ED 350 314 TM 019 009

AUTHOR	Lang, Michael H.; And Others
TITLE	The Effect that Varying the Test Mode Had on School Effectiveness Ratings.
PUB DATE	Apr 92
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Testing; *Criterion Referenced Tests; *Elementary Schools; *Evaluation Methods; Grade 3; Language Tests; Mathematics Tests; *Norm Referenced Tests; Primary Education; Public Schools; Reading Tests; *School Effectiveness; State Surveys; *Test Use
IDENTIFIERS	Educational Indicators; Kappa Coefficient; Louisiana; Residuals (Statistics); Testing Effects

ABSTRACT

The effect of alternating criterion-referenced tests (CRTs) with norm-referenced tests (NRTs) when evaluating schools was studied in a sample of 242 Louisiana public elementary schools accounting for over 18,000 third graders tested in 1989. The study used five separate multiple regression models, each producing studentized residuals used as school effectiveness indicators (SEIs). The SEIs were used to classify schools as effective, average, or ineffective. Each school was classified according to 10 different models. Cross-classification results were analyzed separately for: (1) CRT language arts and NRT language; (2) CRT language arts and NRT reading; and (3) CRT mathematics and NRT mathematics. Each comparison was tested with the kappa z-test. These tests were found to be significant beyond the 0.05 level. Magnitude measures were generally moderately consistent for CRT-NRT comparisons. Findings do not support alternating test modes when evaluating schools. Five tables present study data, and 19 references are included. (Author/SLD)

THE EFFECT THAT VARYING THE TEST MODE HAD ON SCHOOL EFFECTIVENESS RATINGS

by Michael H. Lang
Charles Teddlie
Jeffery Oescher

The Louisiana Department of Education has recently implemented its first school incentive program to acknowledge and reward those public schools which demonstrate progress toward effectively educating their students. The method of determining which schools receive awards was based partly on standardized test scores as applied to school category groups: the highest scoring schools in each category on various tests and other indicators received both monetary and nonmonetary awards.

In addition, the department has begun a school performance comparison program, also based on school category groups. Again, the comparisons are being based partly on standardized test scores.

However, Louisiana, like other states, does not test every grade statewide, nor does it test every grade with the same mode of testing (See Table 1). The Louisiana Educational Assessment Program (LEAP) measures achievement the 4-6-9 grades with norm-referenced tests (NRTs), the 3-5-7 grades with criterion-referenced tests (CRTs), and the 10-11 grades with the Graduation Exit Examination, also a CRT. The Louisiana CRTs are essentially curriculum-based measures.

For apparent financial reasons, incentive awards and school evaluation programs are employing existing state assessment programs as do large scale effective schools studies. Ideally, performance on core subjects in every grade in a given incentive program would be assessed, and each subject/grade would be tested with the same mode (or modes) of assessment. That is, the program would test achievement in every grade with either the NRT mode, the CRT mode, or both.

Table 1

States with School Evaluation Programs and the Tests They Employ at Each Grade

	Grades: 1	2	3	4	5	6	7	8	9	10	11	12
States:												
California	-	-	CRT	-	-	CRT	-	CRT	-	-	-	CRT
Florida[a]	-	-	CRT	-	CRT	-	-	CRT	-	CRT	-	-
Georgia[b]	CRT	NRT	CRT	NRT	CRT	-	NRT	CRT	NRT	CRT	-	-
Louisiana	-	-	CRT	NRT	CRT	NRT	CRT	-	NRT	CRT	CRT	-
Oregon	-	-	NRT	-	NRT	-	CRT	-	-	-	-	-
Pennsylvania[a]	-	-	CRT	-	CRT	-	-	CRT	-	-	-	-
Sth Carolina	CRT	CRT	CRT	NRT	NRT	CRT	NRT	CRT	NRT	-	NRT	-
Texas	-	-	CRT	-	CRT	-	CRT	-	CRT	CRT	CRT	-

Note: Weitman et al., 1990; May, 1990; Roeber, 1989

[a] incentive awards only

[b] categorical comparison program plan only

However, none of the states evaluating individual schools assess every grade, regardless of whether the evaluation consists of an incentive-awards program or some type of categorical comparisons. Of those states evaluating individual schools, only three have statewide testing programs which test more than six grades; those states, Georgia, Louisiana, and South Carolina, vary the mode of testing across grades as demonstrated in Table 1 (Weitman, Garber, Oescher, & Brooks, 1990; Roeber 1989).

