Chapter 7

Interpreting Lertap Results for Cognitive Tests

Contents

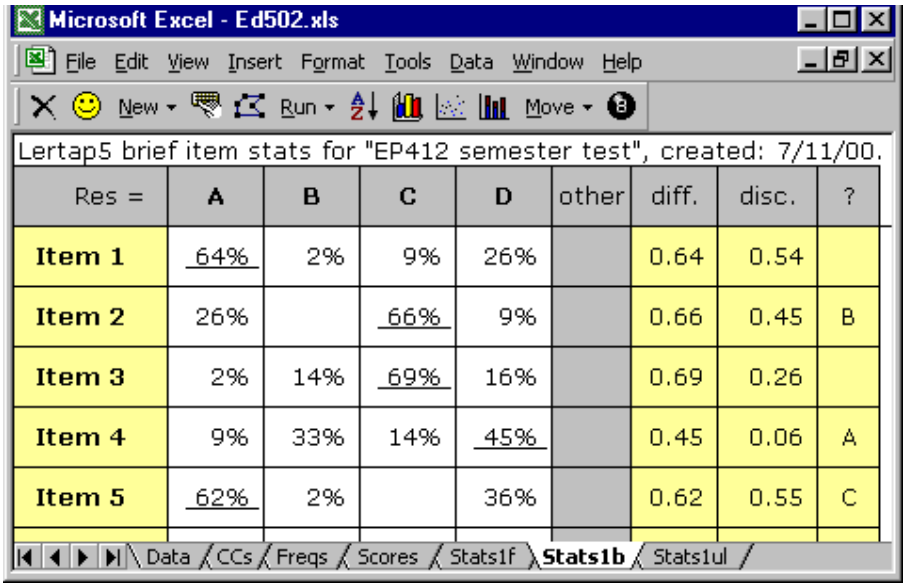
Well, there you are, and you've done it. You've digested all the technical, nitty-gritty material found in previous chapters, prepared and processed your Data and CCs sheets, used "Interpret CCs lines" and "Elmillon item analysis", and found that Lertap has produced its results sheets, worksheets with such beguiling titles as "Stats1b", "Stats1f", "Stats1ul", and "Scores".

How to make sense of all the information now at your fingertips is what this chapter gets into. We'll look at such questions as "How did the students do?"; "Was my test any good?"; "Can I fix the test so that it's better".

How did the students do?

Lertap thinks the results likely to be of most immediate interest are to be found in the brief stats sheet. A worksheet with a name such as "Stats1b" is usually where Lertap puts its focus after you've used the "Elmillon item analysis".

Let's take a look at the Stats1b results from Dirk Hartog's "EP 412, Theories of Learning" class test:



Lertap5 brief item stats for "EP412 semester test", created: 7/11/00.

Res =	A	B	C	D	other	diff.	disc.	?
Item 1	64%	2%	9%	26%		0.64	0.54	
Item 2	26%		66%	9%		0.66	0.45	B
Item 3	2%	14%	69%	16%		0.69	0.26	
Item 4	9%	33%	14%	45%		0.45	0.06	A
Item 5	62%	2%		36%		0.62	0.55	C

Data CCs Freqs Scores Stats1f **Stats1b** Stats1ul

The Stats1b sheet quickly indicates to Dr Hartog how well the students did on each test item. On the first five items, he notes that four were correctly answered by more than 60% of the class, a satisfactory outcome from his point of view[1]. Results from previous testings had indicated that Item 4 was the hardest of the first five items, and once again this turned out to be the case—less than half of the class, 45%, were able to pick out the best answer to this item. Dirk makes a mental note to have another look at this item later.

While thinking of making mental notes, he wishes he could have a printout of the Stats1b sheet. Can do? Of course. This is an Excel worksheet, and Excel has good printing capabilities. He clicks on Excel's File / Print option, then on the Preview button. He makes use of the Page Break Preview option, using his mouse to adjust the page break (this is not so important with the Stats1b sheet where each item's results fit on a single line—on other sheets the Page Break Preview is often very handy).

With his hard Stats1b copy in hand, a copy of the original test paper, and a cup of his favourite brew, Dirk sits down to make notes on his printout. Some of the items were harder than anticipated, more difficult than he had expected them to be. He wishes he had the chance to hit respective topics with the students again. If this were a formative test, he would do exactly that.

In fact, why don't we assume that this was a formative test, or maybe even a criterion-referenced one? If we do this, Dr Hartog might very well stop his analysis at this point, feeling he has sufficient information from the test—a summary of the percentage correct for each item. His review of the percentage figures has enabled him to identify items which seemed harder than he'd like, and his next action will be to make sure the items have no obvious errors to them—he will need to check that he's correctly entered the right answers on his *key card, too.

[1] Results which correspond to the correct answer to an item are underlined.

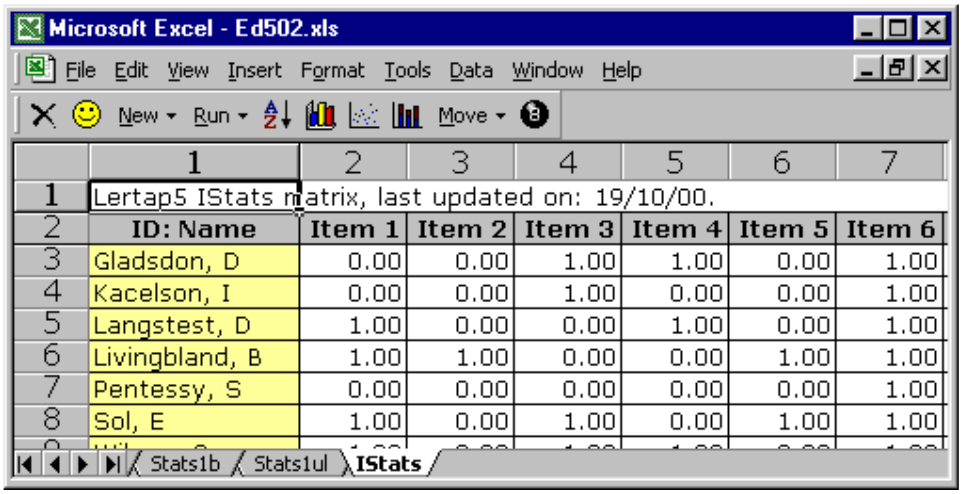
If all seems in order, he'll then review the topics corresponding to the items which students did poorly on, and go over them again with the class.

Can he really do this? He doesn't need to worry about test reliability? What about the other columns in the Stats1b report? He doesn't need to look at the "disc." index, and the "?" column?

He can certainly do this, to be sure, and he might very well. He doesn't need to look at the disc. and ? columns. At this point, he's looking at class-level data, wanting feedback on topics which seem to have been adequately mastered, looking for topics which appear to require additional coverage. Yes, he could quite comfortably and happily end his analyses at this point, no worries.

But let's suppose he wants to go on. Again, were the test a formative one, or a criterion-referenced one, he might want a quick summary of how each student did on each item of the test.

