DOCUMENT RESUME

ED 131 091 TM 005 752

AUTHOR Echternacht, Gary
TITLE Test Bias in the Absence of a Criterion.
PUB DATE [Apr 76]
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Item Analysis; *Test Bias

ABSTRACT

This paper proposes a method of transforming item p-values (the proportion answering a test item correctly) to what are termed "delta" values. First used by Conrad in 1948, deltas are routine statistics computed in all analyses at Educational Testing Service. Using this approach one would conclude no test bias if differences in resulting deltas are constant from item to item for the groups. Under a null hypothesis of no test bias, the sample delta differences should be distributed normally with some unknown mean and unknown variance. If evidence can be gathered to the contrary, the null hypothesis can be rejected and bias concluded. To test the hypothesis of normality, one plots the differences in item deltas on normal probability paper. First, one orders the delta differences. Second, one pairs each difference with the value S/(m+1) where S is the rank of the delta difference and m the number of items. The purpose of S/(m+1) is to anchor the median difference to the 50th percentile. If the differences follow a normal distribution, the plotted points will be on a straight line. A statistical test has been developed by Lilliefors (1967) using the sample mean and variance as a basis for the straightness of the resulting line. That technique was adapted for use with probability paper and formed the crux of the technique. (Author/MV)

Test bias in the absence of a criterion

Gary Echternacht
Educational Testing Service

In the early 1970's testing programs at Educational Testing Service (ETS) became concerned about the possibility of test bias in their tests. Accusations had been made in the general literature to the effect that standardized tests, including those developed by ETS, were discriminatory or biased against certain groups of the testing population. In response to those accusations, most test programs began research aimed at finding empirical evidence that could substantiate claims of test bias. Two approaches were used. Some studies examined test bias in the context of a criterion variable. Because some test programs found it difficult or even impossible to collect adequate criterion data due to the effect of extensive prior selection or small sample sizes, several studies of bias were funded using approaches not requiring a criterion. I will describe one such method used to ascertain the extent of bias in some of these studies.

When we first began to develop a methodology for studying test bias in the absence of a criterion, we considered several alternatives. The most straightforward approach was to examine differences in item p-values (the proportion answering an item correctly) for the groups. One would conclude that there was no test bias if the item p's differed by the same constant for each group. Conceptually, however, this apprcach resulted in some problems. For example, if the actual difference between p-values was .20 in the population, one would necessarily conclude bias if there were any very easy or very difficult items--i.e., where p-values tended to be around .1 for the higher scoring group or .9 for the lower scoring group. Because test developers favored including a few very easy items and a few very difficult items in their tests, this approach was not considered.

There was also a statistical problem associated with the first approach--the heterogeneity of variance encountered in performing statistical tests. Although making an arcsine transformation avoided this problem, the difficulties associated with the ranges in p-values persisted. Thus, this alternative was no longer considered.

There did appear to be a way to get around the problem of easy and difficult items. This could be accomplished by transforming p-values to what are termed "delta" values. First used by Conrad (1948) in his *Psychological Monograph*, deltas are routine statistics computed in all analyses at ETS. Essentially, the results of making the delta transformation is to transform p-values to variables ranging from minus infinity to plus infinity, with mean 13 and standard deviation 4. Details of the transformation can be found in either Conrad or my *Educational and Psychological Measurement* paper (see Echternacht, 1974). Using this approach one would conclude no test bias if differences in resulting deltas were constant from item to item for the groups.

I would like to mention another approach developed by Richard Potthoff (1966) that was highly regarded and used in some of our studies. In general, it relaxed the requirement for unbiasedness that item p-values differ by a constant by requiring only that one group score consistently higher than the other group on all items. At the same time, it required as a condition for unbiasedness that if one group scored higher on one item compared to a second item, then the other group should also score higher on the first item vis-a-vis the second.

My first choice of approach was Potthoff's method because it seemed to allow for some vagueness in the mathematical definition of bias.

