8. IDENTIFYING ITEM BIAS

The past twenty years has witnessed increasing concern about test bias. This has produced a substantial amount of literature. A few of these articles actually deal with the critical issues in test bias, but most of what has been published is ill-suited to actual practice.

Psychometricians have tried to deal with the technical issues of test bias from many perspectives. This chapter looks at item bias from the point of view of Rasch measurement and shows how item bias can be detected and dealt with in test practice. The techniques we describe are straightforward and easy to apply. They work with most measurement applications. Were these techniques to be routinely used, whatever item bias actually existed would be clearly identified and could be easily monitored and controlled.

There has been a fundamental error in thinking about bias which has lead to confusion over what bias is and, hence, how to detect it. This error occurs whenever the detection of any "difference" at all in test scores is immediately assumed to signify bias. The error typically occurs when contrasting samples are compared and found to be different in their measures. Examples of this confusion are differences in mean test scores between demographic types like males and females or blacks and whites. When such a difference is identified, the accusation is made that "bias has been found to exist." But "differences" of this kind do not signify bias.

The fallacy in such accusations can be illustrated by simple examples which show that "differences" in measures are not proof of measurement bias. Suppose we weigh two groups: professional football linemen and professional jockeys. When we compare the mean weight of these two groups a great difference in pounds will be found and indeed is expected. Would we then infer that this observed difference in weight indicates that the scale used for weighing these equally professional athletes was biased against jockeys? Similarly, if we compared the height, weight or general skills of 18 year olds with those of 8 year olds, would any differences found in favor of 18 year olds be taken to indicate bias in the measuring instruments? If the average height of 8 year olds is less than that of 18 year olds, is the ruler biased? Of all the numerous practical illustrations of this type that we could cite, none would cause us to conclude that the observed differences were indicative of biased measuring tools. Hence we must realize that differences in measures do not necessarily signify bias. We must look further into the question of bias for its *necessary* indicators.

The phenomena that is actually indicative of bias is significant and persistent *interaction* between some *but not all* persons and some *but not all* items. When a measuring process encounters unexpected differential effects within the replications necessary to estimate a measure, this *unmodeled interaction* is an indication of possible bias. *Differential interaction* between some items and some persons produces results which cannot be predicted within the intended frame of reference. Interaction confounds the intended interpretation of test scores. Interaction confuses interpretation because we can no longer base our measures upon the replications of the variable implied by the measuring instrument, but must, instead, take into account a second, poorly defined variable which differentially affects the manifest relations between some persons and some items.

Sometimes these interactions are substantial enough to spoil the resulting test scores and sometimes they are not. Suppose we give a sixth grade student an easy arithmetic word problem to read and solve. If he fails to give a correct answer, is it due to a reading problem, or to difficulty with arithmetic or to both or to something else? To identify an answer as incorrect without reviewing the probability of it's being incorrect and, when the answer is improbable, diagnosing the reasons for this unexpected incorrect answer, is incomplete. No count of right or wrong answers can, in itself, yield information about the reason for an improbable error. An improbable error (or success) implies the possibility of an interaction between person and item with respect to some secondary variable also active in the testing situation. When such confusion occurs, how can we detect it? What can we do about it? Here is how to proceed.

We want to find out if any items in a particular test are biased, say, against girls (or boys). Here are the steps to follow:

1. Examine the items carefully for sex-linked content and then classify them according to "theory" as a) those expected to favor boys, b) those expected to favor girls and c) those expected to be neutral. This is an important first step. If we really have no idea what we are looking for, we will surely have difficulty finding it. Worse, we will be seduced into mistaking accidental and transient irrelevancies for enduring effects.

2. A sample of girls and a sample of boys must take the test, if they have not already done so.

3. A separate calibration of the test in question is done for each sample - one for boys, another for girls. (Test calibration is explained and demonstrated step-by step in Wright and Stone, 1979, pages 28-62.)

4. The calibrated item difficulties from the separate analysis of each sample (a boy item calibration and a girl item calibration for each item) are centered and plotted against each other.