To get results for each student at an item level, the "Output item scores matrix" capability on Lertap's Run menu is useful. It produces a report such as the following:



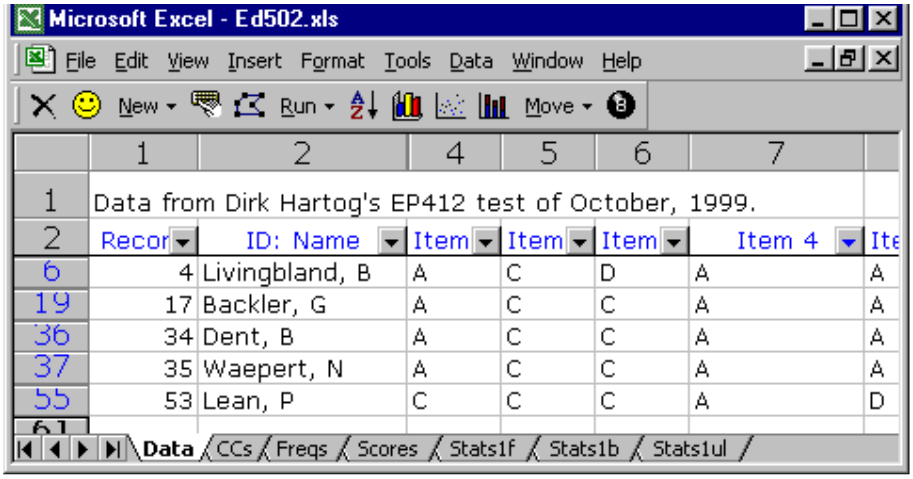
	1	2	3	4	5	6	7
1	Lertap5 IStats matrix, last updated on: 19/10/00.						
2	ID: Name	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
3	Gladsdon, D	0.00	0.00	1.00	1.00	0.00	1.00
4	Kacelson, I	0.00	0.00	1.00	0.00	0.00	1.00
5	Langstest, D	1.00	0.00	0.00	1.00	0.00	1.00
6	Livingbland, B	1.00	1.00	0.00	0.00	1.00	1.00
7	Pentessy, S	0.00	0.00	0.00	0.00	0.00	1.00
8	Sol, E	1.00	0.00	1.00	0.00	1.00	1.00

Stats1b Stats1ul IStats

A report such as this, found on the IStats worksheet, quickly indicates the number of points each student earned on each item. For example, Miss Gladsdon appears to have mastered three of the first six items on the test, Items 3, 4, and 6—her score of "1.00" for these items indicates that she got them right.

Such reports take a bit of time to digest when there are many test items, but criterion-referenced and mastery tests are often short, and in such cases the IStats report is very useful.

While we're working at the level of item responses, it's worth mentioning that Excel has an "autofilter" feature which can often be put to good advantage. For example, suppose Dirk Hartog wanted to find out who it was, exactly, that took option A on Item 4. He can go to the Data worksheet and use Data / Filter / AutoFilter to get a screen which looks like the following:



The autofilter options put small arrows in the column header row. The screen shot above resulted after Dirk clicked on the arrow next to Item 4, and then clicked on A from a list which appeared. Excel displayed all records having an A in the column. In this example, Dr Hartog became a bit concerned with what he found: of the five students selecting option A on Item 4, he knew that three of them were among the strongest in the class. This led him to think that he might do more than review topics with the class—he started to consider going over some of the test items with students in order to try and discover why the most capable sometimes chose one of the distractors.

Now, at this point, we have once again reached a spot where Dr Hartog could pack up his bags, as far as Lertap goes. He has thus far expressed an interest in looking at item-level data. He first wanted to know how many people correctly answered each item, and he used the Stats1b report for this. Then he wanted to see individual results per item, for which he turned to the IStats worksheet. This latter worksheet, IStats, is produced by using Lertap's Run, "Output item scores matrix", option; it is not one of the reports produced by using "Elmillon item analysis".

What about digging deeper? Let's say we wanted to advise Dirk on the quality of his test, from a measurement, or psychometric, point of view. If he wanted to use the test scores which Lertap produces, would he be justified in doing so? Does the evidence suggest that the test results are sufficiently free of error?

Was my test any good?

The assessment of the quality of a test generally involves determining its reliability and validity. However, the questions we posed above exemplify a useful fact: much can be gained by looking at results on an item by item basis. If an instructor believes that the questions she's used in her test validly (that is, truly and fairly) reflect her teaching and learning objectives, she can often draw important pedagogical observations, and conclusions, by looking at, for example, the number of students getting each question correct. Data analysis of this sort may lead her to redouble her efforts in certain topic areas, and to review selected

items with students in order to find out more about the thinking behind their answers.

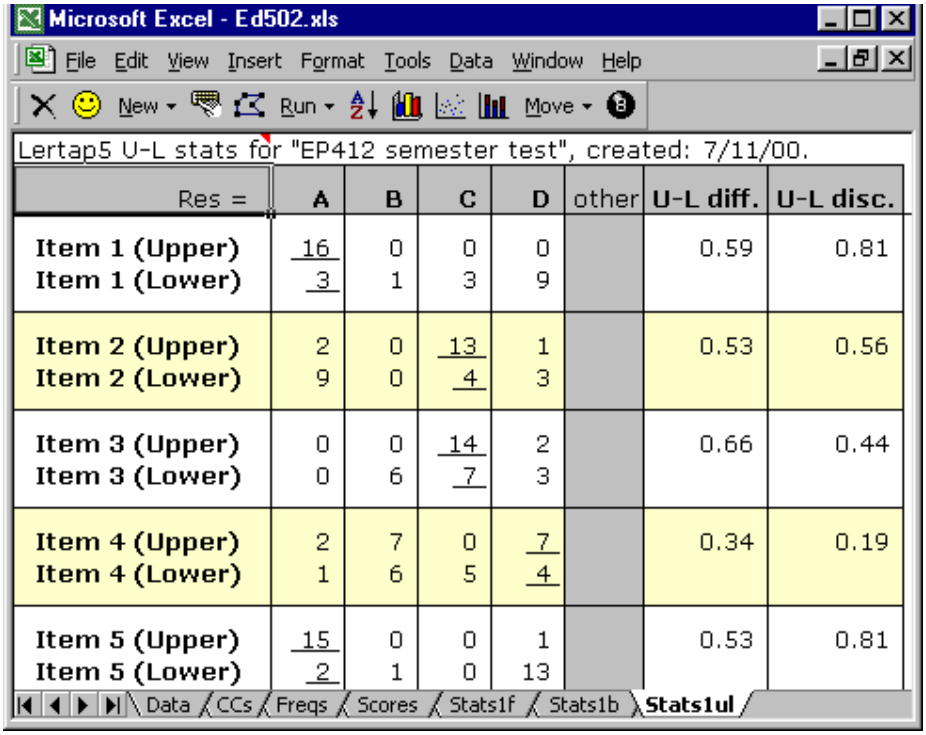
But there are many times when instructors wish to do more with Lertap's results.