The data base for states to consistently evaluate schools on achievement across all grades may not exist.

LITERATURE REVIEW

The effective schools research began in the 1970s with the hypothesis by some educators that schools can educate their populations regardless of the background of children that they serve (Edmonds & Freideriksen, 1979). The research movement built momentum in the 1980s as definitions of effective schools and various methods of isolating such schools began to emerge. Emerging with these definitions were improved methods of comparing schoolwide achievement, including school categorization and regression analysis. Generally, research has used regression analysis to control for student background (Lang, 1991) while practice appears to have primarily used categorization techniques (Weitman et al., 1990).

Several comparative studies in the past two decades explored various methods of controlling such background characteristics when attempting to measure school effectiveness. Though measures of pupil background differed, most operational definitions used in those methods contained combinations of socioeconomic (SES) variables and/or previous achievement test scores (Good & Brophy, 1986).

Of those studies, the regression model has demonstrated more success than the others in controlling for pupil background characteristics (Lang, 1991). Conceptually, the regression model appears to be more difficult for the non-technical decision makers to comprehend; but, practically, the categorical model can become cumbersome rather quickly as additional dimensions (variables) are included in the model.

Most of the more recent regression studies employed residuals as measures of school effectiveness. Mandeville and Anderson (1987) termed those measures the "School Effectiveness Indices" (SEIs). The independent variables (IVs) in those studies were generally pupil background variables; the dependent variables (DVs) were school achievement test results.

A present concern of effective school literature is regression model's capability to produce stable effective school classifications (Mandeville & Anderson, 1987; Good & Brophy, 1986; Rowan, Bossert, & Dwyer, 1983). Good and Brophy (1986) noted that there was no tendency demonstrated for schools to be classified as effective (or as ineffective) from year to year. Mandeville and Anderson (1987) conducted such a longitudinal study to determine the stability of the regression model on matched populations followed over several years. Their results also demonstrated a lack of stability, thus supporting Good and Brophy's concerns. In addition, Levine and Lezotte (1990) reached the similar conclusion after reviewing existing research. Of the five longitudinal studies cited in recent school effectiveness reports, all reported problems with stability (Mandeville & Anderson, 1987; Good & Brophy, 1986).

Underlying this issue of stability is another issue--consistency. Consistency is defined in this study as the quality of a regression model to accurately isolate effective and ineffective schools at one point in time. Conceptually, consistency is a necessary, but not a sufficient, condition of stability.

If the regression model is allowed to vary between grades, schools, or studies because of the availability of data, then consistency is threatened. If the resulting classifications are inconsistent, then the stability of such results from one point in time to the next is in jeopardy.

Complications associated with model variations on school effectiveness classifications was noted by Levine and Lezotte in a 1990 monograph on effective schools:

> "Researchers who have carefully examined the data in school effectiveness studies generally have concluded that many schools identified as particularly successful according to a particular measure such as reading scores or sub scores at a particular grade do not stand out as unusually successful with respect to other grade levels, other subject areas, and alternate performance measures (norm-referenced or criterion-referenced) in the same subject or related area." (page 4).

The essence of which Levine and Lezotte (1990) have abstracted from previous research was that wherever the evaluation model was varied, the resulting school-effectiveness ratings or classifications were not consistent. This study suggests that such design-inconsistencies pose major threats to longitudinal stability.

Research on techniques of isolating the effectiveness of individual schools has evolved over a two-decade period since the Dyer, Lynn, and Patton (1969) study had attempted to control for student background variables with the regression model. Within that time frame, researchers conducted numerous studies on effective schools, employing various techniques of which the regression model was most frequently used (Lang, 1991).