5. An identity line is drawn through the origin of this plot with slope one.

6. Statistical control lines are constructed around this identity line to guide interpretation and the plot is examined to see whether any items fall outside the control lines and hence are statistically identified as possibly biased. (See Wright and Stone, 1979, pp. 94-95 and Wright and Masters, 1982, pp. 115-117.)

We will illustrate these steps by examples designed to give the reader visual experience with the configurations that usually occur.

Figure 8.1 is a plot of two such item calibrations. The items expected to favor boys are indicated by triangles. The items expected to favor girls are indicated by circles. In Figure 1, the item plots center around the identity line. The items expected to be biased are not separated from each other. All items are within the 95% control lines. There is no indication of item bias in this plot which brings together the separate item calibrations for boys and girls. We must conclude that these data provide no reasons to suspect item bias with respect to sex.

Figure 8.2 is a different plot of item difficulties for boys and girls. In Figure 2 we can see two distinct item streams. One large item stream containing items favoring boys and also girls runs slightly above the (dotted) identity line. A second smaller stream of items favoring girls runs well below the dotted identity line.

To clarify what has occurred we draw a second (solid) identity line (also with slope one) through the middle of the larger stream of mixed items. Now we add control lines at two standard errors out around the solid identity line. This helps us to see the statistical separation of the two item streams. A difference is clearly indicated. There is an interaction between item content and sex which makes scores on the original mixture of items ambiguous. However, the majority of items in the larger stream might be used to provide unbiased measures on a "new" variable defined now by the particular items in the larger stream.

In Figure 8.3 we have another situation. Now we have three streams of items. One stream of items is above the identity line and favors boys as expected, another stream of items is below it and favors girls also as expected. Finally, a third stream of mixed items follows the identity line. Each stream of items is clearly distinct from the other. The question before us is: Which item stream defines the variable that we intend? The answer cannot come from the statistics. We must review the prior intention which motivated the composition of these items in order to make a sensible decision. We must decide which of the three streams of items contains the content which best, by our definition, defines the variable we intend. Once we have made this decision, the other items will become, by our definition, deviant from the frame of reference of this intention and hence "biased."

Figure 8.1 demonstrates what we will see when two samples produce no evidence of bias because all items plot along the expected identity line.

Figure 8.2 shows a larger stream of items slightly above the original identity line and a smaller stream of items below it. The simplest conclusion is that the smaller stream of items is biased with respect to the variable defined by the larger stream of items along the identity line.

In Figure 8.3, the situation is more complicated. We must decide which two of the three streams of items are deviant. We must decide which item stream marks out our intended variable. Does our intended variable remain with the original identity line or does it follow one of the offset streams of items? The example in Figure 3 causes us to realize that sometimes we will be forced to go beyond our statistics to outside criteria in order to establish a basis for judgment. Statistical analysis can show us what we have observed, but we must go beyond the data to make a criterion decision.

Our next example is from real data. It is a practical situation involving public school achievement test scores. Figure 8.4 is a plot of item calibrations made from two classrooms. One class is at Grade 2 and the other class is at Grade 3. Both classes took the same arithmetic computation test. The plot of item calibrations for the two samples, Grade 2 vs. Grade 3, shows two items clearly differing from the overall cluster of items.