A common objective is to devise, or identify, a test which will allow an instructor to differentiate among the students, separating those who have achieved at a satisfactory level from those who have not. In order to accomplish this, a test's items are expected to do the same on an individual basis—to help ferret out those who've got the goods from those still looking for them.

All of Lertap's statistical reports sheets can aid in the process of indexing item and test quality. They do so by using two different approaches, ones which have been developed over considerable time to become some of today's standard measurement tools.

The U-L method

The upper-lower method of assessing item quality is reflected in the Stats1ul report, an example of which is shown below:



Res =	A	B	C	D	other	U-L diff.	U-L disc.
Item 1 (Upper)	16	0	0	0		0.59	0.81
Item 1 (Lower)	3	1	3	9			
Item 2 (Upper)	2	0	13	1		0.53	0.56
Item 2 (Lower)	9	0	4	3			
Item 3 (Upper)	0	0	14	2		0.66	0.44
Item 3 (Lower)	0	6	7	3			
Item 4 (Upper)	2	7	0	7		0.34	0.19
Item 4 (Lower)	1	6	5	4			
Item 5 (Upper)	15	0	0	1		0.53	0.81
Item 5 (Lower)	2	1	0	13			

In this sort of analysis, Lertap forms two groups—those who have done well on the criterion, and those who have not. It calls the first group the "upper" one, while the second is labelled "lower".

The standard criterion for determining upper and lower levels is the test score, which is called an "internal" criterion.

These are the steps the program goes through in order to define the groups: (1) it computes a test score for each student; (2) it sorts these scores from highest to lowest; (3) it picks off the top 27% of the results, and stores them in a

column of a hidden worksheet called "ScratchScores"; (4) finally, it picks off the bottom 27% of the results, and stores them in another column of this hidden worksheet. That's it—the two groups have been defined, and are ready for use by Elmillon, Lertap's item analysis program.

When Elmillon comes in, it looks in the upper group to find the number of students who chose option A on Item 1, how many chose option B on Item 1, how many selected option C, and how many went for the last option, D. It then does the same in the lower group, finding how many selected each of the four options for Item 1. Once it has this information, it determines the U-L diff. figure by adding up the number in both groups who selected the right answer, and dividing this figure by the total number of students in both groups. In this example, the correct answer to Item 1 is A; 16 students in the upper group selected this option, as did 3 in the lower group—this gives 19. There are 16 people in the upper group, and 16 in the lower—this gives 32. Divide 19 by 32 to get 0.59, the proportion of students in the two groups who identified the correct answer.

The U-L disc. figure is then derived for Item 1. It's the proportion in the upper group who answered the item correctly (16/16), less the proportion in the lower group who answered the item correctly (3/16). This gives 13/16, or 0.81, the value seen above in the worksheet.

What to make of these results? Well, what's the question we want to answer at this point? It's: Did Item 1 work as wanted?

What was wanted? Items which could help identify which students did well on the criterion measure, and which students did not. If an item is successful in this task, we would expect to find that most of the strong students are able to discern the correct answer to the item, while the others, those in the lower group, are not. If this happens, we say that the item is discriminating, that is, providing us with a means which we can use to distinguish who has, and who has not, done well on the criterion.

Was Item 1 a good one? Yes. All of the strongest students got the item right, and most of the weaker students did not.

A key to having good items is to have distractors which effectively work as foils to draw off the weaker students. We want the distractors to fool the bottom group, but not the top one. In this sense, Item 1 again comes through well. Those in the top group were not foiled by the distractors, but most of those in the lower group were. Option D seems to have been a particularly good distractor as 9/16, or 56%, of the weaker students went for it.

Item 2 wasn't quite as good. Two of the three distractors, A and D, fooled a few people in the top group. Option B was not an effective distractor at all—no-one selected it. The proportion in the top group who got Item 2 right is 0.81 (13/16), while the proportion in the lower group is 0.25 (4/16), giving a U-L disc. value of 0.56.

Let's jump down to look at Item 4. Here a substantial number of students in both groups were effectively distracted by option B, and the proportion in the upper group who identified the correct answer was low, 0.44 (7/16). We would say that Item 4 did not discriminate well in this group of students.

What does the literature say about U-L indices?

Ebel & Frisbie (1986, Chapter 13) state that any item with a U-L disc[2]. value below .20 is a poor item, "to be rejected or improved by revision". "Very good items" will have values of 0.40 and up.

Hopkins (1998, Chapter 10) agrees that items with a U-L disc. of 0.40 provide "excellent discrimination", but suggests that this index may go as low as 0.10 and still indicate "fair discrimination".

Linn & Gronlund (1995, Chapter 12) do not suggest a minimum value for the U-L disc. figure, stating that a "low index of discriminating power does not necessarily indicate a defective item".

Mehrens & Lehmann (1991, Chapter 8) write "In general, a discrimination index of 0.20 is regarded as satisfactory".

Oosterhof (1990, Chapter 13) states "On teacher-constructed tests, an item discrimination about 20% is generally considered sufficient. An item discrimination index above 40% is quite high and is equivalent to the level of discrimination found on many commercially developed tests".