Unfortunately, to implement this approach, extensive computer programming was required, and the results were difficult to interpret to test developers. For example, the only conclusion one could provide test developers was that the test was either biased or not biased. Test developers wanted more. They wanted to know which items were biased or at least where they might devote their resources toward removing any bias. They also wanted a methodology they could apply in a variety of situations without consultation. Because there did not seem to be a way to accomplish the objectives of the test developers, the third approach--viz., differences in delta being constant--was adopted.

Under a null hypothesis of no test bias, the sample data differences should be distributed normally with some unknown mean and unknown variance. If evidence can be gathered to the contrary, the null hypothesis can be rejected and bias concluded. To test the hypothesis of normality, one plots the differences in item deltas on normal probability paper. First, one orders the delta differences. Second, one pairs each difference with the value S/(m + 1) where S is the rank of the delta difference and m the number of items. The purpose of using S/(m + 1) is to anchor the median difference to the 50th percentile.

If the differences follow a normal distribution, the plotted points will lie on a straight line. A statistical test has been developed by Lilliefors (1967) using the sample mean and variance as a basis for judging the straightness of the resulting line. That technique was adapted for use with probability paper and forms the crux of the technique.

The tables and figures following this paper illustrate the technique with test data from the reading comprehension section of the test used in graduate business school admissions. The data presented compares the performance of black and white males.

The first table shows the item numbers, differences in deltas, the rank of the difference, and the value of S/(m + 1).

In the figure following the table, the differences in deltas have been plotted against the values of S/(m + 1) for the thirty items.

In the following figure a solid line has been added to represent the plot of the hypothesized normal distribution with mean -12.4 and standard deviation 7.2. The sample mean is plotted at 50 and the sample standard deviation is added to the mean and plotted at 84. A straight line is then drawn between these two points representing the illustrated line.

Confidence bands are then drawn on the basis of tabled values in Lilliefors' paper. Lilliefors gives a critical value of .161 for a sample of 30 at the .05 level of significance. In the scale used in this example 16 is used as the critical value. The technique then calls for measuring 16 units on the horizontal axis in each direction for a number of points and subsequently connecting these points to form a significance band. It is best to construct limiting lines at the end of each significance band. First draw lines at 100 x (critical value) and 100 (1 - [critical value]). Horizontal lines are drawn from the point where the vertical line crosses the solid lines. These are useful for determining the shape of the significance bands at the endpoints.

If any plots fall outside the significance bands, one concludes bias. In this example no bias was found.

References

Conrad, H. S. Characteristics and uses of item analysis data. *Psychological Monographs*, 1948, *62*(Whole No. 295).

Echternacht, G. J. A quick method for determining test bias. *Educational and Psychological Measurement*, 1974, *34*, 271-280.

Lilliefors, H. W. The Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 1967, *62*, 399-402.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. Institute of Statistics Mimeo Series No. 479. Chapel Hill: University of North Carolina, Department of Statistics, 1966.

Table 1

Delta Differences

(White Males - Black Males)

m = 30 items

Rank(s)	S/(m + 1)	Difference x 10	Summary Statistics
1	3	-26	$\overline{X} = -12.4$
2	6	-21	S = 7.2
2	6	-21	
2	6	-21	
2	6	-21	
6	19	-20	
7	23	-18	
7	23	-18	
9	29	-17	
9	29	-17	
11	35	-16	
12	39	-15	
12	39	-15	
12	39	-15	
15	48	-14	
16	52	-13	
17	55	-12	
18	58	-10	
18	58	-10	
18	58	-10	
13	58	-10	
22	71	- 9	
23	74	- 8	
27	87	- 7	
28	90	- 4	
29	94	- 3	
30	97	- 2	