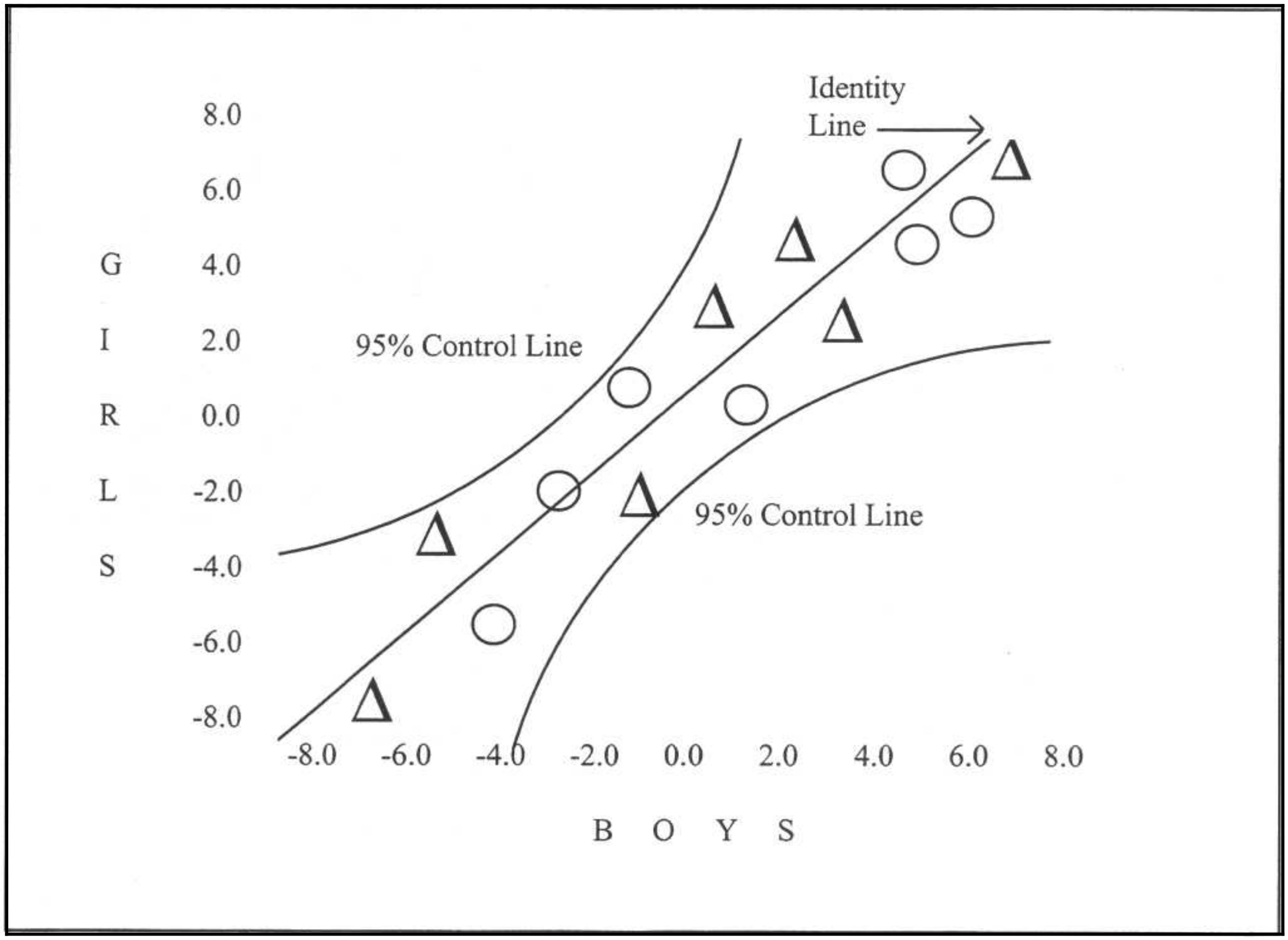
For these data we have some important external criteria, namely the content of the items. The computation skills required for most of these items are addition and subtraction of whole numbers without regrouping. The two deviant items, however, have common characteristics. They both involve subtraction with regrouping. This arithmetic operation marks a difference between the two grades. Within the frame of reference of whole number addition and subtraction without regrouping, there is an interaction between these two subtraction with regrouping items and grade level which makes these two items "biased."

These two items are biased against those second graders who have not yet learned how to do regrouping. The few second graders who were successful on these two items are ahead of their peers. For the others, these two items are almost impossible.

These examples make clear that a practical strategy is required to determine whether and how "bias" is evident. We have used Rasch measurement to illustrate how this can be done. It is especially important to be clear about our intentions prior to analysis in order to use the intended meaning of the items to help us understand the results of our analysis. Identification of bias is possible only when procedures like the one described have been applied. External criteria are needed to interpret results. But the criteria selected must be unequivocal in their application to the problem or they cannot be useful.