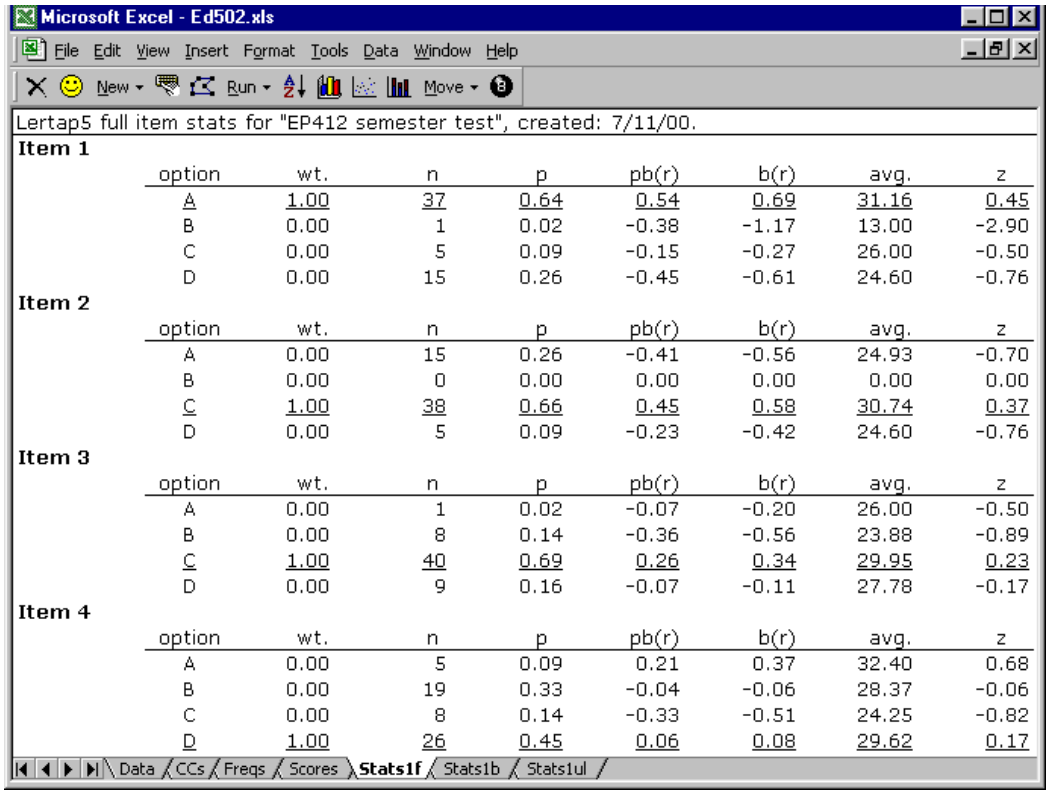
What about the item difficulty index, U-L diff.? There is a relationship between the two U-L indices, diff. and disc. U-L disc. values will not be high unless diff. values are at a certain level. Thus, looking at the disc. index is sometimes sufficient: if the disc. value is good, the diff. level will likely be adequate too. Some authors do provide desirable levels for diff. figures. Mehrens & Lehmann (1991, p.164) suggest these "ideal average difficulty" figures for a "maximally discriminating test": for multiple-choice (M-C) items with five options, diff. should be about 0.70; for M-C items with four options, diff. values should be around 0.74; M-C items with three options should have diff. values around 0.77, while true-false items should have diff. values around 0.85. Hopkins (1998, p.257) states "The maximum measurement of individual differences by an item is at a maximum when the item difficulty level is .5 ....". Allen & Yen (1979, p.121) write "...for a four-option multiple-choice item ... the optimal difficulty level is about .60 .... Generally, item difficulties of about .3 to .7 maximize the information the test provides about differences among examinees...."

The correlation method

Correlation coefficients have long played a prominent role in tests and measurement. Their job is to express the degree to which two "variables" are related. In the context of our present discussion, the variables are item responses and the criterion measure, the subtest score. When we use a correlation coefficient in this setting, we're asking if the people who did well on an item also did well on the criterion.

[2] Ebel & Frisbie, and most other texts, call the U-L item discrimination index "D".

Lertap likes to compute correlation coefficients. In its full stats reports, such as seen in the Stats1f worksheet, Lertap uses two correlation coefficients to signal the extent to which item responses correlate with the criterion score (the subtest score). Have a look:



Microsoft Excel - Ed502.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 full item stats for "EP412 semester test", created: 7/11/00.

Item 1

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	1.00	37	0.64	0.54	0.69	31.16	0.45
B	0.00	1	0.02	-0.38	-1.17	13.00	-2.90
C	0.00	5	0.09	-0.15	-0.27	26.00	-0.50
D	0.00	15	0.26	-0.45	-0.61	24.60	-0.76

Item 2

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	15	0.26	-0.41	-0.56	24.93	-0.70
B	0.00	0	0.00	0.00	0.00	0.00	0.00
C	1.00	38	0.66	0.45	0.58	30.74	0.37
D	0.00	5	0.09	-0.23	-0.42	24.60	-0.76

Item 3

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	1	0.02	-0.07	-0.20	26.00	-0.50
B	0.00	8	0.14	-0.36	-0.56	23.88	-0.89
C	1.00	40	0.69	0.26	0.34	29.95	0.23
D	0.00	9	0.16	-0.07	-0.11	27.78	-0.17

Item 4

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	5	0.09	0.21	0.37	32.40	0.68
B	0.00	19	0.33	-0.04	-0.06	28.37	-0.06
C	0.00	8	0.14	-0.33	-0.51	24.25	-0.82
D	1.00	26	0.45	0.06	0.08	29.62	0.17

Data CCs Freqs Scores Stats1f Stats1b Stats1ul

The pb(r) and b(r) columns have the correlation coefficients, with "pb(r)" being a point-biserial correlation, and "b(r)" the biserial equivalent. These coefficients answer this question: How did the people who selected an item option do on the criterion measure? If they did well on the criterion, both pb(r) and b(r) will be "high", where high may be taken as anything over 0.30 for pb(r), and anything over 0.40 for b(r)[3].

With these figures as guidelines, what do we make of Item 1? Those 37 people who took option A, the right answer, did well on the criterion measure. Those who selected the other options, the distractors, did not do as well—their correlations with the criterion measure are negative. Note how the "avg." column confirms this: the average criterion score for the 37 people who got the item correct was higher than the average criterion score for those who chose one of the distractors. The signs of the z-scores reflect the signs of the correlation coefficients; these z-scores indicate how far the "avg." figure for each option was from the criterion score's mean, with negative z-scores being below the criterion mean.

If an item is discriminating, helping to identify who's strong on the criterion measure and who's not, the pb(r) and b(r) figures for the correct answer will be high, while corresponding figures for the distractors will be negative. The avg. value for the right answer will be higher than the avg. values for the distractors,

[3] It is possible for biserial values to have a magnitude greater than 1.00.

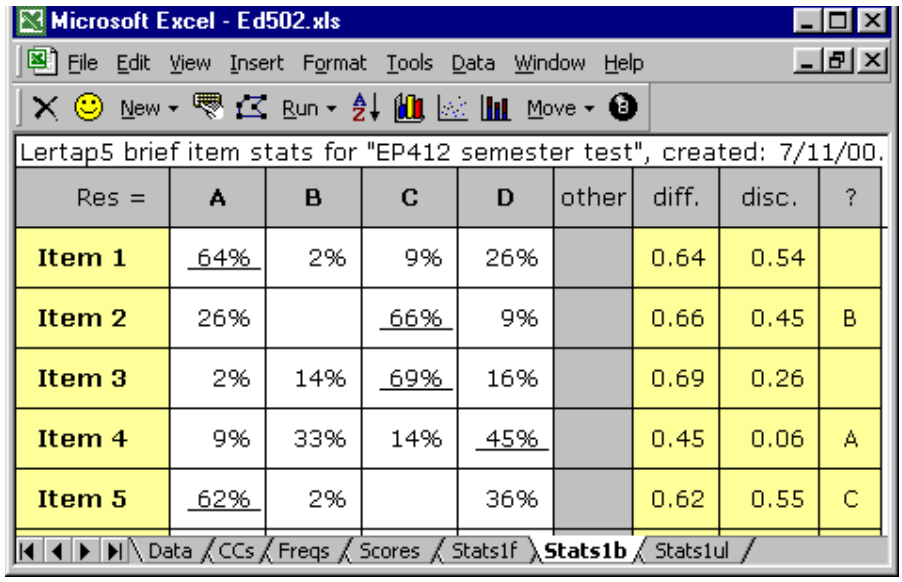
and all of the z-scores should be negative, except for one, that pertaining to the correct answer.

