DOCUMENT RESUME

ED 300 427 TM 012 430

AUTHOR Ligon, Glynn
TITLE A Light Look at Some Heavy Issues in Public School Evaluation.
INSTITUTION Austin Independent School District, Tex. Office of Research and Evaluation.
REPORT NO ASID-87.27
PUB DATE Apr 88
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communication Problems; Elementary Secondary Education; *Evaluation Problems; Evaluation Utilization; Evaluators; *Program Evaluation; *Public Schools; Research Reports; Technical Writing; User Satisfaction (Information)
IDENTIFIERS *Evaluation Reports

ABSTRACT

Examples of the difficulties that public school program evaluators can have in explaning perplexing results are provided. Arguments abound over the best way to report a finding that has multiple interpretations or presentation options, and the enigma of realizing that one way to report a finding is not the only way, or even necessarily the best way, can be the bane of the evaluator. Twenty examples cover apparent paradoxes and contradictions in the reporting of evaluation findings. These are generally problems of wording, problems that often have seemed too frivolous for textbook consideration. Evaluation findings are seldom simple and straightforward. Reporting them requires a practical communication style that is focused on the intended audience. The key to describing an evaluation question is understanding and addressing the question that is being asked. As obvious as this seems, some reports never directly answer the question that was to be addressed. (SLD)

A Light Look at Some Heavy Issues in Public School Evaluation

Glynn Ligon
Austin Independent School District
Austin, Texas

Paper Presented at the Annual Meeting
of the American Educational Research Association

New Orleans, Louisiana April, 1988

Publication Number 87.27

87.27

A Light Look at Some Heavy Issues in Public School Evaluation

Glynn Ligon, Austin (TX) Public Schools

Evaluating public school programs is serious business. However, within the maze of traditional inferrential, Bayesian, Rasch, parametric, nonparametric, and descriptive statistics, there are some technical issues that can be viewed from a lighter perspective. The purpose of this paper is two fold. First, we evaluators need a chance to share our favorite stories about the difficulties we have had trying to explain some very perplexing findings;or about the second thoughts we have had about the way we reported a finding. We can seldom joke in the cafeteria about Simpson's paradox or the insanity of a national goal to have every student achieve above the 50th percentile without having to spend the rest of our lunch hour explaining normal distributions to a bunch of central office administrators who would confuse an F test with a grading standard, or a p value with a measure of a type A lunch.

In fact it is time for us evaluators to step back from our self-proclaimed state-of-the-art methodologies and to remind ourselves that, with all the sophisticated computer analyses we are capable of, there are some phenomena that haunt our reporting. Some results just seem to defy our best efforts to explain them straightforwardly to our audiences. Other results afford the evaluator wide authority to portray a finding truthfully as either positive or negative.

This paper is drawn from the author's 15 years of experience in writing, editing, and proofing evaluation reports. The verbal contortions that we evaluators tie ourselves up in when we try to explain certain perplexing results and the enigma we corner ourselves with when we realize that the way we reported a finding is not the only--or even the best--option available have fascinated me for years. I just enjoy hearing people argue over the best way to report a finding that has multiple interpretations or presentation options. In the hope of sharing some of these experiences with other evaluation report writers, 20 examples of these issues are described.

Conclusion. I know that conclusion means the end, but I do not want to lose the attention of any readers who do not make it through all 21 examples. So, here goes the conclusion.

1. Evaluation findings are seldom simple or straightforward. Evaluators want to be able to reduce our findings to a simplistic level that can be accepted by our audiences. We need to be constantly aware that our findings are complex and to watch for ways we are oversimplifying in our reporting.

2. Reporting evaluation findings is an art that requires a practical communication style focused on the intended audience. The audience for a public school evaluation report is usually not sophisticated enough, interested enough, or patient enough to learn from a university-research, professional-journal communication style. Despite this generalization, our audiences are diverse enough to defy formula report writing. We must develop the art of communicating in a style that is clear, straightforward, and jargon free if our findings are to become easily accessible and usable to our audiences.

3. The evaluator must be aware of and constantly learn about these and many other issues in order to communicate findings effectively. Most of the examples in this paper are not found in college texts on psychometrics and statistics--possibly because the authors consider these to be too frivolous for a text book. Experience and reading a variety of evaluation reports are the best teachers.

4. The key to how to describe an evaluation finding is in understanding and addressing the question that is being asked. This may seem obvious, but many of the evaluation reports I have read over the years either bask in the glory of the detailed analyses presented without realizing the distraction and obfuscation created, or they present a single measure of an outcome and miss the richness of their data. Incredibly, some reports wander around their analyses and never get to the point--never directly answer the primary question that was to be addressed.

THE _SSUES

1. The Difference Between A and B is Twice as Much as the Difference Between B and A.

A member of the Board of Trustees recalls that a previous report said that Program A cost 100% more than Program B; now we are reporting that the difference between the two is 50%. Were we wrong? Certainly not! All six of the alternative conclusions shown below are correct.

Program A: $2000 per student

Program B: $1000 per student

Conclusions:

1. The difference between the two programs is 100%.
2. The difference between the two programs is 50%.
3. Program A is 100% more costly than Program B.
4. Program B is 50% less costly than Program A.
5. Program A costs twice as much as Program B.
6. Program B costs half as much as Program A.

Clearly, we have the choice in reporting this comparison of the wording that represents the difference between the two programs as being somewhat larger or somewhat smaller.

2. Small Change Can Be Valuable.

One staff member reports that the dropout rate declined 1% from the previous year, but another reports that the decline was 9%. Which staff member is wrong? Although neither is wrong in terms of everyday language usage, the 9% figure is more accurate. Conclusions 2 and 3 below are clearly preferable and more precise than 1.

Dropout Rate:	1985	1986
	11%	10%

Conclusion:
1. The dropout rate declined 1%.
2. The dropout rate declined 9%.
3. The dropout rate declined one percentage point.

3. The Silent Majority is Alive and Living in Apathy.

One evaluator says that fewer than half of the teachers approve of the new policy, but the Superintendent says that more teachers approve than disapprove. Of course, the Superintendent is correct, but is the evaluator incorrect? No. The interpretation by each gives a different impression of the data shown below.

Item: Do you agree with the new policy?

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
10%	25%	50%	10%	5%

Conclusions:
1. 35% agree.
2. Fewer than half agree.
3. 15% disagree.
4. Fewer than half disagree.
5. More than twice as many agree as disagree.
6. Half are undecided.

4. We Did Not Mean to Agree.

The preliminary report says that all four groups agree that Program A is successful, but the final report says that the four groups differ widely in their opinions. Was the preliminary report wrong?

Group	Mean	Strongly Agree	Agree	Undecided	Disagree	Disagree
1	3.0	40%	10%	0%	10%	40%
2	3.0	10%	40%	0%	40%	10%
3	3.0	5%	20%	50%	20%	5%
4	3 0	3%	7%	80%	7%	3%

Conclusions:
1. There are no differences among the groups in their average response.
2. There are key differences among the groups in how they responded.

5. You Can't Tell the Program Without the Players.

The program staff members say that the program costs $1900 per student, but the evaluation report shows a per student cost of $9500. Who is correct?

Total Program Cost: $19,000,000

Cost per Student	Number of Students
$1900	10,000 Enrollment (number of students enrolled at any time, cumulative count)
$2533	7,500 Average daily membership
$2714	7,000 Average daily attendance
$2235	8,500 Peak membership
$9500	2,000 Full-time equivalents

Obviously, we can represent the cost of a program in many different ways depending upon our choice of how to count the students served. In fact each of these calculations is legitimate for particular questions. To compare with the cost per hour or day of regular instruction, using FTEs makes sense. For staffing, using peak enrollment makes sense.

6. Only Say "Only" When There is Only One Opinion.

Reporting evaluation results requires a cautious choice of words. For many reports, there are audiences who look for bad news and audiences who look for good news. As we know, people can take the same numbers and interpret them both ways. For example, a dropout rate of 26% can be seen as positive if it represents a decline from previous rates, or as negative if one wishes for a much lower rate. An evaluator may be making an error by using the qualifier "only" in front of 26% rather than leaving the interpretation up to the reader.

In the second conclusion shown below, the 41% rate could be viewed as excellent if the program participants are high-risk students who would have been retained if they had not participated.

Conclusions:
1. Only 26% of our high school students drop out.
2. Only 41% of the students in Program A were promoted to the next grade.

7. Odd Numbers and Decimals are not Approximations.

In reading reports, few things break my concentration and mystify me more than reading statements like those below. Approximations are supposed to be numbers that have not been carefully calculated to be exact or have been rounded to a whole, even number like 10, 50, or 90. Exceptions might be 25 and 75 because they represent quarters in our numerical system. However, if someone goes to the trouble to calculate a number with two decimal places, then that number just is not an approximation to anyone other than a physicist. Certainly, a number like 3511 can be an estimate of a population variable from a sample, but when it is written up, why not round it off to 3500 for the text of the report?

Conclusions:
1. About 47% disagree.
2. Approximately 3511 students graduate each year.
3. Seniors average about 3.11 in GPA.

8. Looking Out for Number Five.

Rounding numbers is a mental challenge to many people. However, the most perfect rounder can fall victim to the fives. Any time one encounters a five when rounding, it would be prudent to resort to the original number and divisor. The problem comes from rounding a number that has been previously rounded.

Original Number	Round 1	Round 2
49.49	49.5	50
49.49		49
49.49	49.5	50
50.49		50

9. N-Significant Differences Should be Noted.

We all know that sample size determines the degree of a difference required for statistical significance. I propose that researchers write the word "significant" as "sigNificaNt" when sample sizes are sufficiently large, so we can be alerted to the fact that the difference may in reality have a small educational significance even though the difference is statistically reliable. When sample sizes vary and the results should be interpreted accordingly, a researcher should write "sigNificant." When the sample sizes are small, and the difference must be relatively great to be statistically significant, then the researcher should use the traditional "significant." This way those of us in the know can interpret the educational significance of findings without having to read the technical notes. Why not?

10. Half Right Ain't Half Bad.

Typically, I have found that teachers reviewing standardized tests for adoption believe that the candidate tests are very difficult for the average student in the intended grade level. The insight that they lack is that the best measurement is achieved when the average student answers about half the items on a test correctly. They also see the more difficult items as being far beyond the reach of their lower achievers and do not understand that these students may not need to answer any of these items correctly to achieve an accurate score.

The context that teachers have is that their own tests are much easier because they grade on the basis of percentage of items correct. Typically, below 70% correct is failing. In fact, in Texas the law is that below 70% is failing. According to a previous study in our District:

- 75% correct on a teacher-made test is average.
- 57% correct on a norm-referenced test is average.

11. The Percentile Really is a Percentage.

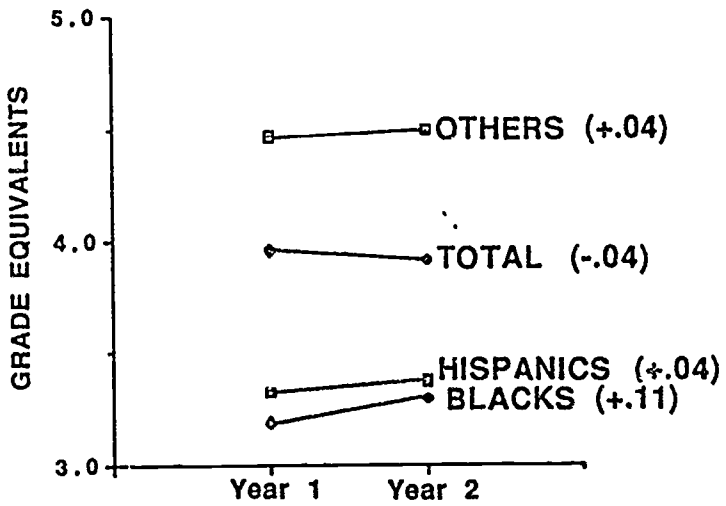
Too often someone who knows a little about percentiles says, "Percentiles are frequently mistaken for percentages. They are not percentages." Well, they are percentages. The distinction is that the naive user of percentiles mistakes them for a percentage of items correct--which they are not. However, they are a percentage--the percentage of other students who scored below a student.

12. Our Choice of Mean or Median Determines Which Students Need to be Taught.

Here is a cold and heartless reality. If a school's or program's goal is stated in terms of raising its mean score, then targeting instruction on the lowest achievers can result in the greatest benefit. If the goal is stated in terms of raising its median score or mastery rate at a set criterion, then targeting instruction on the students achieving around the median/criterion can result in the greatest benefit. Think about it. A mean is influenced by all the scores in the group, and the students with the potential for the greatest gain are the ones the farthest from the ceiling of the test. On the other hand, a median is unaffected by gains made by students who do not move from below the median to above the median, no matter how great their gains.

13. The Superintendent's Disbelief: Three Positives Make a Negative.

Indeed it happened, just the way Simpson described it back in the 50's. When our school system implemented a court-ordered cross-town bussing plan, the District achievement average went down as a result of the higher achieving nonminority students representing a smaller proportion of the total student population, even though the averages for every ethnic group went up.



ITBS GE Medians, Grade 3

Comparison of Changes in Median Grade Equivalent Scores

(ITBS, Grade 3)

ETHNIC GROUP	YEAR 1		YEAR 2		CHANGE	
	Median GE	N	Median GE	N	Median GE	N
Black	3.19	760	3.30	757	+.11	-3
Hispanic	3.33	1078	3.37	1108	+.04	+30
Anglo/Other	4.46	2443	4.50	1917	+.04	-526
Total	3.95	4281	3.93	3782	-.02	-499

14. The Junior High Principals' Dismay: Total Test Score is Lower Than Every Subtest.

Did you realize that if a person's or group's subtest scores are uniformly low, then its total test percentile will most likely be lower than any of the subtest percentiles? The opposite is true for high scores. The best way for me to grasp this is to consider a baseball analogy. A single player who is near the lead in home runs, runs batted in, and batting average may not lead in any category, but being very high in all three is unusual and can earn that person the highest overall ranking across all three areas.

