In addition, the correlations between the residuals from those two prediction equations are given.

It can be seen that the median multiple R for the 1965-68 predictions was .76 and for the 1966-69 predictions .80. These values are very consistent with the Dyer et al. (1969) data. For their matched longitudinal sample, the median multiple R was .80 (over 5 subtests and the composite of the ITBS). Although Dyer, et al, utilized 5th grade scores on the ITBS to predict 8th grade scores on the ITBS, this comparison of their data with the data of this study is useful since both sets of data are based on academic achievement variables.

The last column in Table 1 gives the values of the correlations between the residuals of two consecutive classes. The median R was approximately .28. Dyer, et al. (1969) correlated residuals obtained from prediction equations for random halves of the same class. For the five subtests and the composite score of the ITBS, those values were .62, .66, .76, .84, .67, and .77. The median r was reported as .72.

Thus, the correlations in Table 1 age of a much lower magnitude than the Dyer, et al. values.

Of course, because of a variety of confounding factors, it is not possible to say that the differences in residual correlations are solely due to the fact that these indices were reflecting stability from class to class while those of Dyer et al. were reflecting stability from sample to sample from the same classes. It may be true that using the ITBS scores at grades 5 and 8 for two consecutive classes would produce stability indices much higher than those found in this study. Furthermore, it must be remembered that the stability indices reported in Table 1 probably represent lower bounds to the true stability indices, since no effort was made to check on any systematic or out-of-the-ordinary changes taking place in the schools or communities participating in this study.

The above data relate to the stability of residuals. Of course, the stability of the PIs is directly related to the stability of the residuals. Nonetheless, since the model would use PIs in the decision making aspects, and since the residuals are only one part of the PIs, the stability of PIs for consecutive class will be examined also. Table 2 contains data relevant to the agreement of PIs for two consecutive classes.

It can be seen from Table 2 that perfect agreement was unusual (range from 16% (RNS) to 36% (C) of the schools). However, differences of one unit between PIs probably would not present many problems. That is, considering how the model would be used, differences of one unit would not seem to be crucial. Thus, if a disagreement is defined as PIs differing by more than 1 for the two classes, the percent of disagreements ranged from 12% (C) to 38% (RNS). These percents of disagreement should be regarded as upper bounds, since no effort was made to check on any systematic changes taking place in the 50 schools. Thus, these results seem to provide some evidence that the PIs for the academic area (as measured by standardized achievement tests) are relatively stable.

However, considering the types of decisions that are implicit in this model, it would seem reasonable that PIs be computed for two successive classes before any identification of "over- and under-performing" schools is attempted. If this is done, consistently high or consistently low schools may be identified and studied for possible causes of the observed results. Furthermore, it is possible that widely discrepant PIs for a given cell for two consecutive classes could indicate that some systematic program or environmental change has occurred. In this instance, it may also be fruitful to study these schools for possible factors "causing" the discrepancy.

TABLE 1

Multiple Correlations and Residual Correlations

TEST	Multiple R 1965 Predicts 1968	 1966 Predicts 1969	Residual 65-68 vs. 66-69
Social Studies Background (SSB)	.79	.69	.50
Natural Science Background (NSB)	.69	.80	.16
Correctness and Appropriateness of Expression (E)	.76	.81	.36
Ability to do Quantitative Thinking (Q)	.75	.85	.11
Ability to Interpret Reading Materials in Social Studies (RSS)	.68	.68	.42
Ability to Interpret Reading Materials in Natural Sciences (RNS)	.71	.68	.13
Ability to Interpret Literary Materials (LM)	.87	.74	.23
Vocabulary (V)	.86	.82	.11
Use of Sources of Information (SI)	.76	.83	.37
Composite (C)	.89	.84	.32
Median R	.76	.80	.28

TABLE 2

Agreements or Disagreements Between PIs for Two Consecutive Classes

	PIs Agree	PIs Differ by 1	PIs Differ by 2	PIs Differ by 3	PIs Differ by 4
	N %	N %	N %	N %	N %
SSB	14(28)*	26(52)	8(16)	2(4)	0(0)
NSB	11(22)	22(44)	14(28)	2(4)	1(2)
E	15(30)	24(48)	9(18)	2(4)	0(0)
Q	13(26)	23(46)	9(18)	5(10)	0(0)
RSS	17(34)	15(30)	11(22)	5(10)	2(4)
RNS	8(16)	23(46)	13(26)	4(8)	2(4)
RL	13(26)	29(58)	6(12)	2(4)	0(0)
V	15(30)	27(54)	7(14)	1(2)	0(0)
SI	15(30)	21(42)	11(22)	3(6)	0(0)
C	18(36)	26(52)	4(08)	2(4)	0(0)

*Numbers in parentheses indicate percent of the 50 schools.

References

Dyer, H. S.; Linn, R. L.; and Patton, M. J. Feasibility Study of Educational Performance Indicators: Final Report to New York State Education Department. Princeton: Educational Testing Service, 1967.

Dyer, H. S.; Linn, R. L; and Patton, M. J. 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, 6, 591-606.

Dyer, H. S. Toward Objective Criteria of Professional Accountability in the Schools of New York City. Phi Delta Kappen, 1970a, 52, 206-211.

Dyer, H. S. Can We Measure the Performance of Educational Systems. National Association of Secondary Schools Principals Bulletin, 1970b, 54, 96-105.

Kristof, W. Statistical Inference about Error Variance. Psychometrika, 1963, 28, 129-144.

Kristof, W. Estimation of True Score and Error Variance for Tests Under Various Equivalence Assumptions. Psychometrika, 1969, 34, 489-508.

Lindquist E. F. and Feldt, L. S. Iowa Tests of Educational Development. Chicago: Science Research Associates, 1960.

Lindquist, E. F. and Hieronymus, A. N. Iowa Tests of Basic Skills. Boston: Houghton Mifflin, 1956.

Stanley, J. Reliability. In Thorndike, R. L. (Ed.), Educational Measurement (2nd ed.) Washington, D. C., American Council on Education, 1971, Chapter 13.

Thorndike, R. L. The Concepts of Over- & Under-Achievement. New York: Columbia University Press, 1963.

Wohlfred, G. Toward an Evaluation of Education: A Description of the Quality Measurement Project. (2nd ed.) Albany: New York State Education Department, 1969.

Wohlfred, G. Quality Evaluation Through Nomographs. Albany: New York State Education Department, 1970.