Effective school studies historically have not measured instructional performance of an entire school. Where performance was measured across all grades in a given set of schools, those studies generally alternated modes of assessment across those grades. That is, the studies measured instructional performance from available data on grades and subject areas (Purkey & Smith, 1983).

Mandeville and Anderson (1987) faced that particular issue when they conducted a longitudinal study of matched groups across grade levels. Their study was part of a state-mandated school assessment program based on existing data sets. Their data base involved a sample of elementary schools in South Carolina.

The state tested its first graders in the fall with the Cognitive Skills Assessment Battery and in the spring with the state's Basic Skills Assessment Program (BSAP), the state's CRT instruments. In addition, the state tested its second and third graders with the BSAP and its fourth graders with the Comprehensive Tests of Basic Skills, the state's NRT instruments.

Only the second and third grades were predicted and tested with the same mode, that is, state designed CRTs (BSAP). The other grade levels were not. If based only on the analysis of those grades where modes changed, their conclusion of model instability would be confounded by potential influence of using multiple modes of testing. However, their conclusion found support in the comparison of SEIs for grades 2 and 3 where the IVs and DVs did not vary in mode and where the SEIs demonstrated similar instability.

An important issue related to this study is which mode of testing is most appropriate for school effectiveness studies --

NRTs or CRTs. Nitko (1984) explained that NRTs were useful when the needed information pertains to relative ability or relative attainment and that CRTs were useful when the needed information pertains to a repertoire of knowledge and/or skills. Berk (1984) said that a given test could conceptually have provided both NRT- and CRT-based information. However, he warned that it was unlikely that the same test would have provided maximum information along both modes, that is, both relative ability/attainment and knowledge/skill repertoire. He did note that the two modes used together provide a more complete understanding of an individual or school.

Levine and Lezotte (1990) supported the use of both modes of testing. They cited NRTs as perhaps the only available indicator of a school's comparative performance, but noted that the NRTs have the potential to provide a misleading view of local achievement where such instruments do not match curricula. These authors viewed NRTs as assessing a wide array of skills, perhaps beyond what may have been locally emphasized. According to Popham and Husek (1969), the design of the NRT gave preference to variability over content, whereas the design of the CRT gave preference to content over variability.

RESEARCH QUESTION

This study researched the effect that varying DVs had on the consistency of residual-based school effectiveness ratings. That is, the major interest of this study was whether varying the standardized achievement tests had influenced the SEIs and subsequent school effectiveness ratings or classifications.

> **The question this study raised was whether a CRT used as a dependent variable can produce the same results as a NRT in effective school classification.**

For the research question, the initial hypothesis was that some degree of relationship existed between the two methods of evaluating school effectiveness. That is, the relationship between two sets of school classifications was significantly more than what would have been expected by chance alone. If the hypothesis was accepted, then the two school-rating sets were considered somewhat consistent with each other.

More importantly, the second hypothesis was that the magnitude of the relationship between the two sets of school-effectiveness classifications was sufficient enough that schools can be expected to be correctly classified regardless of which DV was selected. That is, does sufficient consistency exist between the two classification sets to warrant the use of NRTs and CRTs interchangeably.

RESEARCH DESIGN

This study analyzed classification consistency in school effectiveness classifications created by varying regression models. The regression models were used to create SEIs while controlling for identified student background variables. Those SEIs were used to classify schools along three effectiveness categories.

All SEIs were computed by standardizing the school residuals in each regression model. The method of standardizing residuals in this study was studentizing the residuals along a t-score distribution. From these studentized residuals, each school was classified as either effective, average, or ineffective. These classifications were used as the basis for the study's subsequent consistency analyses.

To obtain SEIs for each school, the mean score for every school on each achievement test was predicted from the linear relationship of school-aggregated student background variables and school-aggregated test scores across the data set. The predicted mean score was then subtracted from the actual mean score to produce a raw school residual for each test; the raw residual was then studentized, producing the SEI from which a given school was classified.