Figure 8.1

No evidence of item bias.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

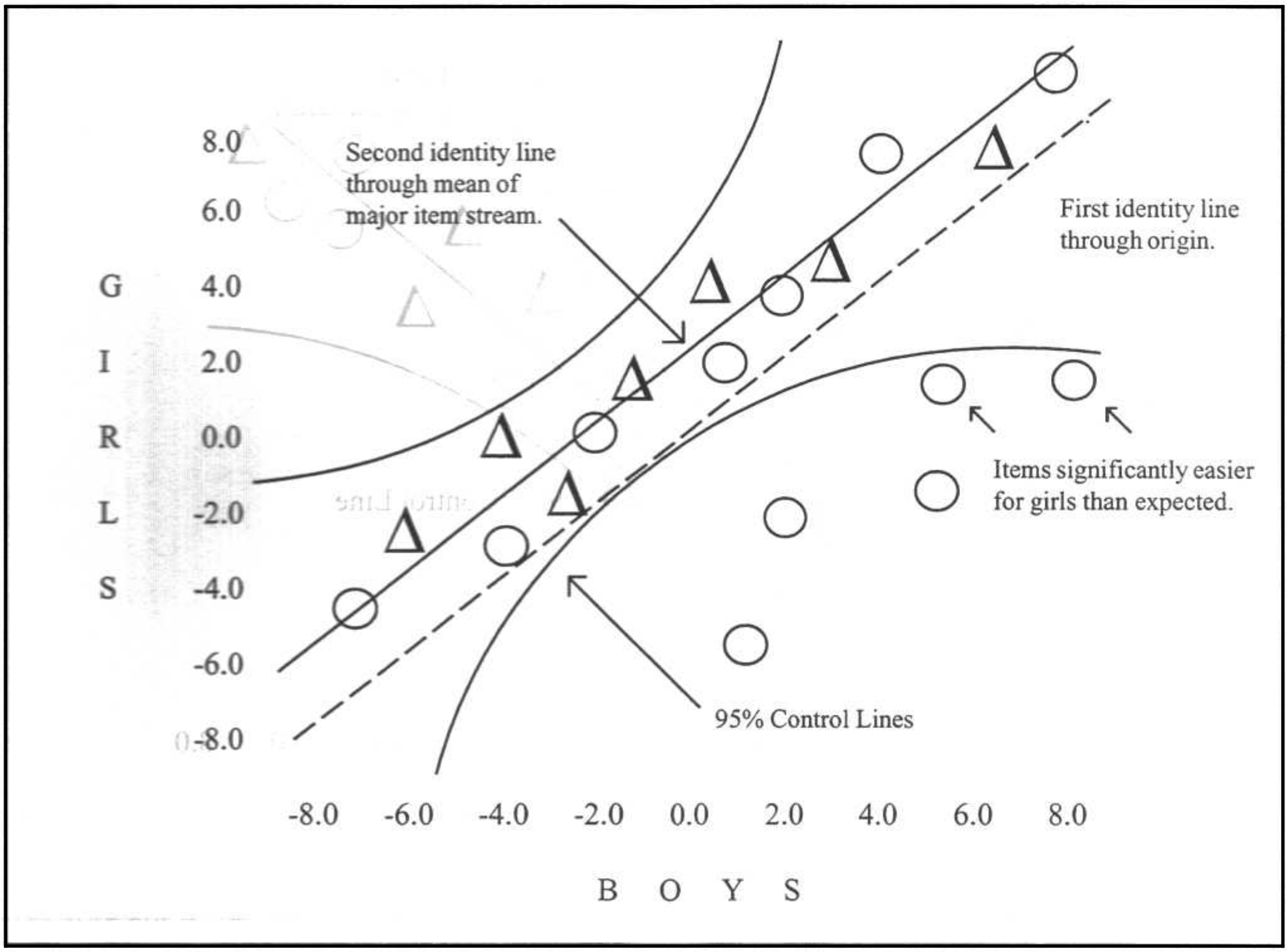
3. Circles = items preclassified "girl favoring."

4. Triangles = items preclassified "boy favoring."

5. For mathmetical specification of control lines see Wright and Stone, 1979, pp. 94-95 or Wright and Masters, 1982, pp. 115-117.

