DOCUMENT RESUME

ED 229 426 TM 830 327

AUTHOR Frechtling, Joy A.
TITLE The Measurement of Effectiveness: Some Methodological Problems.
PUB DATE Oct 82
NOTE 19p.; Paper presented at the Annual Meeting of the Evaluation Network and Evaluation Research Society (Baltimore, MD, October 28-30, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Data Analysis; Elementary Secondary Education; Evaluation Methods; *Evaluation Needs; Grade 5; Longitudinal Studies; Measurement Objectives; *Measurement Techniques; Norm Referenced Tests; Program Improvement; Reading Comprehension; *Research Methodology; *School Effectiveness; Scores; Trend Analysis
IDENTIFIERS Experts; Iowa Tests of Basic Skills; *Measurement Problems; Ranking; Residual Scores

ABSTRACT

This paper highlights measurement issues faced when attempting to assess and interpret results of a school improvement project. Based on the assumption that to measure effectiveness, one must measure a wide variety of school factors, the paper presents a broad perspective on measurement problems and dilemmas in analyzing norm-referenced test data and data obtained through interview, self appraisal, and observation concerning 117 elementary schools. Trend analysis, two forms of residual gains analyses, traditional ranking, and expert judgment methods are compared. Data suggest that school level residual analysis appears to provide the best approach to selecting schools. The individual level residual scores yield a list which overlaps with the school level approach. Trend analysis is the most conservative and yields the fewest schools (which are also identified by residual score analyses). Expert opinion does not correlate positively with residual or trend analyses. Analyses indicated few consistencies over time. The authors conclude with two alternatives--either schools are not consistent in their impacts from year to year or their metric is suspect. (Author/CM)

THE MEASUREMENT OF EFFECTIVENESS
Some Methodological Problems

Dr. Joy A Frechtling, Director
Division of Instructional Evaluation and Testing
MONTGOMERY COUNTY PUBLIC SCHOOLS

Presented at Evaluation '82, Baltimore
October 28-30, 1982

INTRODUCTION

I do not think that I would be overstating the case if I said that research on effective schools is the current raging fad. There is a belief that we have made a breakthrough in determining what makes for an effective school and even a hope that we have found the key for making less effective schools into more effective ones. Although most of the studies on school effectiveness have been conducted in urban, low-scoring schools, higher achieving schools and school districts could hardly be considered disinterested; and concern with assessing effectiveness is evident in every school district and state department of education.

In Montgomery County, Maryland, where Jim and I work, there is also considerable interest in measuring school effectiveness. This interest is, however, mixed with a good deal of concern about whether or not we really know how to measure effectiveness and whether the traditional tools used really tell us what we think they do. Typically, educators have used test scores to decide whether or not a school is effective. For a long time, it was believed that the higher the test score, the more effective the school. I think we've made some progress in moving away from this simplistic and probably erroneous indicator of effectiveness and realize that a single test score tells us as much or more about a student's background, as the effectiveness of the instruction he or she has received. We have developed a number of alternatives to this approach controlling in many ways for "extra-school" factors. But, at this point, we can say only that the alternatives we have come up with are more sophisticated, but not necessarily more accurate.

I take this pessimistic viewpoint because of some work that we have been doing in Montgomery County on methodological issues in determining school effectiveness comparing selected "effective" schools. Last year, several members of my staff rated schools using different analytical approaches. We then compared their results looking for convergence and divergence. In a second analysis we looked more closely at the extent to which the same schools appeared to be effective or ineffective over time. Today, I will briefly discuss the results of each of these analyses.

COMPARISONS AMONG METHODS

First, I will discuss the comparison of methods--five methods were examined; trend analysis, two forms of residual gains analyses, traditional ranking, and expert judgment. Jim Myerberg has already described one of the methods; trend analysis (performance of a matched longitudinal sample using the 8 NCE criterion).

The second and third used residual gain scores, with the unit of analysis being either the individual student or student data aggregated to the school level. The fourth approach, what I will label as "traditional ranking" involved ranking schools according to fifth grade test scores. Finally, a form of "expert judgment" was used in which reading specialists were asked to assess the degree to which schools were effective or ineffective in teaching reading.