What about Item 2? All is well except for option B. A distractor's job in life is to fool people, and option B fooled no-one. Otherwise the item's options show the desired pattern for this sort of analysis: the correct answer has high pb(r) and b(r) values, and its avg. figure is higher than the others. The z-scores for the distracting distractors (options A and D) are negative, which is good.

Item 3? Not too bad. The signs on the correlation coefficients and z-scores are what we want: positive for the correct answer, negative for the distractors. The actual pb(r) and b(r) values are not as high as we might desire; the differences between the avg. values are not as great as those seen in the first two items, but, overall, the pattern isn't too bad.

Item 4 is not as blessed. One of its distractors, option A, has a better pattern than does the correct answer, option D. The avg. criterion score for the five students who selected option A is high—not many people took this option, but those who did were among the strongest of the students—an undesirable outcome for a distractor. This item should be flagged as having a problem to be investigated. Interviewing students will often help to uncover the causes of a problem such as this.

A full statistics report is what we're looking at here, we've got lots of information to look at. It can get to be too much, at times—we might long for something more concise. Enter stage left the corresponding "brief" stats summary:



Res =	A	B	C	D	other	diff.	disc.	?
Item 1	64%	2%	9%	26%		0.64	0.54	
Item 2	26%		66%	9%		0.66	0.45	B
Item 3	2%	14%	69%	16%		0.69	0.26	
Item 4	9%	33%	14%	45%		0.45	0.06	A
Item 5	62%	2%		36%		0.62	0.55	C

Now each item's performance is summarised in a single line. We see the percentage of students who selected each option, can tell if anyone missed out the item (did not respond—signalled by the "other" column), get an index of item difficulty and discrimination, and have a quick analysis of how the distractors functioned.

The diff. figure is the proportion of students who got the answer right. The disc. is the value of pb(r) for the right answer. The ? column indicates if one or more of the item's distractors seemed to fail. A distractor will get an entry in this

column if it wasn't selected by anyone, or if it has a positive pb(r) figure, which means that it was selected by strong students—remember, we don't want the best students to pick the distractors—when they do, Lertap makes note of this by showing the distractor in the ? column.

Lertap allows an item have more than one right answer[4]. It takes the number of right answers to an item as the number of wt. values greater than zero. To this point there's been only one right answer for each item, which is the usual case. When there happen to be multiple right answers, the diff. value is the proportion of students who got the item right, counting over all options having wt.>0, and the disc. value becomes the Pearson product-moment correlation between the item and the criterion. Chapter 10 has more details on the computation of these statistics.

What does the literature say about the correlation method?

It is common to find that the literature will refer to pb(r), the point-biserial correlation coefficient, as the index of item discrimination. Hopkins (1998, footnote on p.270) refers to it as the "standard index", and to the U-L disc. figure as being a sort of "shortcut" to indicating how an item is discriminating. Hambleton, Swaminathan, & Rogers (1991, p.19) refer to pb(r) as the "classical item discrimination". Popham (1978, p.107) refers to pb(r) as perhaps the "most common" index of item discrimination. Haladyna (1994, p.146) writes: "In classical test theory, item discrimination is simply the product moment (point-biserial) relationship between item and test performance".

There are relationships among Lertap's three discrimination indices, U-L disc,, pb(r) and b(r). Hopkins, Stanley, & Hopkins (1990, footnote on p.270) write that U-L disc. values "have been shown to be almost perfectly linearly correlated with biserial coefficients". In turn, the relationship between pb(r) and b(r) is also strong. Lord & Novick (1968, p.340) state that "the point biserial correlation ... is equal to the biserial correlation multiplied by a factor which depends only on the item difficulty". pb(r) values, Lord & Novick go on to state, are "never as much as four-fifths of the biserial" (also p.340). From Magnusson (1967, p.205), we read: "... if the two methods are applied to the same set of data, the biserial coefficient will exceed the point-biserial coefficient by 25% ....". In other words, pb(r) and b(r) are highly related, with b(r) always giving a higher figure than pb(r).

What about acceptable minimum levels for pb(r)? Hills (1976, p.66), writing about the use of pb(r), states that "... many experts look with disfavor in items with correlation discrimination values less than +.30. Teachers will often be not as good at writing items as experts are, however, and acceptable items may have discrimination values in teacher-made tests as low as +0.15 ....". From the citations above regarding the relationship between pb(r) and b(r), we could conclude that Hills might suggest that b(r) values should be +.40, or so, for items authored by "experts", and perhaps may go as low as +.20, or so, for teacher-created items. A disadvantage to using b(r) values is that, unlike conventional product-moment based coefficients, of which pb(r) is one, b(r) coefficients may exceed 1.00 in magnitude.

[4] The *mws card is used to multiply-key items; see Chapter 5.

Which of these methods is best?

Of Lertap's three reports for cognitive tests, which is to be preferred? Given the full statistics seen in sheets such as Stats1f, the brief equivalents in Stats1b, and the U-L results in Stats1ul, is one of these better than the others for assessing the quality of a test's items?

In part the answer to this question depends on the purpose of the test. If the intent is to have an instrument which does best in discriminating among students, then all three sheets will be useful, and can be expected to lead to very similar conclusions. In this situation, which sheets a user drinks his or her coffee with will likely be a matter of personal preference. The use and interpretation of U-L statistics is covered in many texts—these statistics are straightforward and generally perceived as easy to understand. On the other hand, users who are comfortable with correlation methods might be expected to find their home in the full statistics report, or, if they're content with a more concise summary, in the brief statistics companion.

Of course there are times when a test is used not to try and spread students apart, but to assess their performance with regard to some sort of pre-defined standard, or benchmark. Criterion-referenced and mastery tests fall into this category. It is sometimes the case that items on such tests will be either very easy or very hard, and, when this happens, the correlation-based indices will come undone as they depend on having items with middle-level difficulties. In this situation the Stats1f and Stats1b reports may not have much to say, or, worse, may paint a bleak picture (in which case they should be ignored—the correlation-based results might be bleak, but this doesn't mean that the items are necessarily faulty, not at all—in this type of testing some users wouldn't even look at the 1f and 1b worksheets).

Reliability

For some pages now we've been talking about item quality. Earlier in this chapter we mentioned that there are times when test users will not need to move beyond the item level—their main focus is at the item level, and instructors will use item performance statistics to reflect on what they seem to say as regards their teaching. Teachers may also use item-level results to see how individual students did on each test question, perhaps thinking of identifying instructional strategies which will assist each student in areas where s/he has shown a need.

When we talk about test reliability, and validity, we move to a different level. We turn to an inquiry which has to do not with item results, but with the overall test score, with the composite result made by adding together item scores.

How does Lertap measure a test's reliability? It depends on the type of analysis which has been requested. In the conventional situation, Lertap provides coefficient alpha, and the standard error of measurement, as indicators of precision.

Coefficient alpha, also known as Cronbach's alpha, is an index of internal consistency, an indicator of how well item responses intercorrelate. If a test's items correlate well with each other, alpha will be high. The maximum value which alpha can assume is 1.00.