Figure 8.2

Five items biased in favor of girls.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

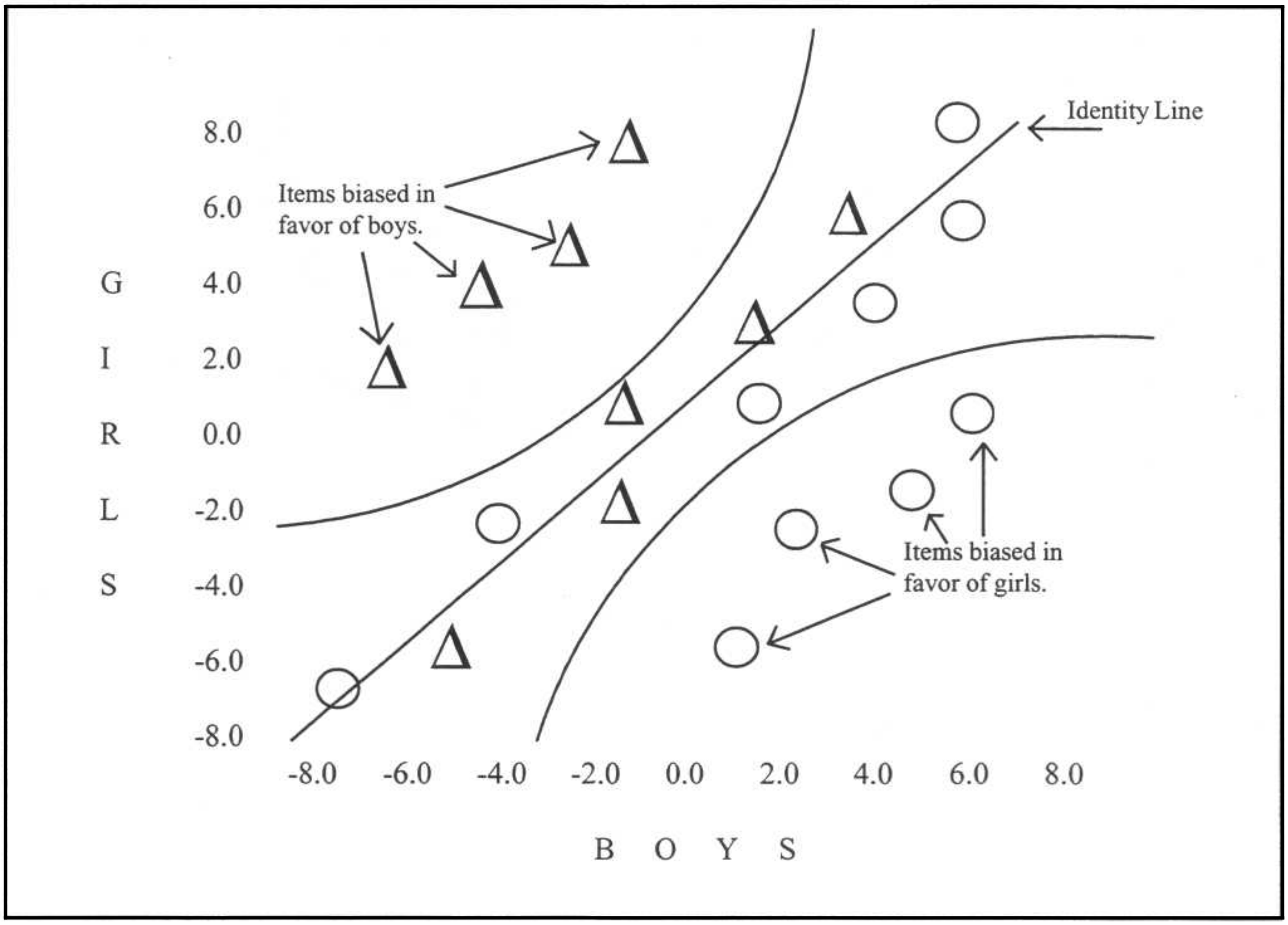
3. Circles = items preclassified "girl favoring."

4. Triangles = items preclassified "boy favoring."

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Maters, 1982, pp. 115-117.