The SEI represented whether that school had performed higher or lower than expected. If its performance was substantially higher than expected (i.e., a high positive SEI), the school was classified as effective. If its performance was substantially lower than expected (i.e., a low negative SEI), the school was classified as ineffective. If its performance reached neither extreme, then the school was considered average.

Three SES variables were employed as IVs in the regression models. Those SES variables included teacher-reported data on level of parent-education (percent of mothers who were college graduates) and parent-employment (percent of fathers who were white collar workers), and student-reported data on school lunch status (percent of students on paid-lunch status). The SES data had originally been collected in categorical format on the student level during the spring CRT administration. It was aggregated to the school level for this study.

Regression procedures required separate procedures for each DV. In conducting separate procedures, the IVs were held constant across all models in order to determine the effect on consistency while the DVs were manipulated. There were five different DVs used in this study.

The DVs in the study were the school mean scores on the CRTs for language arts and mathematics, and the school mean scores on the NRTs for reading, language, and mathematics. The NRTs and the CRTs chosen were those grade appropriate tests which were administered to public schools throughout the state of Louisiana in the spring of 1989.

The measurement instruments used to compute the DVs were the Level 13, Form E, California Achievement Tests (CAT-13) and the Grade 3 Louisiana Educational Assessment Program tests (LEAP-3). The CAT-13 is an NRT instrument; the LEAP-3 is a CRT instrument.

The CAT-13 had been normed for use with third grade students; the LEAP-3 had been designed to measure third-grade language and mathematics skills as stipulated in the Louisiana curriculum guides for those subjects. The LEAP-3 is a grade-level test, not a minimum skills test (Louisiana Department of Education, 1989).

The LEAP-3 is administered annually to all Louisiana public school students in the third grade as a measure of how well individual students, schools, districts, and the state are addressing the grade-level curricula in language arts and mathematics. The CAT-13 is administered annually in many public school districts in Louisiana as a measure of how well third-grader performances relate to a nationally designed norm. Some school districts restrict the testing of the CAT-13 to partial populations, apparently as an aid in placement into remedial and special education classes, though most districts employing the CAT-13 measure their total population.

Nearly 250 Louisiana elementary schools whose third grade populations were tested with both the NRT and the CRT in 1989 formed the study sample. That sample was taken from a larger sample used in a recent study (Oescher et al., 1989) compiled from scores for third grade students in the state's public schools who had taken both NRT and CRT tests. The unit of analysis Oescher et al. study was the student; the unit of analysis for this study was the school.

The final sample was a reduced one reflecting the removal of inappropriate data for school-level analyses. Such data included the following cases: (1) districts which had not attempted to test their total populations with the NRT, (2) schools whose demographic data were in question, (3) schools which had been poorly matched on CRT and NRT scores, and (4) students who had been absent for the administration of the CRTs and had been assigned a zero score in that data set by default.

The number of schools represented in the final data set were

242, accounting for more than 18,000 students. The percent black was 52.9%, the percent white was 44.4%, and the percent of other ethnicity was 2.7%. The proportions of the final sample in terms of gender were 50.5% male and 49.5% female. With regard to ethnicity, the final sample did not reflect the state's population. The black population was oversampled; the white population was undersampled.

To determine the level of school effectiveness, a classification criteria was established for the studentized residuals: +/-0.674 standard error units (se). Those schools with SEIs beyond than +0.674 se for any DV were classified as "effective" for that DV; those schools with SEIs beyond -0.674 se were classified as "ineffective" for that DV. Those schools with SEIs from +0.674 se to -0.674 se for any DV were classified as "average" for that DV.

The reasoning behind the choice of those points were (1) that the outlier status of beyond +/-0.674 se should have been moderate enough as to have minimized the regression effect on subsequent studies of the same schools, (2) that half of the schools were expected to be classified as average, assuming the SEIs to be normally distributed (Glass & Hopkins, 1984), and (3) that the categorical distributions were similar in size (25%-50%-25%) so as to minimize the influence of chance agreement (Reynolds, 1970).