What's a high reliability figure, in practical terms? Hopkins (1998, p.131) says that values of .90 or above can be expected of professionally-developed tests. Linn and Gronlund (1995, p.106) state that "teacher-made tests commonly have

reliabilities between .60 and .85", good enough for what they call "lesser decisions ... useful for the ... instructional decisions typically made by teachers". An interesting statement can be found in Kaplan & Saccuzzo (1993, p.126): "For a test used to make a decision that affects some person's future, you should attempt to find a test with a reliability greater than .95."

A statistic which is just as useful as the reliability coefficient, if not more so, is the standard error of measurement, SEM:

$$SEM = s.d.\sqrt{1 - alpha}$$

In this equation, s.d. is the standard deviation of the test scores.

To show how the SEM is used, look at these results:

reliability (coefficient alpha):	0.73
standard error of measurement:	2.81

The SEM is a practical index of the precision, or accuracy, of the test scores. It's commonly used like this: add and subtract the value of the SEM from a student's score to get a range which indicates where the student's score might fall if our test was perfectly reliable.

Take, for example, a student with a test score of 78. Adding and subtracting 2.81 from this score, and rounding, gives a range of 75 to 81. In the nature of this business, we then say we have a "confidence interval" within which we believe we may find the student's real, or "true", test score, were our test completely reliable.

Here we've added and subtracted one SEM, and the resultant confidence interval is said to be the 68% interval, so called because we've gone one standard error on either side of the student's test score, because we assume these errors to be normally distributed, and because we know that the area under the normal curve, from one standard error below the mean to one standard error above, is 68%. If we wanted a 95% confidence interval, we'd add and subtract two SEMs, getting a range of 69 to 84 in this example.

Note what would happen if we'd set a score of 80 as the minimum test score required to get an "A" grade on this test. If we were naive enough to believe that our test was free of error, was completely reliable, we'd not give a student with a score of 78 a grade of A. This *is* naive—our tests do have errors, they're not 100% accurate, they're not perfectly reliable—withholding an A from someone with a score of 78 would not stand up to challenge.

Many texts expand on the use of the standard error of measurement and confidence intervals. For a particularly good discussion, see Linn and Gronlund (1995, pp.93-98).

Coefficient alpha is a popular index of test reliability. It's very similar to two others: KR-20 and KR-21. The KR indices stem from the work of Kuder and Richardson, work described in just about every test and measurement book we know of (for example, see Ebel & Frisbie, 1986, pp.77-78; Hopkins, 1998, pp.127-129; Linn & Gronlund, 1995, p.89; Mehrens & Lehmann, 1991, p.256; and Oosterhof, 1990, pp.55-56). Alpha is a better index than the KR ones because it allows for items to have more than one right answer. When all items

have only one correct answer, and give the same number of points for the correct answer, then alpha and KR-20 produce exactly the same result. KR-21 is an approximation to KR-20, easy to calculate, but, as noted by Ebel & Frisbie (1986, p.78) it usually "gives an underestimate of the reliability coefficient".

If a test has a high value for coefficient alpha (or KR-20, for that matter), some want to say that it means the test is homogeneous, by which they often mean that it's measuring just one thing, a single factor, a single concept. This is wrong. Here we could do no better than cite Pedhazur & Schmelkin (1991, p.102): *"An instrument that is internally consistent is not necessarily homogeneous"*. High alpha values do not mean that a test's items are all measuring the same thing.

The relationship between reliability and item statistics

The reliability figure we've been talking about is one which is based on item intercorrelations. If an item tends to have high correlations with the other items on the test, and has good discrimination, alpha reliability will be high.

The item difficulty and item discrimination bands found towards the end of the full statistics report (Stats1f), and the ? column seen in the brief statistics sheet (Stats1b), are intricately related to the value of alpha. For a test to have a high alpha figure, the item difficulty bands should have their entries in the .40 to .70 levels, the item discrimination bands should have all of their entries at or above the .30 level, and there should be no entries in the ? column. In the little "alpha figures" table which appears at the end of the full statistics report, there should be no positive values under the ***change*** column.

We could shorten this message: high alpha values are likely to result when items have good discrimination figures. When this is the case, item difficulties are likely to be good, and a review of distractor performance is likely to show that the great majority of them are functioning as wanted.

What about the "criterion-referenced" case?

In criterion-referenced testing (CRT), the focus is on measuring the performance of students in terms of a "clearly defined and delimited domain of learning tasks" (Linn & Gronlund, 1995, p.16). Criterion-referenced tests are "constructed to yield measures that are directly interpretable in terms of prespecified performance criteria" (Hopkins, 1998, p.175).

There are several features in Lertap which are useful in such testing situations.

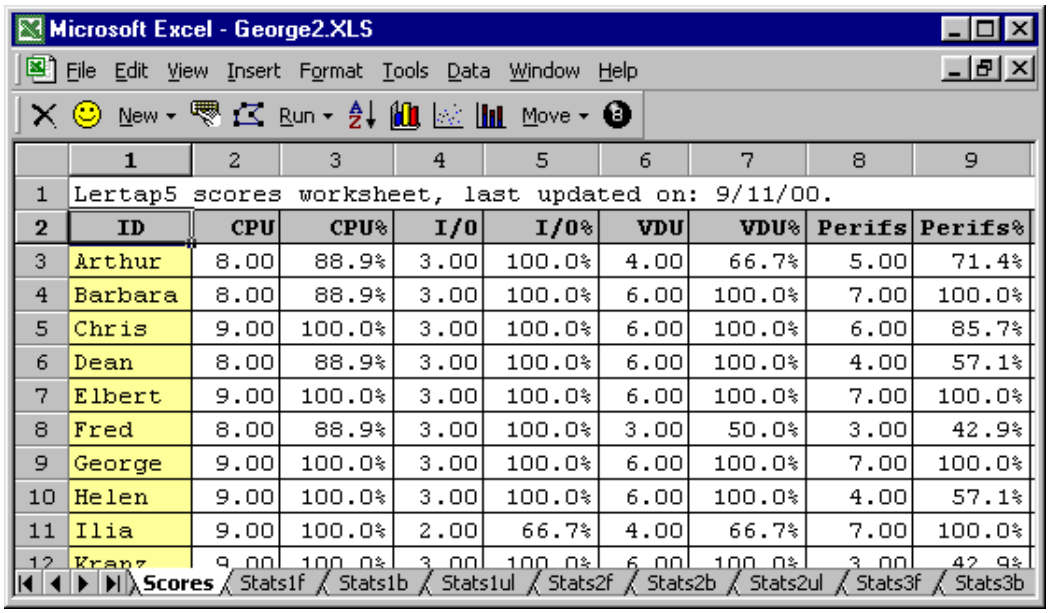
To begin, CRT users will want to have Lertap report scores in terms of percentage correct, as opposed to number of items right. This is easy to do: as mentioned in Chapter 5, using the control word "PER" on a *sub card will see to it that Lertap computes and displays percentage scores. For example, the following card will have Lertap display percentage figures for a test which has something to do with computer-assisted learning, or CAL:

```
*sub name=(CAL history), title=(CALHist), PER
```