Figure 8.3

Must decide which item stream defines the intended variable.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

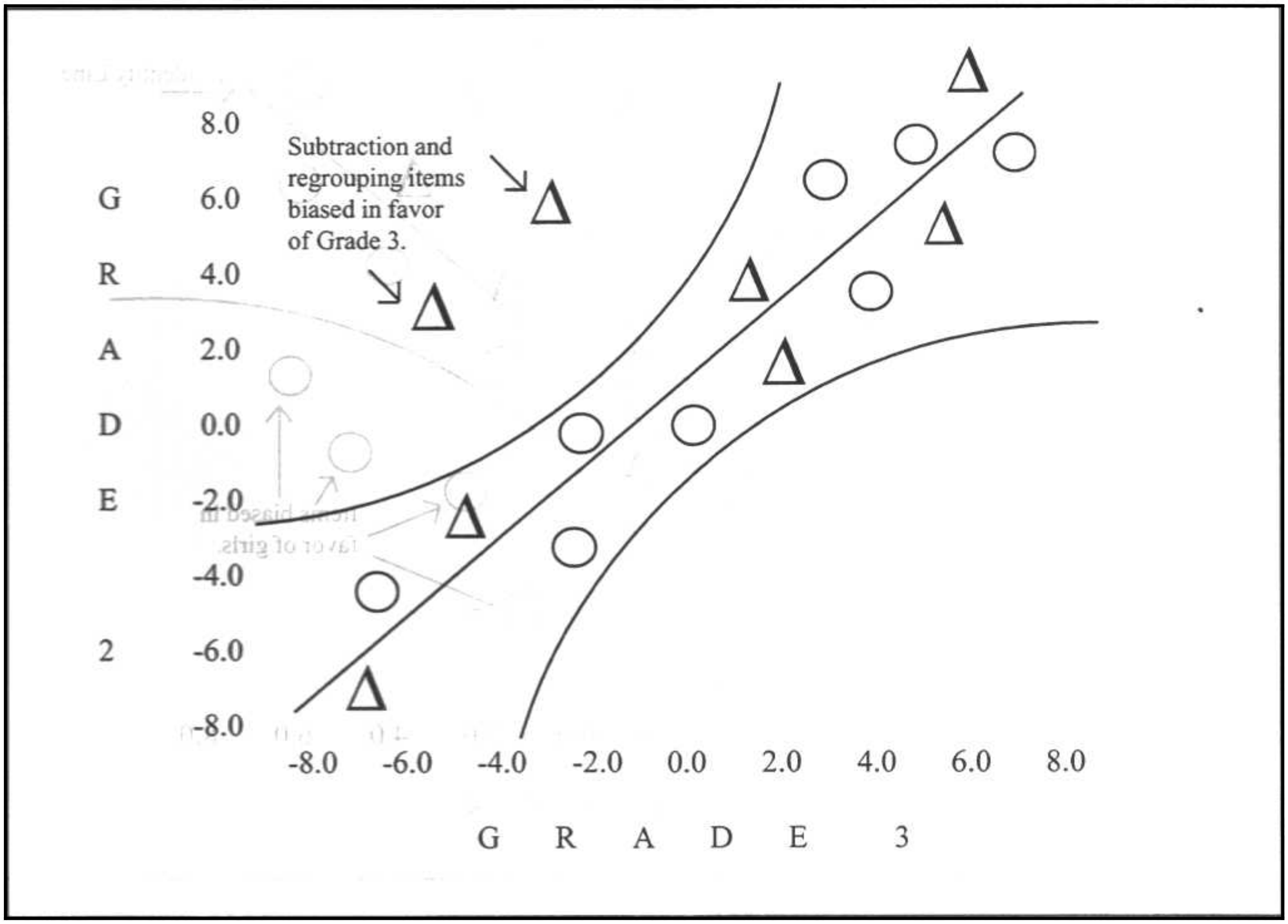
3. Circles = items preclassified "girl favoring."

4. Triangles = items preclassified "boy favoring."

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Masters, 1982, pp. 115-117.

Figure 8.4

Item calibratrions from grade 2 and grade 3 on an arithmetic achievement test.



1. Vertical axis is item difficulty for Grade 2.

2. Horizontal axis is item difficulty for Grade 3.

3. Circles = Grade 2 items.

4. Triangles = Grade 3 items.

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Masters, 1982, pp. 115-117.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware