Item statistics for Jateng exams

12 August 2007

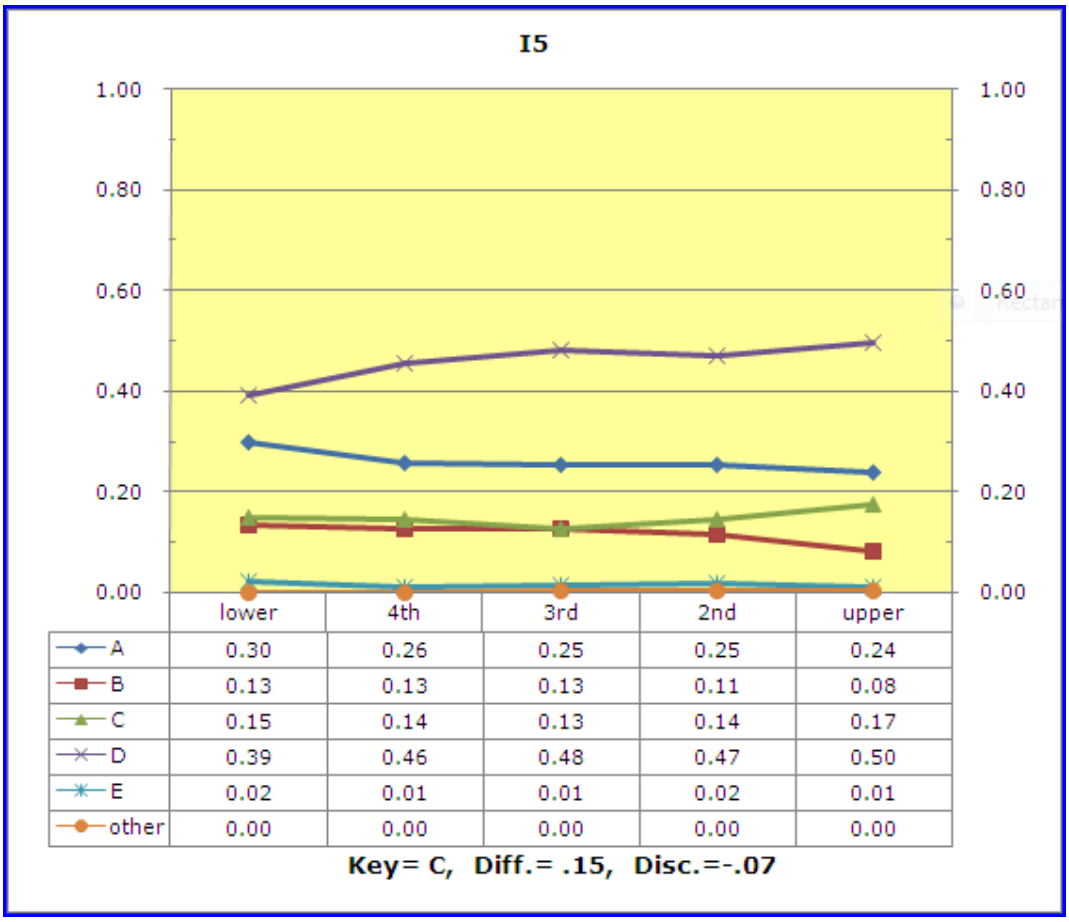
Larry Nelson, Curtin University

If I were asked to comment on the quality of the eight Jateng exams seen at the recent Coffs Harbour workshop, I would say this: if the tests are simply meant to provide information on how much the students knew, then all of the tests probably have value.

On the other hand, if the tests are meant to identify the strongest students, and the weakest ones, then I would say that two of the tests have questionable value: **SMA B. Indonesia**, and **SMK B. Indonesia**. The reliabilities of these two exams are low: 0.58, and 0.71, respectively.

It is commonly felt that an achievement test meant to discriminate among students, that is, to separate students into strong and weak groups, should have a reliability of at least 0.80 – some professional test developers say that this figure should, in fact, be closer to 0.90.

To see why the two **B. Indonesia** tests have limited value when it comes to being able to discriminate among students, consider the following chart of results for Item 5 on the **SMA B. Indonesia** test:



	lower	4th	3rd	2nd	upper
A	0.30	0.26	0.25	0.25	0.24
B	0.13	0.13	0.13	0.11	0.08
C	0.15	0.14	0.13	0.14	0.17
D	0.39	0.46	0.48	0.47	0.50
E	0.02	0.01	0.01	0.02	0.01
other	0.00	0.00	0.00	0.00	0.00