In comparing the results of these different approaches we tried to keep as many things besides method as constant as possible. For example, we made sure that we used the same subscore on the Iowa Tests of Basic Skills as our indicator of effectiveness. This was the Reading Comprehension score. We

also made sure we were looking at test scores from the same cohort of students. While these controls sound so obvious as to be trivial, comparative studies in the past have not always taken these precautions either through oversight or because for some reason it has proven impossible. We also used the same criterion for determining outliers--those which we were going to call "effective: and "ineffective." That is, we standardized the test scores and took as our outliers those schools with a "Z" score of ± 1.38. This criterion was admittedly somewhat arbitrary. It was selected because it gave us what might be considered a "face valid" number of outliers--about 10 percent of our elementary schools at either tail. We might discuss sometime whether it was, in fact, an appropriate choice. At least we can say it was consistent.

Where expert opinion was used, we tried to exert some control by asking our experts to focus on effectiveness in the area of reading. Other aspects of this method clearly differed, however, from the other four and no attempt was made to control for them. For example, it is likely that the expert judgments took into account the overall performance of the school in reading (rather than focusing on 5th grade performance) and included an assessment of the schools' performance over more than a single year.

Effective Schools

Exhibit 1 shows the schools selected as "effective" by each of the methods employed. This exhibit shows:

- Overall 47 of the 117 (40%) elementary schools examined were nominated by one or more of the methods
- Only 11 (9%) were nominated by more than one method
- The methods differed widely in the number of schools identified as "effective", from a low of "3" for the trend analysis to a high of 27 for the Expert method.

Further analyses examined the correlations between the individual methods of selecting effective schools (Phi). In calculating these correlations we decided to treat effectiveness as a dichotomous rather than a continuous variable. That is, we looked only at the degree to which the alternative methods resulted in the nomination of the same schools as effective. Clearly, a viable alternative would have been to consider the entire continuum of schools. Exhibit 2 presents the findings.

Exhibit 2 shows that all the methods are significantly correlated except the cross-sectional ranking method and individual level residual score analysis. The strongest positive relationship (.38) was found between the trend analysis and the individual level residual analysis. The strongest negative relationship (-.49) was found between expert judgment and the individual residual score analysis. Expert judgment was, however, significantly negatively related to all three of the indices considered. We were somewhat surprised to note the low correlations between the cross-sectional method and the residual score analyses. In previous studies, this relationship was found to be stronger.

Exhibit 1

Schools Selected as Effective Using Each of the Five Methods
N=117

School #	Trend Analysis	Expert Opinions	School Level Residuals	Cross-Sectional	Individual Level Residuals	Total
2			X		X	2
4		X				1
7		X				1
8		X				1
10					X	1
11				X		1
12	X		X		X	3
14				X		1
16		X			X	2
17		X		X	X	3
19					X	1
20		X				1
26		X	X		X	3
27		X				1
28		X				1
30		X				1
33					X	1
34		X				1
41		X				1
43					X	1
45		X				1
47					X	1
48				X		1
54		X				1
60					X	1
61		X				1
62			X			1
63		X				1
68					X	1
69			X	X	X	3
77	X	X	X	X	X	5
78		X				1
81		X				1
86	X				X	2
91					X	1
98		X				1
100		X				1
103				X		1
105		X			X	2
106					X	1
113		X	X		X	3
116					X	1
119		X		X	X	3
122		X				1
123		X				1
124		X				1
127					X	1
Totals	3	27	7	8	22	47

Exhibit 2

Correlations Between the Alternative Methods for Selecting Effective Schools

Method[1]	A	B	C	D	E
A		-.13*	+.38**	+.11*	+.28**
B			-.12*	-.18**	-.49**
C				+.13*	+.33**
D					+.03
E					

*P .05
**P .01

[1] A = Trend Analysis
B = Expert Opinion
C = School Level Residual Scores
D = Cross Sectional
E = Individual Level Residual Scores

Finally, we looked at the degree to which each of the approaches correlated with a composite effectiveness score. The latter was determined by summing the number of nominations and dividing the schools into two groups: those nominated once (N=26) and those nominated twice or more (N=11). Exhibit 3 presents the results of the Phi analysis.