Lengthy CRT instruments often have subsets of test items measuring in distinct areas. In Lertap, each of these areas will become a subtest. Consider the following[5]:

```
*sub res=(1,2,3,4,5), name=(CPU components), title=(CPU), PER
*sub res=(1,2,3,4,5), name=(I/O devices), title=(I/O), PER
*sub res=(1,2,3,4,5), name=(VDU characteristics), title=(VDU), PER
*sub res=(1,2,3,4,5), name=(Peripheral devices), title=(Perifs), PER
```

These four *sub cards might be found in a criterion-referenced test used to assess knowledge of computer components. Lertap's Scores report would look like this:



	1	2	3	4	5	6	7	8	9
1	Lertap5 scores worksheet, last updated on: 9/11/00.								
2	ID	CPU	CPU%	I/O	I/O%	VDU	VDU%	Perifs	Perifs%
3	Arthur	8.00	88.9%	3.00	100.0%	4.00	66.7%	5.00	71.4%
4	Barbara	8.00	88.9%	3.00	100.0%	6.00	100.0%	7.00	100.0%
5	Chris	9.00	100.0%	3.00	100.0%	6.00	100.0%	6.00	85.7%
6	Dean	8.00	88.9%	3.00	100.0%	6.00	100.0%	4.00	57.1%
7	Elbert	9.00	100.0%	3.00	100.0%	6.00	100.0%	7.00	100.0%
8	Fred	8.00	88.9%	3.00	100.0%	3.00	50.0%	3.00	42.9%
9	George	9.00	100.0%	3.00	100.0%	6.00	100.0%	7.00	100.0%
10	Helen	9.00	100.0%	3.00	100.0%	6.00	100.0%	4.00	57.1%
11	Ilia	9.00	100.0%	2.00	66.7%	4.00	66.7%	7.00	100.0%
12	Kranz	9.00	100.0%	3.00	100.0%	6.00	100.0%	3.00	42.9%

Scores | Stats1f | Stats1b | Stats1ul | Stats2f | Stats2b | Stats2ul | Stats3f | Stats3b

In this simple example, using PER and multiple subtests has resulted in a report which concisely profiles student performance. Elbert has done well in all areas; Fred is weak in VDU and Perifs.

[5] Here response options consisted of five digits, hence res=() declarations are necessary. Note that each of the *sub cards had a *col card before it, and a *key card after it—these are not shown.

At the item level, CRT users often like to contemplate response frequencies, and item difficulty. The Stats1b report is useful in this regard:

Microsoft Excel - George2.XLS

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 brief item stats for "CPU Components", cre

Res =	1	2	3	4	5	other	diff.
CPU1	96%		3%		1%		0.96
CPU2	5%	24%	1%	1%	69%		0.69
CPU3	92%	1%	3%	3%	1%		0.92
CPU4	14%	75%	2%	1%	8%		0.75
CPU5	13%	4%		76%	6%	1%	0.76
CPU6	3%	87%	4%	2%	3%	1%	0.87

Scores Stats1f **Stats1b** Stats1ul Stats2f Stats2b St

The Stats1b report shows that the class did well on items CPU1 and CPU3, but there may be a problem in the area addressed by item CPU2. Note that here we have intentionally omitted the two last columns of the brief stats sheet, disc. and ?. Item discrimination is not always wanted in the CRT case, but, when it is, Lertap provides it. Here, for example, is the Stats1ul report:

Microsoft Excel - George2.XLS

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "CPU Components", created: 9/11/00.

Res =	1	2	3	4	5	other	U-L diff.	U-L disc.
CPU1 (Upper)	27	0	0	0	0		0.94	0.11
CPU1 (Lower)	24	0	2	0	1			
CPU2 (Upper)	0	0	0	0	27		0.70	0.59
CPU2 (Lower)	3	11	1	1	11			
CPU3 (Upper)	27	0	0	0	0		0.85	0.30
CPU3 (Lower)	19	1	3	3	1			
CPU4 (Upper)	0	27	0	0	0		0.70	0.59
CPU4 (Lower)	10	11	1	1	4			
CPU5 (Upper)	0	0	0	27	0		0.70	0.59
CPU5 (Lower)	8	2	0	11	5	1		

Scores Stats1f Stats1b **Stats1ul** Stats2f Stats2b Stats2ul Stats3f

Only one of the five items shown in the table above, CPU1, is not discriminating.

The top group does well on all five items, but the weaker students displayed adequate performance only on item CPU1. (A question for readers: Why don't the U-L diff. values equal the Stats1b diff. values?[6])

Would we say that item CPU1 is a bad item? No, not necessarily, not in the CRT case. The results indicate that almost everyone mastered the content that it tests for, which might be a pleasant finding indeed. In the CRT case, the question for items is not always how well they discriminate, but how well the students did on them. CRT items which everyone gets right, or wrong, are not discriminating items, but they're invaluable in indicating what students know, or don't know.

The mastery case

Sometimes our cognitive tests are meant to determine who has mastered a content area, and who has not. The examples we've just worked through on criterion-referenced testing are 100% relevant to this objective, but we now step up the action a bit by saying we want to use a cutoff score, a minimum score which students must reach in order to be said to have mastered the material on which they've been tested.

Lertap provides support for mastery test analyses by (1) computing Brennan's (1972) generalised index of item discrimination; (2) undertaking a Brennan-Kane (1977) variance components analysis to derive estimates of test dependability and measurement error; and (3), computing an index of classification consistency, using a procedure recommended by Peng and Subkoviak (in Subkoviak,1984).

How to activate a mastery test analysis? Use the Mastery control word on a *sub card, as seen here:

```
From EP 412 class of Semester 2, 1999.
*COL (C4-C48)
*SUB Name=(412 semester test), Title=(EP412), Mastery
*KEY ACCDA CCDBB DBCDB ADDAC BDDCC BDCBA BCCBD CBCCC ACDCA
```