The design for the consistency analyses of the study's comparisons crossed the results of the mean-based CRT-determined

SEIs with that of the mean-based NRT-determined SEIs in three separate contingency tables:

(1) classifications based on NRT language arts SEIs crossed with those based on CRT language arts SEIs;

(2) classifications based on NRT reading SEIs crossed with those based on CRT language arts SEIs; and

(3) classifications based on NRT mathematics SEIs crossed with those based on CRT mathematics SEIs.

All three contingency tables were 3-by-3 in design for each level of school effectiveness: effective, average, and ineffective. The purpose of the contingency tables was to compare the results of the two classification models.

The comparisons were tested to determine if significant consistency existed. The consistency of the school effectiveness classifications were measured for each issue using the kappa z-test of agreement to determine if varying the DV significantly affected classification decisions. Additionally, magnitude measures of agreement were computed for each comparison to determine the degree of consistency.

The most straight-forward measure of agreement is the unweighted agreement ratio. The unweighted agreement ratio served in this study as a measure of absolute agreement. It is the percent of classifications with which two models concur; it is the sum of the diagonal cells divided by the total units in the analysis. With a possible range from zero to one, the ratio gauges the numerical proportion of identical classifications to the total classifications.

This statistic was employed as the measure of absolute agreement. With regards to this type of magnitude measure, all agreements were absolute, there were no partial agreements; hence, all disagreements were also absolute.

The weighted agreement ratio is a variation in which the elements in off-diagonal cells are weighted inversely as to their degree of disagreement. Regarding a three-level contingency table, neither agreement or disagreement is absolute with the weighted agreement ratio. Perfect agreement cells were weighted with a 1.0, and the perfect disagreement cells were weighted with a 0; the other cells which represent partial disagreement (or agreement) were weighted with 0.5.

A third variation of percent agreement is the kappa coefficient. That statistic controls for chance agreement expected from the distribution of the data. It employs the table's row and column totals (marginals) in determining expected agreement. This study employed a weighted kappa coefficient which was an extension of the weighed agreement ratio. The general range of kappa is +1.0 for perfect agreement to 0 where the agreement ratio equals expected chance agreement. Kappa values are negative where the agreement ratio is less than what is expected by chance (Reynolds, 1977).

For significance testing, the kappa $\underline{z}$-statistic was chosen as the measure of consistency because it was not as sensitive to the sample size as measures of association and because it controlled for chance consistency. A significant $\underline{z}$-test means that the two classification distributions demonstrate some agreement; an insignificant test means that the two distributions are independent of one another--there is no significant agreement beyond what would be expected by chance. The _z_-statistic is computed by dividing the kappa coefficient by its standard deviation (Reynolds, 1977).

FINDINGS

The crossing of NRT and CRT modes demonstrated significant agreement along effective school classifications for all three pair-wise results considered. However, the degrees of magnitude as measured by the kappa coefficient and the weighted agreement ratio were somewhat limited for the classification comparisons in this study.

The NRT reading and CRT language arts classification comparison produced the most consistent results (kappa=.626). Alternating the NRT reading test with the CRT language arts test produced consistent results in approximately five of every eight schools evaluated by both instruments (See Tables 2, 3, & 4).

The other two comparisons demonstrated consistent results in slightly more than a one of two cases (controlling for expected agreement). The kappa coefficients for both of the other comparisons follow: NRT language and CRT language arts, .541; and NRT mathematics and CRT mathematics, .560.

Table 2
Contingency Table Comparison of School Classifications
by NRT Language SEIs & CRT Language Arts SEIs

CRT-Based Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
NRT-Based Results:								
Ineffective	44	(18.2)	17	(7.0)	0	(0.0)	61	(25.2)
Average	19	(7.9)	80	(33.1)	22	(9.1)	121	(50.0)
Effective	0	(0.0)	25	(10.3)	35	(14.5)	60	(24.8)
Column Total	63	(26.0)	122	(50.4)	57	(23.6)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.541	.526	.372
Kappa Z-Statistic	2.06***	1.51	1.13
Unwghted Agreement Ratio	.657	----	----
Weighted Agreement Ratio	.829	.775	.710

* probability < .001
** probability < .01
*** probability < .05

With regard to raw agreement, weighted agreement ratios ranged from .829 to .862, depending on the comparison. However, chance agreement was not controlled with the weighted agreement ratio. Regardless, a considerable amount of discrepancy existed between the two classification models (See Tables 2, 3, & 4).