TOTAL PERCENTILE	ALL OR MEAN SUBTEST PERCENTILE	DIFFERENCE
99	97	+2
95	92	+3
90	87	+3
85	83	+2
80	78	+2
75	73	+2
70	69	+1
65	64	+1
60	59	+1
55	55	0
50	50	0
45	45	0
40	40	0
35	36	-1
30	31	-1
25	26	-1
20	22	-2
15	17	-2
10	13	-3
5	8	-3
1	4	-3

ACHIEVEMENT LEVEL	PERCENTILE RANK SUBTEST 1	SUBTEST 2	SUBTEST 3	SUBTEST 4	TOTAL TEST
LOW	12	11	11	15	9
AVERAGE	51	51	51	51	51
HIGH	90	90	90	87	93

15. Chapter 1 Chagrin: Losing Ground with Percentile Gains

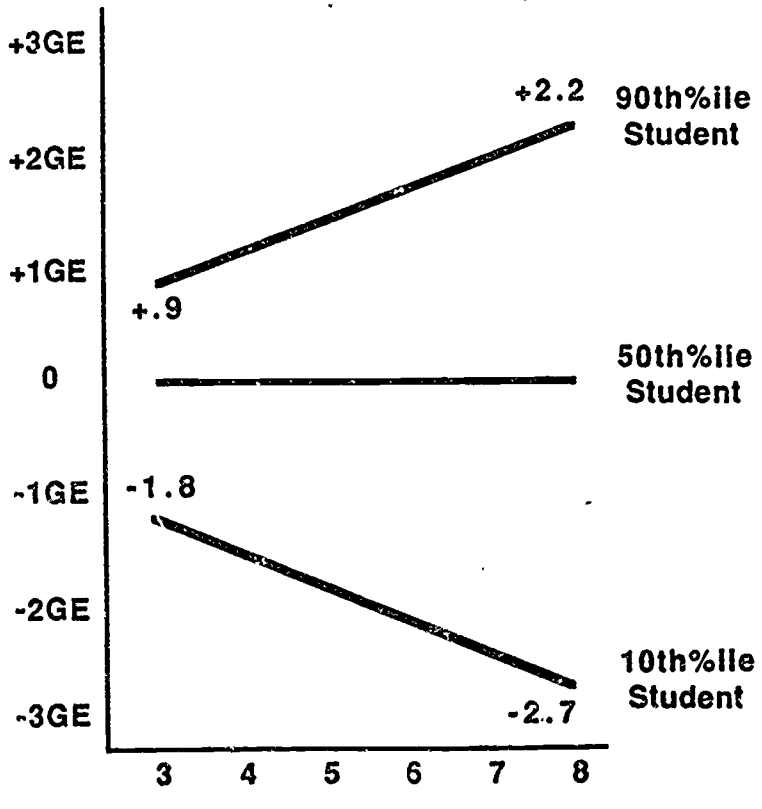
All evaluators dream of finding that miracle--especially a program that works miracles with low-achieving students. Unfortunately, evaluation reports that use percentiles or NCEs exclusively may give the impression to some that the participants are closing the gap between themselves and average achievers when indeed the gap is increasing. The bottom line is that a low-achieving student such as the one represented in the example below can make a percentile gain but end up farther behind in relationship to grade level.

Student A

Year 1 (Grade 4)	27th Percentile	3.0GE	1.8GE < Grade Level
Year2 (Grade 5)	28th Percentile	3.9GE	1.9GE < Grade Level
		+.9GE	-.1GE Loss