Figure 1. Plots of delta differences on probability paper

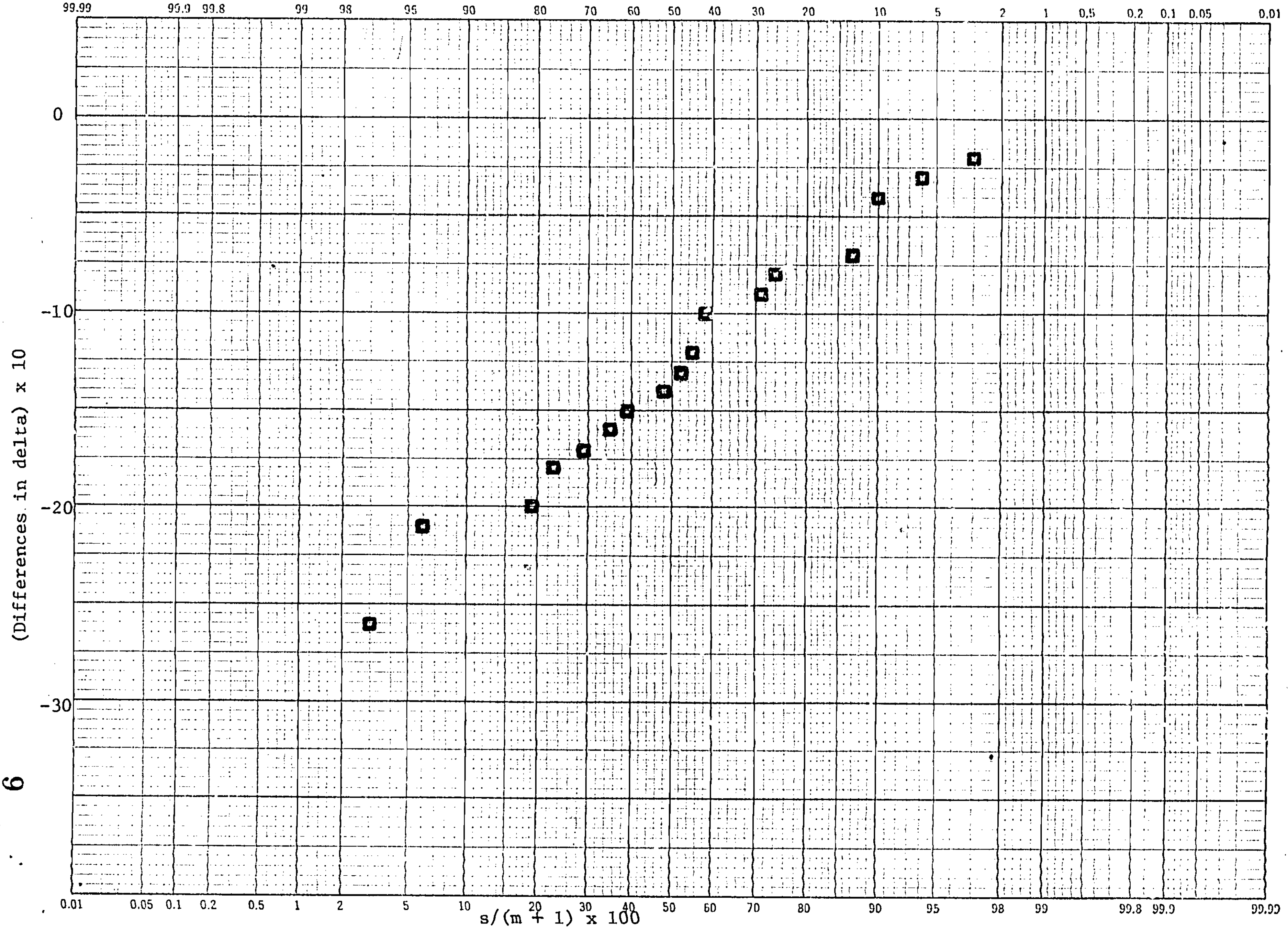