This study was an exploration into the effect that varying the dependent variables had on consistency. Tests of significance and weighted magnitude measures tell the scientist much about that effect. However, to the decision maker who must ultimately select a school classification model from an imperfect world of data, what is important is the proportion of schools that will be classified consistently by a given a model.

Table 3
Contingency Table Comparison of School Classifications
by NRT Reading SEIs & CRT Language Arts SEIs

CRT-Based Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
NRT-Based Results:								
Ineffective	44	(18.2)	15	(6.2)	1	(0.4)	60	(24.8)
Average	19	(7.9)	91	(37.6)	15	(6.2)	125	(51.7)
Effective	0	(0.0)	16	(6.6)	41	(16.9)	57	(23.5)
Column Total	63	(26.1)	122	(50.4)	57	(23.5)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.626	.564	.580
Kappa Z-Statistic	2.33^{**}	1.63	1.70^{***}
Unwghted Agreement Ratio	.727	----	----
Weighted Agreement Ratio	.862	.799	.810

* probability < .001
** probability < .01
*** probability < .05

A measure of absolute consistency provides that information. It is the unweighted agreement ratio without any controls for expected agreement. For this study, the measure of absolute agreement has provided the proportion of schools consistently classified and the data from which to compute the proportion of schools inconsistently classified.

The range for the unweighted agreement ration was .657 to .727. That range means that more than 1 out of every 4 schools were inconsistently classified where CRTs and NRTs were alternated. The important question for decision makers is whether such a level of inconsistency is too great to tolerate alternating modes of testing.

Table 4

Contingency Table Comparison of School Classifications by NRT Mathematics SEIs & CRT Mathematics SEIs

CRT-Based Results:	Ineffective		Average		Effective		Row Total	
	n	%	n	%	n	%	n	%
NRT-Based Results:								
Ineffective	42	(17.4)	17	(7.0)	1	(0.4)	60	(24.8)
Average	15	(6.2)	89	(36.8)	22	(9.1)	126	(52.1)
Effective	2	(0.8)	18	(7.4)	36	(14.9)	56	(23.1)
Column Total	59	(2′.4)	124	(51.2)	59	(24.4)	242	(100.0)

Statistical Results

Comparison Levels:	Effective Average Ineffective	Average Ineffective	Average Effective
Statistics:			
Kappa Coefficient	.560	.572	.460
Kappa Z-Statistic	2.11***	1.72***	1.34
Unwghted Agreement Ratio	.690	----	----
Weighted Agreement Ratio	.839	.804	.758

* probability < .001
** probability < .01
*** probability < .05

CONCLUSIONS

Though utilization of both NRT and CRT modes of testing in this study resulted in significant agreement along effective school classifications for all three pair-wise comparisons, the degree of magnitude was somewhat limited for every classification comparison. Such results can hardly justify alternating CRT and NRT instruments for that third-grade population for any test combination except perhaps that of the NRT reading and the CRT language arts instruments. The two modes of testing are measuring schools effectiveness differently in too many cases.

Though this study researched the effect that alternating test modes within a grade has on school evaluations, the results also raised concerns about the effect that alternating of modes of testing across grades has had in previous studies or programs. Such was the unavoidable design problem wherever existing data sets have been employed, as in the case of the Mandeville and Anderson (1987) longitudinal study.

Generally, one can expect a certain degree of instability when different tests are employed, whether they be cross-mode or cross-grade. However, employing another test which is varied on two fronts (i.e., both different mode and different grade) increases the potential for instability.

If substantial inconsistency exists within a given grade, how much more inconsistency exists across grades? Moreover, what effect on longitudinal stability does alternating modes of testing have on school evaluations?