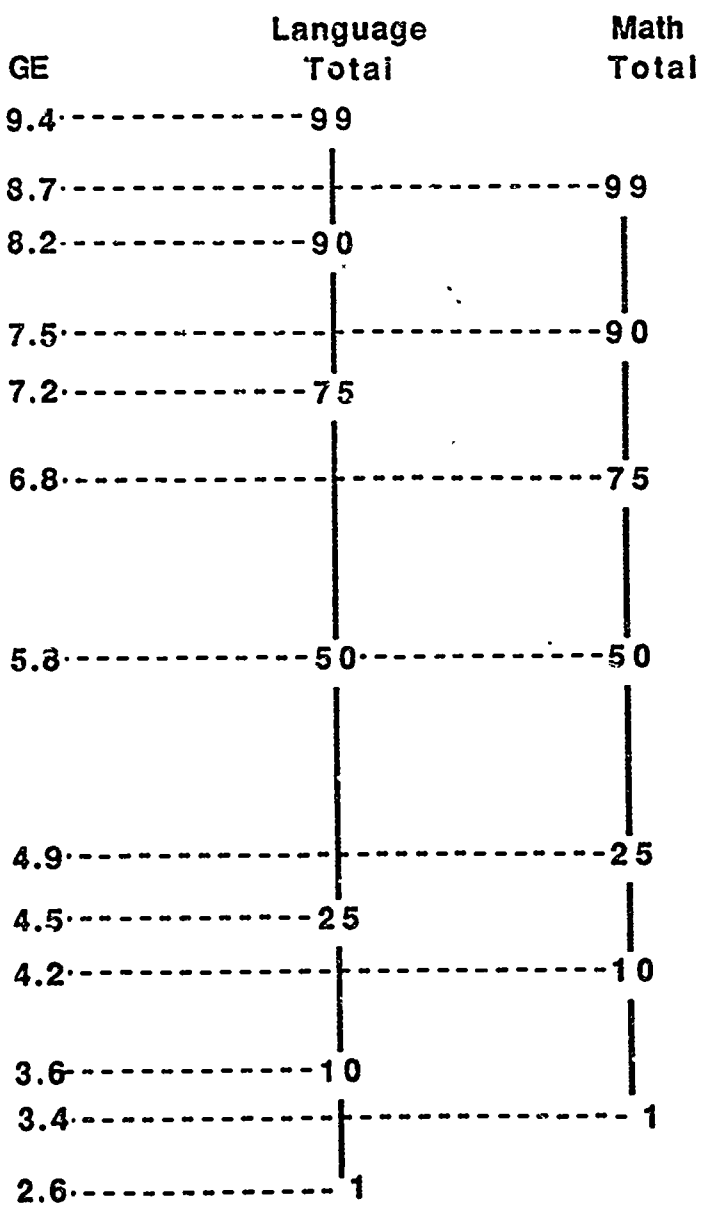
Student B

Year 1 (Grade 4)	78th Percentile	5.1GE	1.3GE > Grade Level
Year 2 (Grade 5)	77th Percentile	6.2GE	1.4GE > Grade Level
		+1.1GE	+.1GE GAIN



16. The Parent Trap: Six of One is More Than Half a Dozen of the Other.

Parents can easily be confused when a student makes the same percentile score in both language and mathematics but they find out that the student is functioning half a grade level higher in language than in mathematics.



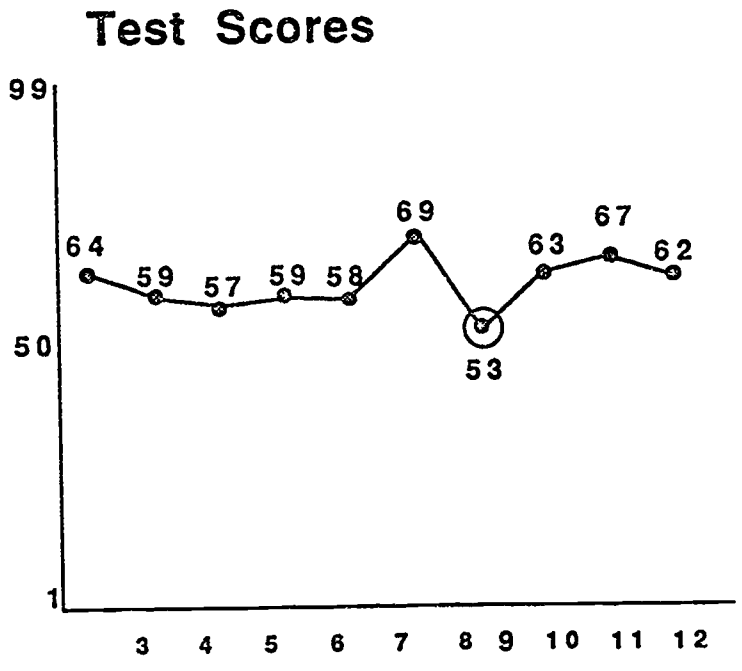
17. Evaluating the Principal: The Highest Score May Not be the Best Score.

Even though we all know that context variables can influence achievement test scores more than the effect of school instructional variables, it is always good to have some clear examples of this issue.