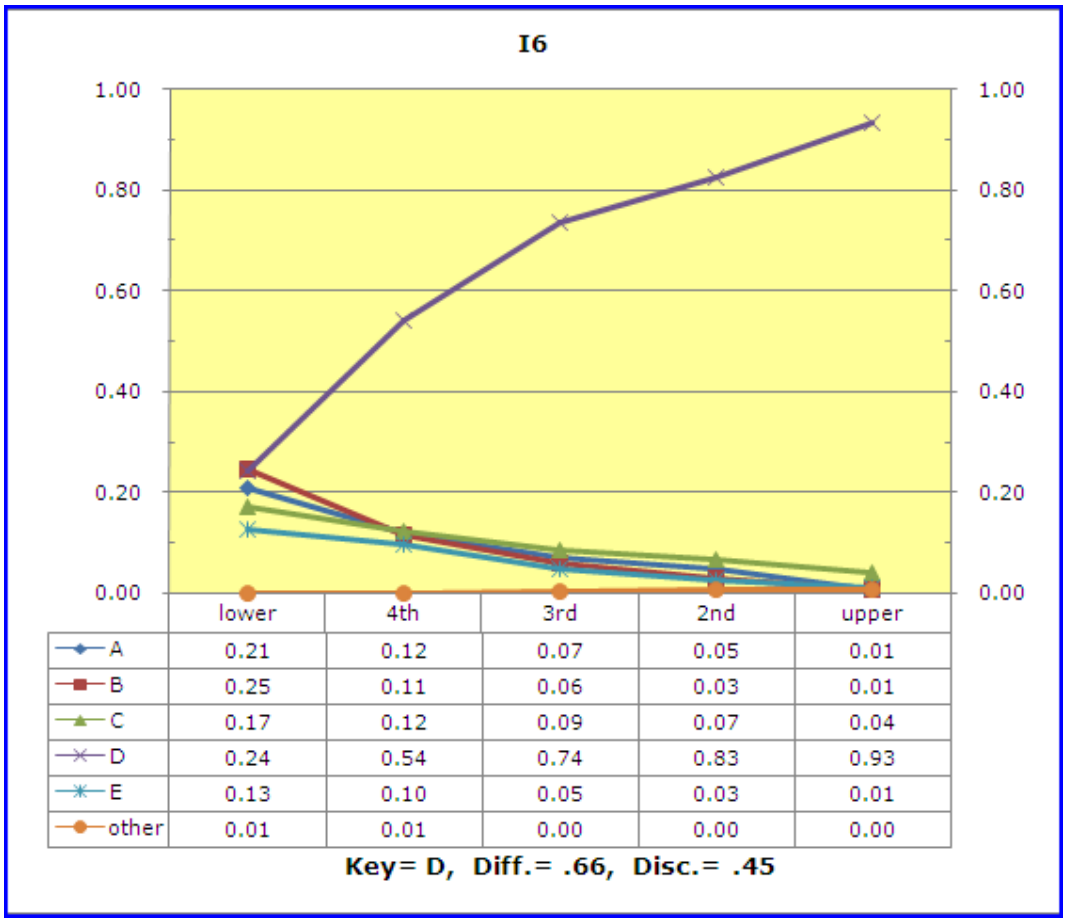
Key= C, Diff.= .15, Disc.=-.07

In the chart above there is one trace line for each item option, plus a line for those students who did not answer the item ('other'). The 'lower', '4th', '3rd', '2nd', and 'upper' labels refer to groups of students; 'lower' is the bottom group, those with the lowest test scores, while 'upper' is the top group, those with the highest test scores. The '3rd' group corresponds to those students whose test scores were in the middle of the results. The '4th' group has the next-to-lowest test scores, while the '2nd' group has the next-to-highest test scores.

The correct answer for Item 5 is 'C'. Look at the trace line for option C in the chart above. It is essentially flat. In the 'lower' group, 15% identified C as the correct answer. However, in the 'upper' group, the best students, only 17% identified C as the correct answer. Half of the 'upper' group felt that option D was the right answer, not C.

If the purpose of the test is to pick out the best students from the weakest ones, Item 5 does not work as wanted. When an item is discriminating, there should be a very noticeable difference as the trace line for the correct answer goes from left ('lower') to right ('higher').

As an example of a discriminating item, consider the chart for Item 6 on the **SMA Matematika** test:



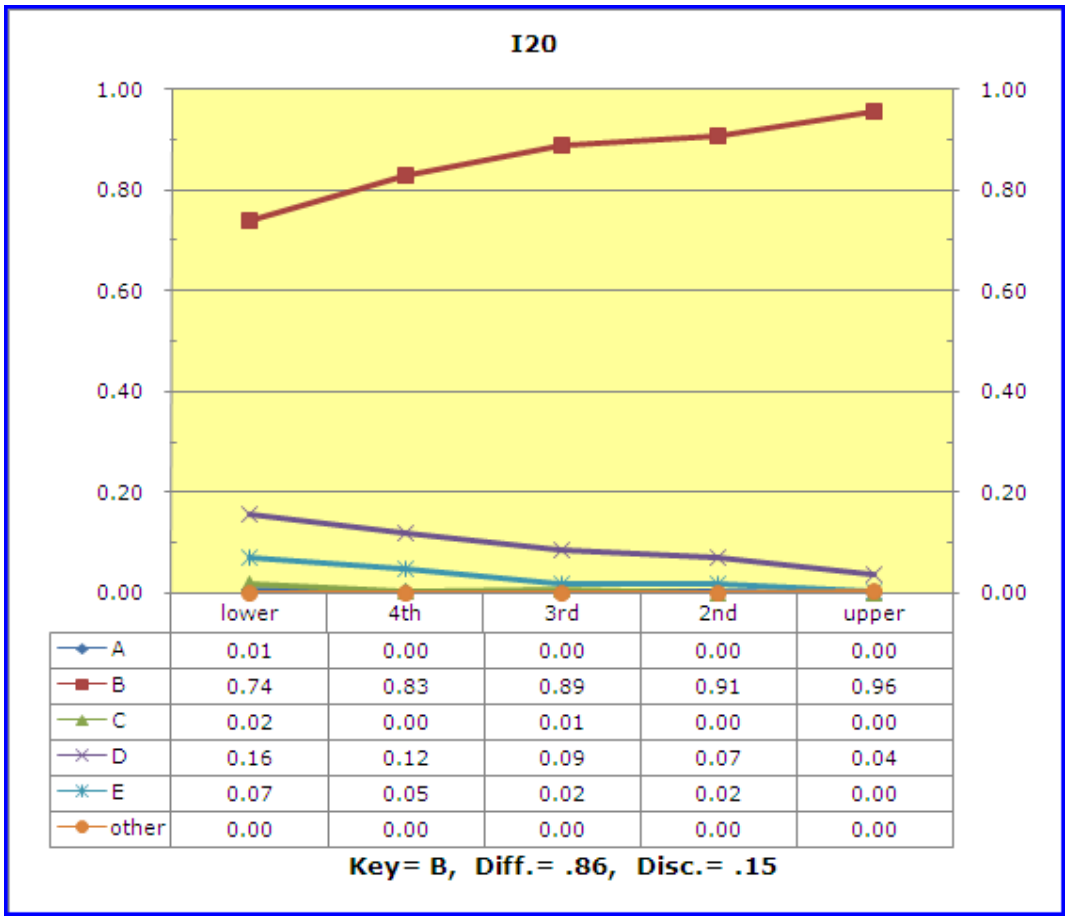
	lower	4th	3rd	2nd	upper
A	0.21	0.12	0.07	0.05	0.01
B	0.25	0.11	0.06	0.03	0.01
C	0.17	0.12	0.09	0.07	0.04
D	0.24	0.54	0.74	0.83	0.93
E	0.13	0.10	0.05	0.03	0.01
other	0.01	0.01	0.00	0.00	0.00

Key= D, Diff.= .66, Disc.= .45

The correct answer for Item 6 is 'D'. Look at the trace line for option D in the chart above. It rises sharply as it goes from left to right. In the 'lower' group, 24% chose option D. In the 'upper' group, 93% chose option D. This is the type of pattern expected of an item which is able to discriminate

among students, that is, an item capable of separating the strongest students from the weakest ones.

Most of the items in the **SMA B. Indonesia** test have trace lines which come close to being flat. The following chart is typical of the items in this test:



	lower	4th	3rd	2nd	upper
A	0.01	0.00	0.00	0.00	0.00
B	0.74	0.83	0.89	0.91	0.96
C	0.02	0.00	0.01	0.00	0.00
D	0.16	0.12	0.09	0.07	0.04
E	0.07	0.05	0.02	0.02	0.00
other	0.00	0.00	0.00	0.00	0.00

Key= B, Diff.= .86, Disc.= .15

While the trace line for Item 20's correct option does rise from left to right, the rise is not dramatic. In the 'lower' group, 74% got this item correct; in the 'upper group', 96% got the item correct. When items are discriminating, the percent correct in the 'lower' group should be around 20% or less, while in the upper group the percent correct should be close to 100%.

In order for a test to have a high reliability figure, all items in the test should be good at discriminating, they should all have a chart like that shown above for Item 6 in the **SMA Matematika** exam.

What are the practical consequences of low test reliability? Increased measurement error. A test with low reliability usually cannot be used as an accurate measure of student achievement. It is for this reason that I say, above, that 'two of the tests have questionable value'. The **SMA B. Indonesia** and **SMK B. Indonesia** exams have test items with relatively poor response charts; this is particularly true in the case of **SMA B. Indonesia**.

As a final comment, I could mention that high test reliability is not required when a test is not meant to discriminate among students. Teachers might be interested not in reliability, but in the percent-

age of students able to answer a test item correctly. The chart for Item 20 above indicates that the item did not discriminate among the students, but a teacher might not care about this if he, or she, just wanted to look at the percentage of students able to identify the correct answer. In the case of Item 20, 86% of all the students got the item right, a fact which might tell the teacher that the content area covered by the item had been mastered by the great majority of students. (The **Diff.=.86** figure at the bottom of the chart means that 86% got the item right.)

Mastery and criterion-referenced tests will very often have low reliability figures, but still be very useful tests.

I would be happy to provide more comments on the Jateng exams, time permitting. If the exams are meant to be mastery tests, for example, there are other statistics which could be computed to reflect on overall test quality.