This study concludes that cross-mode instruments should not be employed on an alternating basis in evaluating school effectiveness. The consistency coefficients in this study are not of sufficient magnitudes to support alternating test modes across grades. Without sufficient consistency, the stability of any longitudinal studies or evaluations will be in question.

> **The data set for states to consistently evaluate schools on achievement across all grades does not exist.**

IMPLICATIONS

Statewide school evaluation programs need to do more than employ existing testing programs if their results are not to be challenged on consistency and stability grounds. Those programs may need to affect an expansion to their states' existing testing programs to guarantee consistency and to increase stability in school evaluations.

Expansion of an existing testing program raises another issue--the NRT versus CRT dilemma. Levine and Lezotte (1990) recommended employing both instruments. They suggested that the NRTs are the only indicators of comparative school performance and that the CRTs are the best indicators of curricula performance. Berk (1984) noted that the two modes used together provide a more complete understanding of an individual or school.

Though employing both modes has support in literature, available finances may dictate otherwise. Furthermore, time normally allocated to classroom instruction may not be available for additional testing of each grade. The cost-benefit ratio should be considered before selecting a dual-mode testing program for each grade.

Regarding the use of regression model for evaluating schools, the NRT appears to be a more suitable instrument both in terms of design and expense. The regression model is a relative model; that is, school effectiveness classifications are relative to the performance of all schools with regard to whatever control factors are employed as IVs. As Popham and Husek (1969) noted, the design of NRT instruments favors relative performance, not absolute performance. The two authors noted that NRTs are designed to enhance variability which augments an instruments' capability to discriminate.

For this study, that enhanced variability in student test scores also increased variability in school-aggregated test scores and in their resulting residuals when SES was controlled. Table 5 provides the standard deviation (i.e., standard errors) and the variance (i.e., mean square error) for the raw residuals.

Table 5

Variance Found in Raw Residuals for the Study's Regression Models

Dependent Variables	Std. Dev.	Variance
CRT Math. Mean	3.91	15.29
CRT Lang. Mean	4.05	16.42
NRT Math. Mean	17.01	289.22
NRT Lang. Mean	14.34	205.53
NRT Read. Mean	17.85	318.50

With the regression model, there is no absolute criterion of effectiveness as there is with CRT instruments. Hence, the employment of a CRT instrument with the regression model bypasses the intent and design of CRTs as measures of performances on absolute criteria. The CRT instruments are generally not designed to maximize differences in relative performances as demonstrated in Table 5. Instead, they are primarily designed around content issues (Popham & Husek, 1969).

On the other hand, the weak point of commercial NRT instruments is that whatever curriculum match exists beyond what would be expected from a generic test is happenstance. Rowan et al. (1983) suggested that the underlying cause of instability in longitudinal school evaluation programs may have been caused by using tests which do not match the curricula. CRTs, however, are designed either to reflect curricula, as is the case of the Louisiana testing program, or to reflect a set of minimum curriculum skills, as was the case of the South Carolina testing program (May, 1990) employed in the Mandeville and Anderson (1987) study.

Nevertheless, where regression models are employed to classify schools along levels of effectiveness, the instruments most appropriately designed to measure relative performance among schools appear to be NRTs. Since they are designed to maximize differences in relative performances, NRT instruments are conceptually better fitting instruments for the regression model than are their CRT counterparts. In addition, the NRT instrument provides the least expensive solution to testing every grade in the same mode. Which NRT test to employ, however, is an issue of best curriculum match.

Conceptually, the employment of CRTs in evaluating schools would best be implemented if a non-relative model were used to control for whatever background variables the evaluators deem important. The criteria for effective and ineffective classifications should be absolute for a given school if optimal use of a CRT is to be realized. Until CRTs are developed for each grade and until an absolute school evaluation model is developed, grade appropriate NRT instruments appear to be the most appropriate tools in evaluating schools where acceptable curricula match can be found, particularly where the regression model is employed.

REFERENCES

Berk, R.A. (1984). Selecting the index of reliability. In R.A. Berk (ED) A Guide to Criterion-Referenced Test Construction. Ed. R.A. Berk, Baltimore: John Hopkins University Press.