The data suggest that the strongest correlation is between the school level residual analysis and the composite score (.62). Nearly as strong was the relationship between the composite score and individual residual score (.59). The trend analysis and composite score also showed a strong relationship (.47). The other two methods showed considerable weaker correlations, although they remained significant.

Exhibit 3

Correlation Between Each Method and The Composite Effectiveness Score[1]

	Composite Effectiveness Score
A[2]	+.47**
B	-.07
C	+.62**
D	+.28**
E	+.59*

**P .01

[1]The outcome measure "Composite Effectiveness Score" is defined by the total number of nominations received by each school. We have chosen to divide the nominal schools into 2 categories: those receiving 1 nomination and those receiving 2 or more.

[2]A = Trend Analysis
B = Expert Opinion
C = School Level Residual Scores
D = Cross Sectional
E = Individual Level Residual Scores

Ineffective Schools

Exhibit 4 presents the findings for schools judged to be "ineffective." It should be noted that expert opinion was not used as a means for selecting ineffective schools because it was deemed preferable not to ask the specialists to single out schools as being particularly ineffective in teaching reading. The exhibit shows:

- Overall 30 of the 117 (26%) elementary schools were nominated by one or more of the methods.
- Only 7 (6%) were nominated by more than one method.
- The methods differed in the number of schools identified as ineffective. The trend analysis yielded only one school. The individual residual score method yielded the largest number - 23.

It is interesting to note that there are four schools which appear on both lists -- the effective schools and the ineffective schools. In all four cases, the schools had been nominated as effective by expert judgment.

Correlations among the measures are presented in Exhibit 5. This matrix is somewhat spotty because of the absence of expert opinion and the elimination of the trend analysis, since it yielded only one case. It is, however, worth pointing out a couple of findings. The school level residual analysis appears to be uncorrelated with either the cross-sectional or individual level residual analysis--a finding which is not consistent with relationships among the methods for selecting of "effective schools". Interestingly, the cross-sectional and individual residual analyses are strongly negatively correlated.

Exhibit 6 presents the correlation between each of the measures and the composite ineffective scores. As with the analysis of effective schools, the highest correlation is between the composite score and the school level residual analysis. The cross sectional method is unrelated to the composite score.

Exhibit 4

Schools Selected as Ineffective Using Each of the Five Methods
N=117

School #	Trend Analysis	Expert Opinions	School Level Residuals	Cross-Sectional	Individual Level Residuals	Total
3					X	1
12				X		1
15			X	X	X	3
22	X				X	2
29				X		1
30				X	X	2
35					X	1
42			X			1
51			X			1
57				X		1
61					X	1
63					X	1
65					X	1
67			X	X	X	3
70				X		1
72					X	1
75					X	1
79			X		X	2
80					X	1
81					X	1
84				X	X	2
85				X		1
89					X	1
93			X		X	2
108					X	1
110					X	1
114					X	1
124					X	1
126					X	1
115					X	1
Totals	7	0	6	9	23	30

Exhibit 5

Correlation Between the Alternative Methods of Selecting Ineffective Schools
N=30

Method[1]	A	B	C	D	E
A[2]					
B					
C				.04	-.12
D					-.50**
E					

**P .01

[1] A = Trend Analysis
B = Expert Opinion
C = School Level Residual Scores
D = Cross Sectional
E = Individual Level Residual Scores

[2] Since the trend analysis yielded only one case we have eliminated it.

Exhibit 6

Correlations Between Each Method and The Composite Ineffectiveness Score[1]

	Composite Ineffectiveness Score
A[2]	
B	
C	+.51**
D	+.03
E	+.30**

**p .01

[1]This score is derived by summing the number of nominations received for each school. The resulting group is then divided into two groups: those receiving 1 nomination and those receiving 2 or more.

[2]A = Trend Analysis
B = Expert Opinion
C = School Level Residual Scores
D = Cross Sectional
E = Individual Level Residual Scores

Conclusions

Of the methods examined here the school level residual analysis appears to provide the best approach, to selecting schools. The individual level residual score yields a list which to a large extent overlaps with that produced by the school level approach but also contains a number of others. It has been suggested that this is because it fails to account adequately for error variance in the individual scores.