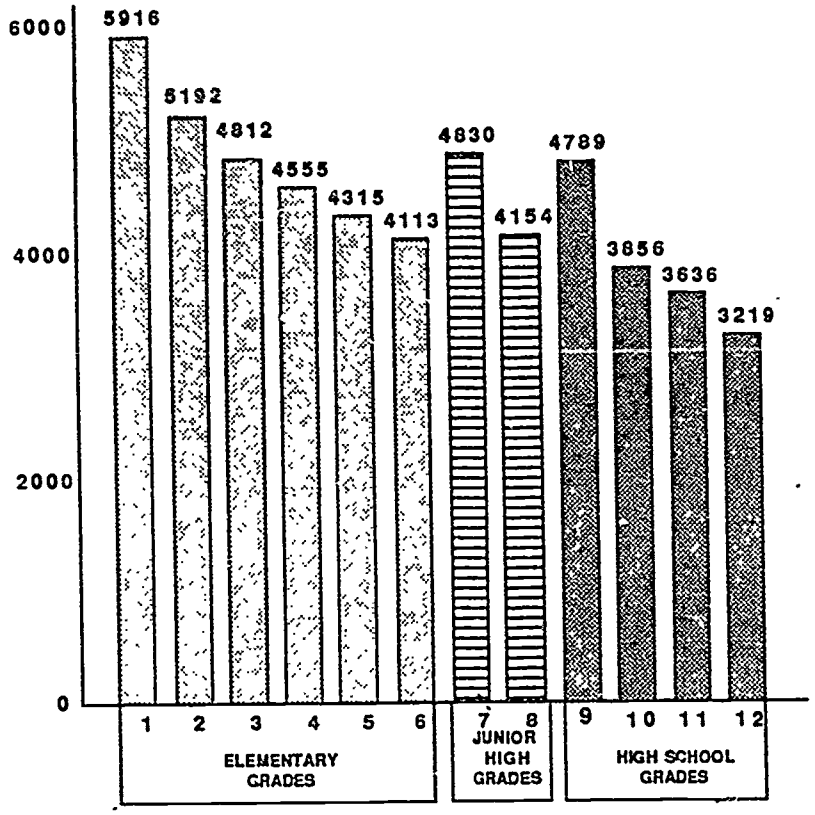
	% of Students Gaining One or More Years in GEs		% of Students Gaining More Than Predicted Based on Prev. Achievement, Income, and Other Factors	
	%	Rank out of 60 Schools	%	Rank out of 60 Schools
School A	**48.6**	**39**	**47.0**	**40**
School B	**42.5**	**50**	**60.7**	**3**

18. Caution--Hazardous Grade: Ninth Graders at Risk

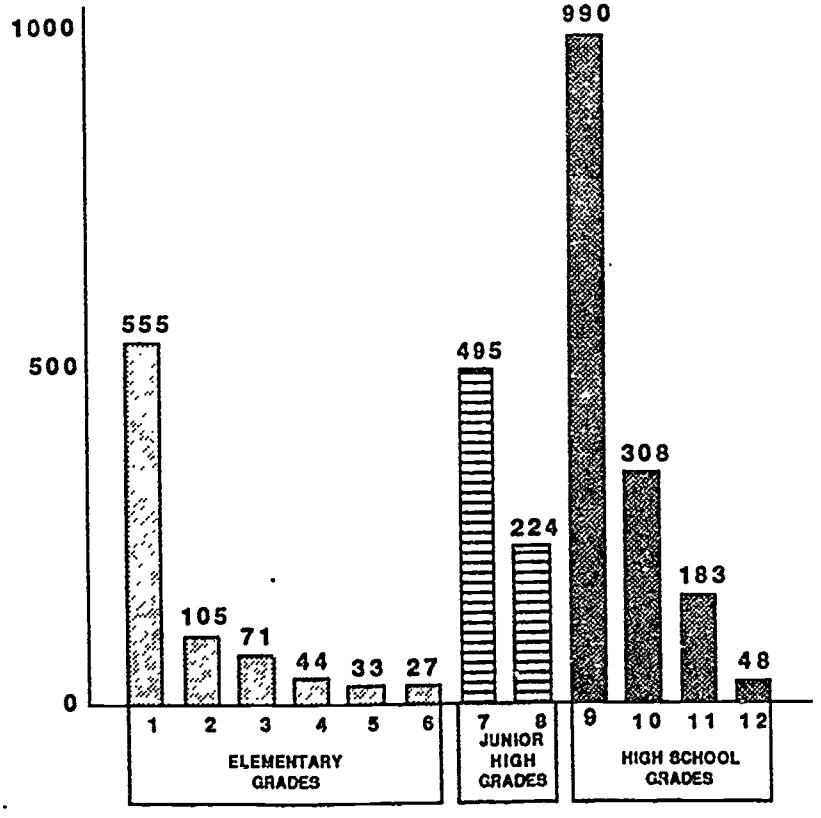
Individual grade levels often present anomalies in the comparison of averages across grade levels.