Dyer, H.S., Linn, R.L., Patton, M.J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6(4), 591-605.

Edmonds, R.R., Frederiksen, J.R. (1979). Search for effective schools: The identification and analysis of city schools that are instructionally effective for poor children. ED 170396.

Glass, G.V., & Hopkins, K.D. (1984). Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall.

Good, T.L., & Brophy, J.E. (1986). School effects. Third Handbook of Research on Teaching, Ed. M. Whitrock, New York: Macmillan.

Lang, M.H. (1991). Effective School Status: A Methodological Study of Classification Consistency. Dissertation, Louisiana State University, Baton Rouge.

Levine, D.U., & Lezotte, L.W. (1990). Unusually Effective Schools. Austin: The National Center for Effective Schools Research & Development.

Louisiana Department of Education (1989). Louisiana Educational Assessment Program Annual Program Report, 1988-89 School Year. Baton Rouge.

Mandeville, G.K., & Anderson, L.W. (1987). The stability of school effectiveness indices across grade levels and subject areas. Journal of Educational Measurement, 24(3), 203-216.

Mandeville, G.K., & Heidari, K. (1988). Measuring school effectiveness using hierarchical linear models. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans. ED303477.

May, John (April, 1990) Real world considerations in the development of an effective school incentive program. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Mitko, A.J. (1984). Defining "Criterion-referenced Test." In R.A. Berk (ED) A Guide to Criterion-Referenced Test Construction. Ed. R.A. Berk, Baltimore: John Hopkins University Press.

Oescher, J., Paradise, L.V., & Kirby, P.C. (1989). Comparison study of Grade 3 norm-referenced and criterion-referenced test results. Report prepared for the Louisiana Department of Education, Baton Rouge.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement 6(1), 1-9.

Purkey, S.C., & Smith, M.S. (1983). Effective schools: A review. The Elementary School Journal, 83(4), 427-452.

Reynolds, H.T. (1977). The Analysis of Cross Classifications. New York: Free Press.

Roeber, E.D. (1989). Survey of large-scale assessment programs. A annual survey for the Association of State Assessment Programs, Lansing.

Rowan, B., Bossert, S.T., & Dwyer, D.C. (1983). Research on effective schools: a cautionary note. Educational Researcher, 12 (4), 24-31.

Weitman, C.J., Garber, D.H, Oescher, J., Brooks, C. (1990). Louisiana School Incentive Program: Policies and Recommendations 1990. Report for the Louisiana Department of Education. Louisiana Tech University.

THE EFFECT THAT VARYING THE TEST MODE HAD ON SCHOOL EFFECTIVENESS RATINGS

by Michael H. Lang
Charles Teddlie
Jeffery Oescher

Presented at the April 1992 AERA Annual Meeting in San Francisco.

ABSTRACT

This study investigated both the effect of alternating criterion-referenced tests (CRT) with norm-referenced tests (NRT) in evaluating schools.

The sample included 242 Louisiana public elementary schools (18,000 third graders tested in 1989). The study employed five separate multiple regression models, each producing studentized residuals used as school effectiveness indicators (SEIs). The independent variables for all models were student's free lunch status, mother's educational level, and father's employment level. The dependent variables were school mean scores for CRT language arts and mathematics tests, and NRT reading, language, and mathematics tests.

The study used SEIs to classify schools as effective, average, or ineffective. It classified each school according to ten different models using +/-.674 se as the criteria.

The study separately analyzed appropriate cross classification results: (1) CRT language arts & NRT language, (2) CRT language arts & NRT reading, and (3) CRT mathematics & NRT mathematics.

The study tested each comparison with the kappa z-test; it measured agreement with the weighted kappa coefficient (chance-controlled agreement), the weighted agreement ratio (adjusted agreement), and the unweighted agreement ratio (absolute agreement).

The study found the kappa-z tests significant beyond the .05 level. It found that magnitude measures were generally moderately consistent for CRT-NRT comparisons. The study concludes that findings do not support alternating tests modes in evaluating schools.