The trend analysis is the most conservative, yielding the fewest schools. Nonetheless the schools identified by this method are also identified by the residual score analyses. It may well be that it is only a matter of criterion that separates the trend and school level approaches from converging more completely. We explored this by doing a supplementary analysis where, instead of the 8 NCE criterion, we used the 1.38 Z criterion that had been employed in the residual gains methods - using this standard the trend analysis approach naturally yielded more schools. Two of these, however, were not identified by the residual score analysis and may be cases of regression to the mean. We need to further explore this method, and the strengths and weaknesses of using it with different criteria for school selection. While it may be lacking in a certain amount of elegance, the approach has a considerable amount of appeal because it is so easy to understand and apply.

The role of the other two analyses appears to be far less clear. The data certainly suggest that the cross sectional approach can be misleading. And, this study supports others which have suggested that it is better to avoid making judgments regarding school effectiveness on such data.

The usefulness of expert opinion remains a question. As this study shows, expert opinion does not correlate positively with the residual or trend analyses explored here. On the other hand, as we shall see in the next section, these methods do not correlate well with each other from year to year--a rather disturbing finding for those trying to identify effective schools! The possibility should not, therefore, be dismissed that expert opinion is providing useful information and its lack of correlation with the other methods does not necessarily mean that expert opinion can or should be dismissed.

CONSISTENCY ACROSS YEARS

The second series of analyses looked at consistency across years. If a school really is effective (or ineffective), it would be expected that analyses would show the school to be an outlier with some consistency. We, therefore, looked at consistency over time using both the NCE trend analysis and the residual gains methods. Exhibit 7 shows the findings from the trend analysis, considering cohorts. The black boxes show scores that are high, the shaded boxes those that are low. Few consistencies over time are found. Further, when correlations were computed by John Larson of my staff for two cohorts using residual gains analyses, the findings were similar. Across the two cohorts, the correlation is only .24 for reading scores and .32 for mathematics. If we consider these data in terms of variance explained, we come up with a whooping four to nine percent.

We are left with two alternatives--we can conclude either that schools are not consistent in their impacts from year to year or that our metric is suspect, if not faulty. Intuitively, we have to suspect the metric. Unfortunately, we

EXHIBIT 7

Schools With Substantial Longitudinal Trends in Each of the Last Four Years - First Quarter

1978-79					1979-80					1980-81					1981-82				
No.	RC	TL	TM	C	No.	RC	TL	TM	C	No.	RC	TL	TM	TB	No.	RC	TL	TM	TB
19			▨	▨	15	▨		▨	▨	23		▨			26				
29					48					33					29			▨	
37					53					40					36				
22					35					39					40				
56					50			■		46					45				
31					39	▨				33	▨				26				
22					55					53					50				
38					42			▨		20					25				
44					45					58					67				
27		■			55					48					44				
52					47					51			■		52				
54					50					39	▨				38				
22			▨		21			▨		24			▨	▨	17				
42					40					67					52				
41		▨	▨	▨	40			▨	▨	37			▨	▨	25	▨	▨	▨	▨
55					60					67			■		61				
44		■			36			■		30					23				
25					30			▨	▨	48			▨	▨	27			▨	
42					42					35					21	▨			
68					54		▨		▨	78					65				
12					13					17					10				
45					58					55					54				
35					29					31					35				
34					36					27			▨		34				
22					35					46					43			▨	
32			■		42					35					31				
37		▨			58					58					37				
46					70					65			▨		50			▨	
27					46		■			38					42				
43					41					56					65				
53					44					54					59				

■ - School longitudinal trend was at least 8 NCE points higher than the county trend.

▨ - School longitudinal trend was at least 8 NCE points lower than the county trend.