Figure 2. Plots with hypothetical normal distribution

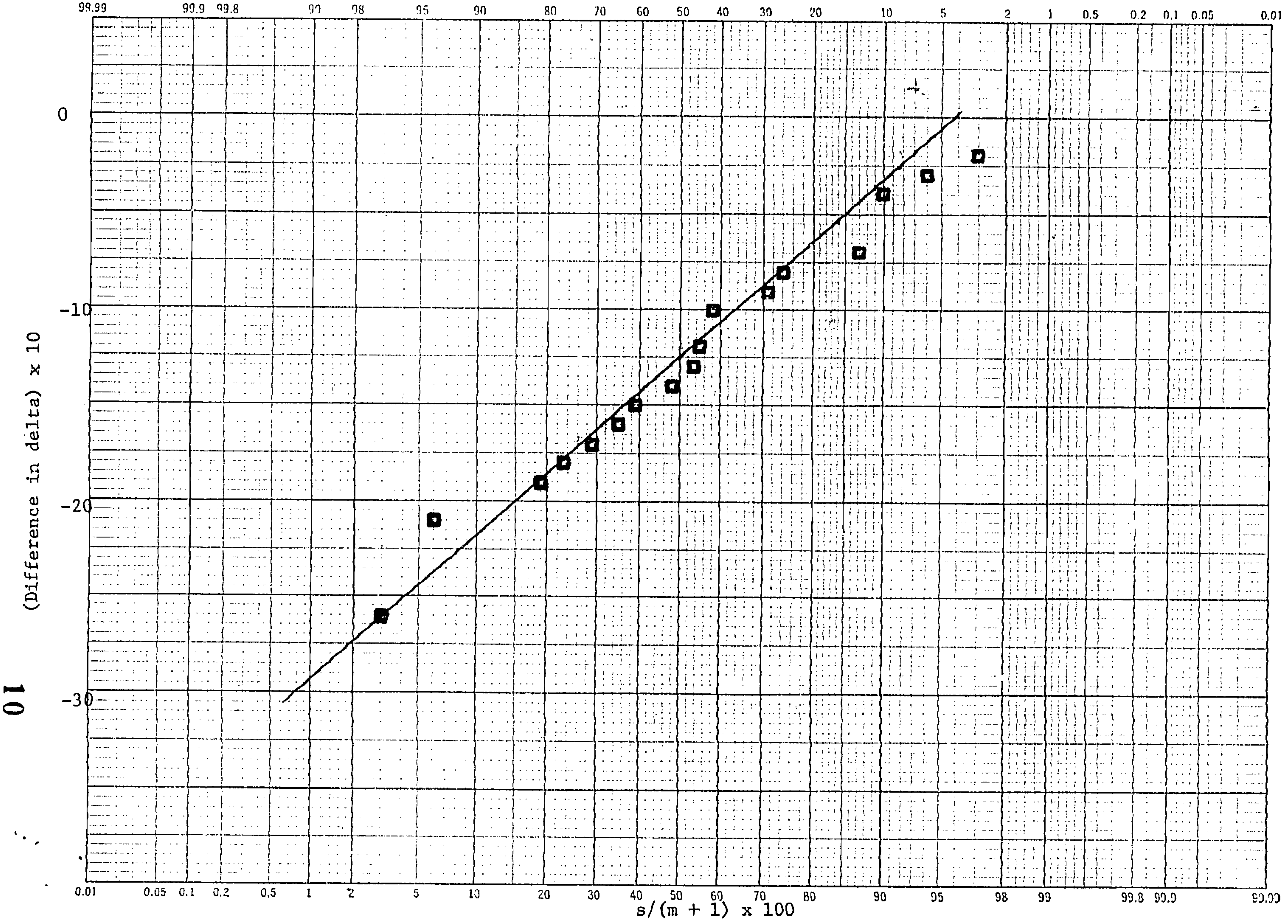
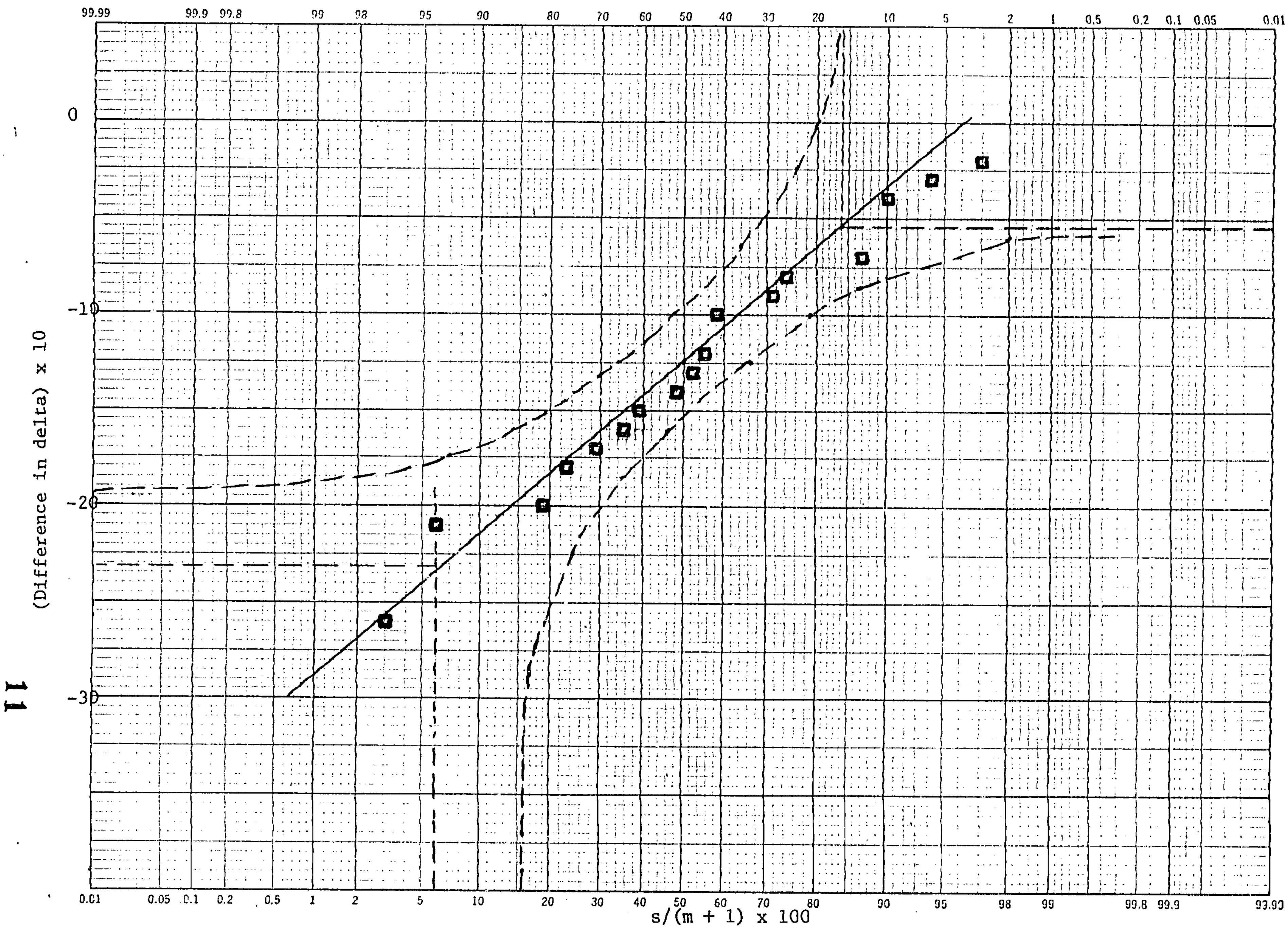


Figure 3. Plots with confidence bands added


(Difference in delta) x 10
s/(m + 1) x 100