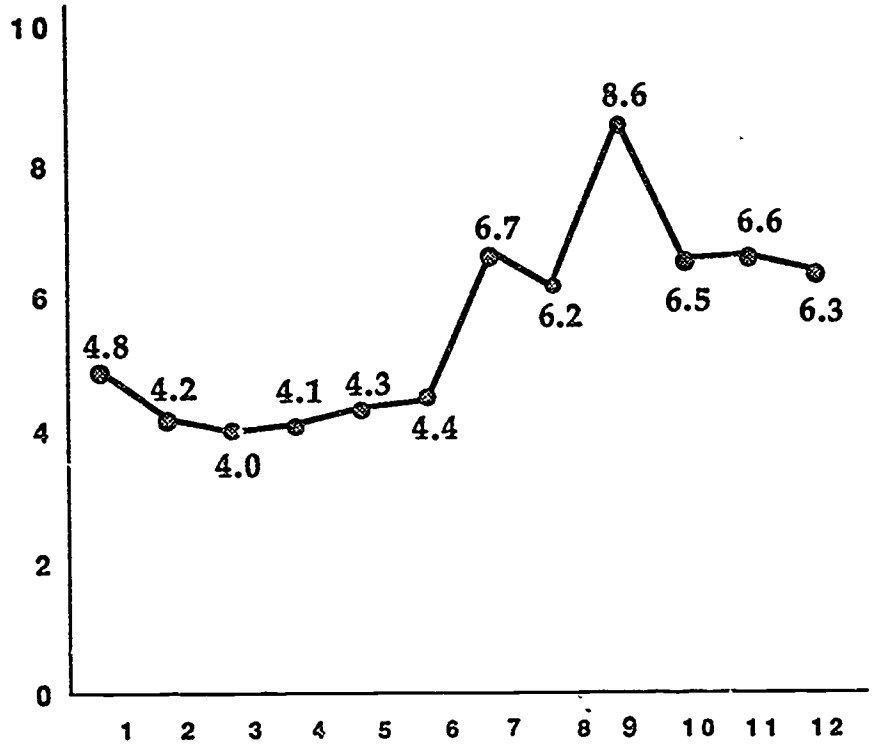
Enrollment



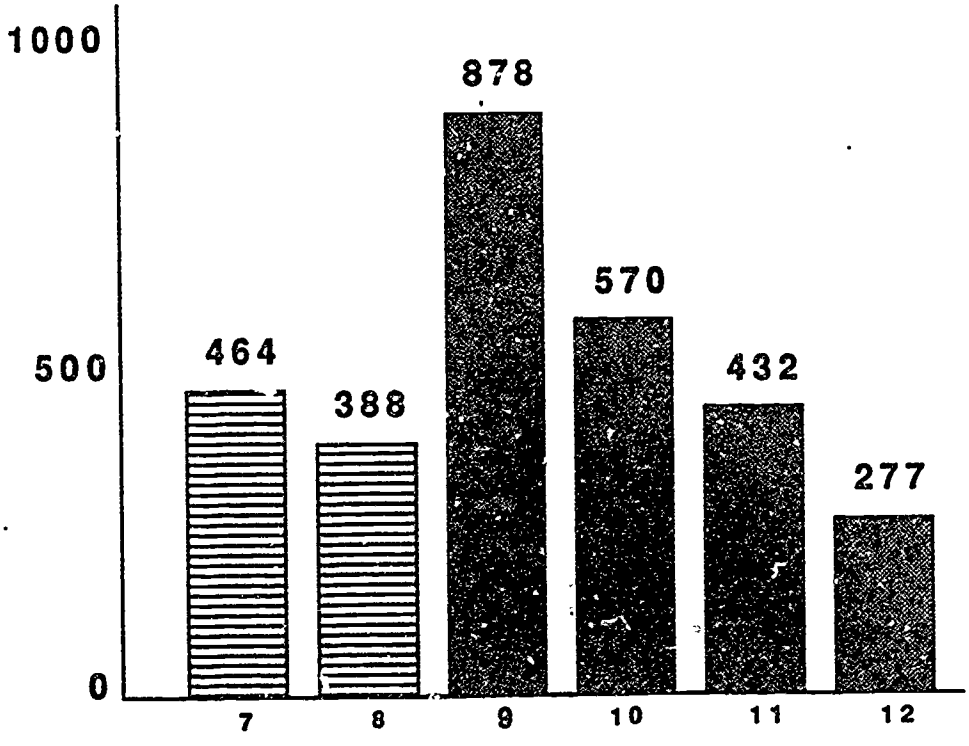
Retainees



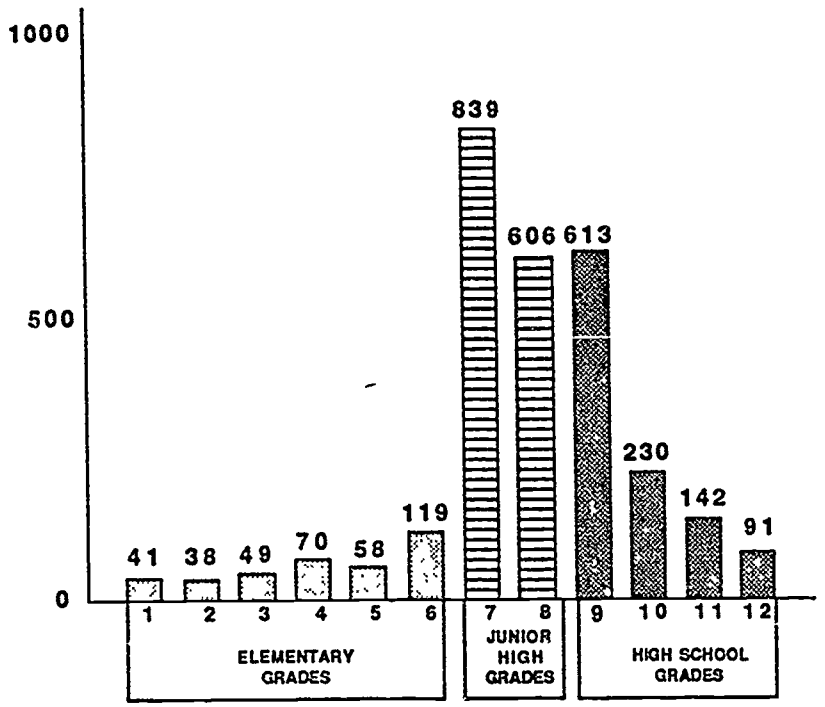
Absences



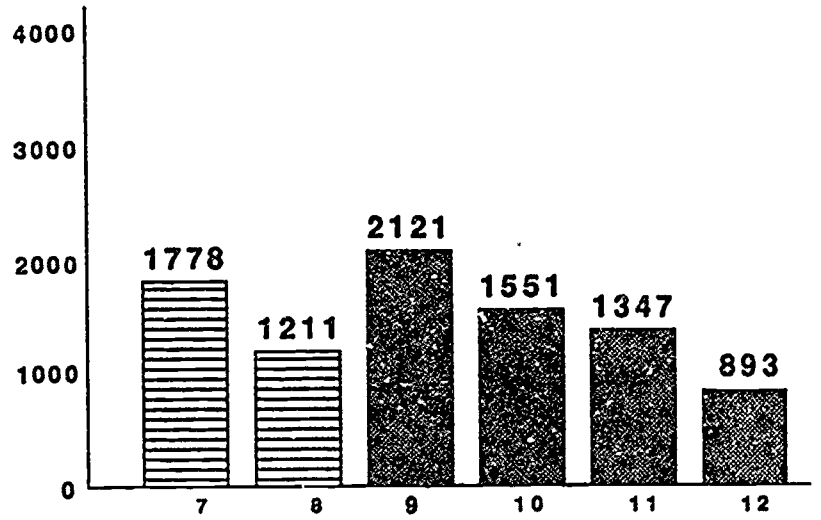
Dropouts



Students Disciplined



Students Who Earn Failing Grades



19. TEAMS Quirk: 100% Mastery but Still a Bottom 25% School

In Texas, we have a statewide minimum skills test (TEAMS, Texas Educational Assessment of Minimum Skills) that has been used to rank schools to determine the ones most in need of improvement. Incredibly, the ranking has not been done on the basis of percentage of students mastering the TEAMS, a criterion referenced test, but on a mean standard score. This mismatch creates the interesting situation of allowing a school to have 100% mastery by its students but still fall in the targeted bottom 5% of schools.

TEAMS MASTERY

School A	School B
700	690
705	690
710	690
715	700
720	740
725	750
730	760
734	770
Mean= 717	Mean= 724
100% Mastery	62.5% Mastery

20. Pinning TEAMS to the MAT: Equating is an Annual Affair.

The Friends for Education were surprised to find Texas reporting a statewide average above the national average, but they had not examined how Texas manufactured its average. The TEAMS was equated to the MAT6 in 1985. Since that time, the State has seen an impressive gain in the TEAMS skills, but continues to use the original equating. The assumption would have to be that Texas students have made equivalent gains across all skill levels, not just the minimum skills measured by TEAMS, in order to make the original equating still useful.