No. - Number Tested
TL - Total Language Composite
RC - Reading Comprehension
TM - Total Math
TB - Total Battery

EXHIBIT 7 (Continued)
Schools With Substantial Longitudinal Trends in
Each of the Last Four Years - Second Quarter

1978-79					1979-80					1980-81					1981-82				
No.	RC	TL	TM	C	No.	RC	TL	TM	C	No.	RC	TL	TM	TB	No.	RC	TL	TM	TB
–					16					17					13				
52					73					61					58				
38					24					40					18				
59					45					70					56				
27					30					33					29				
54					56					50					46				
62					91					65					58				
62					74					80					80				
31					65					66					58				
53					52					61					57				
56					47					57					49				
25					40					32					28				
33					47					71					51				
33					34					35					19				
17					50					49					43				
39					66					65					59				
15					23					18					22				
60					65					79					72				
28					35					43					23				
44					56					50					50				
30					36					34					36				
42					64					56					43				
31					15					25					26				
45					48					31					30				
57					61					55					52				

■ - School longitudinal trend was at least 8 NCE points higher than the county trend.
▨ - School longitudinal trend was at least 8 NCE points lower than the county trend.

No. - Number Tested
TL - Total Language
C - Composite

RC - Reading Comprehension
TM - Total Math
TB - Total Battery

EXHIBIT 7 (Continued)
Schools With Substantial Longitudinal Trends in
Each of the Last Four Years - Third Quarter

1978-79					1979-80					1980-81					1981-82				
No.	RC	TL	TM	C	No.	RC	TL	TM	C	No.	RC	TL	TM	TB	No.	RC	TL	TM	TB
15					15					12					-				
59					52					41					54				
48					31					51					53				
33					43					50					49				
35					20					22					33				
66					48					57					46				
49					52					72					52				
32					47					42					28				
28					49					29					27				
75					69					71					65				
15					31					23					36				
68					60					61					35				
38					26					21					34				
37					42					45					35				
23					26					29					17				
74					85					87					84				
40					52					42					38				
57					63					74					60				
39					49					50					49				
29					18					29					20				
22					45					45					41				
27					17					26					21				
-					-					23					35				
63					69					87					66				
42					42					52					30				
45					45					49					38				
33					38					43					40				
38					36					40					26				

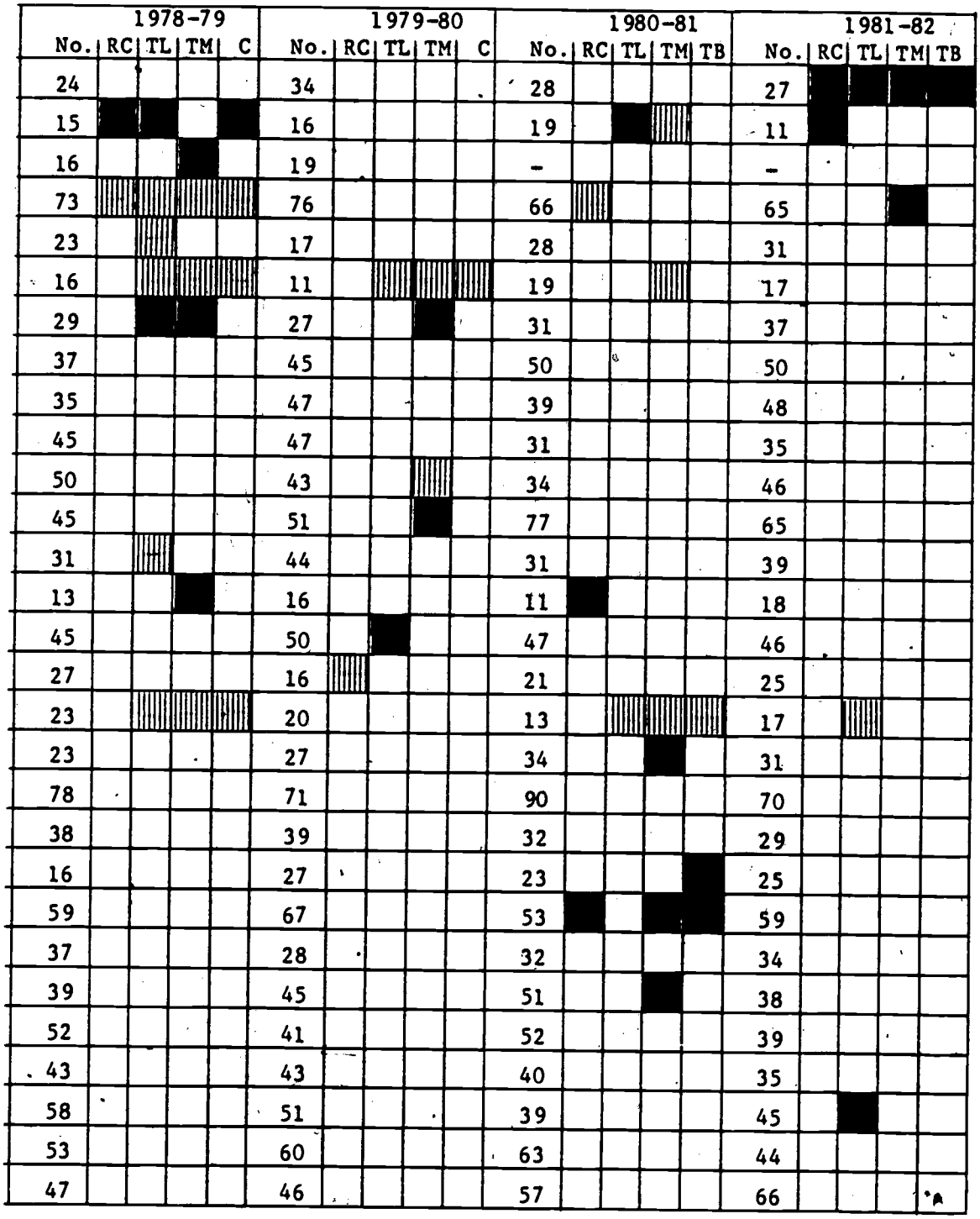
- School longitudinal trend was at least 8 NCE points higher than the county trend.
- School longitudinal trend was at least 8 NCE points lower than the county trend.

No. - Number Tested
TL - Total Language
C - Composite

RC - Reading Comprehension
TM - Total Math
TB - Total Battery

EXHIBIT 7 (Continued)

Schools With Substantial Longitudinal Trends in Each of the Last Four Years - Fourth Quarter



1978-79					1979-80					1980-81					1981-82				
No.	RC	TL	TM	C	No.	RC	TL	TM	C	No.	RC	TL	TM	TB	No.	RC	TL	TM	TB
24					34					28					27				
15					16					19					11				
16					19					-					-				
73					76					66					65				
23					17					28					31				
16					11					19					17				
29					27					31					37				
37					45					50					50				
35					47					39					48				
45					47					31					35				
50					43					34					46				
45					51					77					65				
31					44					31					39				
13					16					11					18				
45					50					47					46				
27					16					21					25				
23					20					13					17				
23					27					34					31				
78					71					90					70				
38					39					32					29				
16					27					23					25				
59					67					53					59				
37					28					32					34				
39					45					51					38				
52					41					52					39				
43					43					40					35				
58					51					39					45				
53					60					63					44				
47					46					57					66				

- School longitudinal trend was at least 8 NCE points higher than the county trend.
- School longitudinal trend was at least 8 NCE points lower than the county trend.

No. - Number Tested
TL - Total Language
C - Composite

RC - Reading Comprehension
TM - Total Math
TB - Total Battery

have not found a more satisfactory substitute. In any case, these findings are of great concern to us in our jobs and, in addition, make us regard existing research on school effectiveness with some degree of skepticism.

520b/3

TM 830 327

ABSTRACT

THE MEASUREMENT OF EFFECTIVENESS
Some Methodological Problems

By Dr. Joy A. Frechtling

This paper highlights measurement issues faced when attempting to assess and interpret results of a school improvement project. Based on the assumption that to measure effectiveness, one must measure a wide variety of school factors, the paper presents a broad perspective on measurement problems and dilemmas in analyzing norm-referenced test data and data obtained through interview, self appraisal, and observation

The central question is the following: How does one move from correlation analysis to assessment of change, when measuring instruments as well as methodology are far from satisfactory?

705b/75