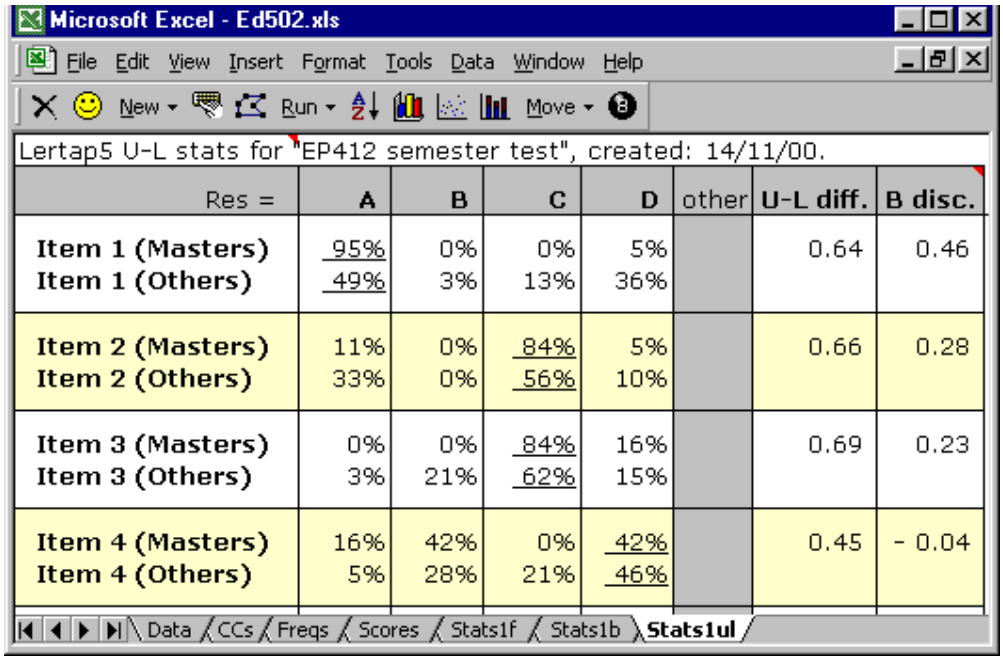
These cards will produce a mastery analysis with the cutoff score set at 70%, Lertap's default value. To set the cutoff score at another level, change the *sub card as exemplified below:

```
*SUB Name=(412 semester test), Title=(EP412), Mastery=80
```

Now the cutoff score has been set to 80%. Note that Lertap automatically outputs percentage correct scores when the Mastery control word is found on a *sub card; there's no need to use the PER control word.

[6] The U-L statistics do not involve the whole class—in this case, the class consisted of 99 students, but the U-L groups involve only 54 cases.

In the case of mastery analyses, the Stats1ul report changes from its normal (nonmastery) appearance. What was the U-L disc. is now the "B disc." index, after Brennan (1972):



Microsoft Excel - Ed502.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "EP412 semester test", created: 14/11/00.

Res =	A	B	C	D	other	U-L diff.	B disc.
Item 1 (Masters)	95%	0%	0%	5%		0.64	0.46
Item 1 (Others)	49%	3%	13%	36%			
Item 2 (Masters)	11%	0%	84%	5%		0.66	0.28
Item 2 (Others)	33%	0%	56%	10%			
Item 3 (Masters)	0%	0%	84%	16%		0.69	0.23
Item 3 (Others)	3%	21%	62%	15%			
Item 4 (Masters)	16%	42%	0%	42%		0.45	- 0.04
Item 4 (Others)	5%	28%	21%	46%			

Data CCs Freqs Scores Stats1f Stats1b **Stats1ul**

The B disc. figure is interpreted in a manner analogous to the U-L disc. figure, and (in fact), it's computed in the same way, that is, by subtracting the item's difficulty in the "Others" group from the difficulty found in the "Masters" group.

The big difference between a masters analysis and an ordinary U-L analysis is in the formation of the two groups. In the masters case, the "upper" group, the "Masters", is comprised of all those students whose percentage correct score was equal to or greater than the score corresponding to the cutoff percentage. In the mastery test analysis case, the "lower" group is called "Others", and in it will be found all students whose percentage correct score was below the cutoff.

This means that no-one is missed out. Lertap's mastery test analysis includes all students, not just the top and bottom 27% normally found in a conventional U-L analysis[7].

What would we make of the four items summarised above? Well, if we're wanting items which will help identify those who've mastered the material and those who haven't, we're likely to want to use all the stats available. The discrimination values for three of the four items are okay. Option B on Item 2 was not an effective distractor. More of the weaker students got Item 4 correct than did the "Masters"—there may be something wrong with option B on Item 4 as a high proportion of the Masters selected it.

[7] Note that this means that the U-L diff. value is now the item's real difficulty index as all cases are involved.

But wait, there's more to look at. After all the item statistics have been presented, the Stats1ul report has some small tables.

Summary group statistics				
	n	avg.	avg%	s.d.
Masters	19	34.6	77%	2.7
Others	39	25.8	57%	3.9
Everyone	58	28.7	64%	5.4

This was an Upper-Lower analysis based on a mastery cutoff percentage of 70.

The summary group statistics table shows how many people ended up in each of the two groups, and what their respective test score statistics came out to be. The average of the Masters group was 20 percentage points higher; the "avg." column shows a difference of about 9 points between the two groups, which we can interpret as meaning the Masters got, on average, 9 more questions correct (there were 45 questions on the test).

Variance components			
	df	SS	MS
Persons	57	37.91	0.67
Items	44	118.69	2.70
Error	2508	446.24	0.18
Index of dependability:		0.732	
Estimated error variance:		0.005	
For 68% conf. intrvl. use:		0.070	

The variance components table seen in the Stats1ul mastery report is based on the work of Brennan and Kane (see Brennan, 1984, and Brennan & Kane, 1977). Their model for the sources of variances underlying observed test scores differs from the model used in the classical true-score case. Brennan & Kane add another component, one which reflects the variance introduced by sampling items from the mastery test's domain. As a result their estimate of measurement error is higher than that found in Lertap's Stats1f report, and the Brennan-Kane index of dependability, which is analogous to coefficient alpha in the classical model, is usually lower. For example, the Stats1b report gives these values for the same data:

reliability (coefficient alpha):	0.73	
index of reliability:	0.86	
standard error of measurement:	2.81	(6.2%)

In this example it seems that Brennan and Kane's dependability index is the same as alpha, not lower, but this is due to the rounding caused by Lertap's display. Taking both values out to more significant figures[8], the Brennan and Kane index is 0.7323, while alpha is 0.7325.

More difference can be seen in the two error figures. The standard error of measurement is 6.2%, or, as a proportion, 0.062. This is the value to add and subtract from each student's percentage correct score, or proportion correct, to get the 68% confidence interval discussed above. A student with a test score of

[8] To see more significant digits, just click on the corresponding cell in the worksheet, and then look in Excel's formula bar.

29 has a percentage score of 29/45, or 64.4%. Adding and subtracting 6.2% gives a 68% confidence interval of 58.2% to 70.6%.

Using the Brennan and Kane estimate of error results in a larger confidence interval. Lertap's results show that we should add and subtract 0.070 from a student's proportion correct score, or 7% from the percentage correct score, which gives a 68% confidence interval of 57.4% to 71.4%.

Note that both approaches produce a confidence interval which spans the cutoff value of 70%. From what we know of the error involved in our testing, it's of a magnitude sufficient to rule out saying that the student whose test score is 29, 64.4%, is not a potential "master". In other words, classifying the student with a score of 29 as a nonmaster may be wrong.

Which of the two error figures should be used, the standard error of measurement, based on coefficient alpha, or the error variance from Brennan and Kane's approach? The latter. Brennan and Kane's. In the dinkum[9] mastery test situation, where items are sampled from a domain, their estimate of error is better, and fairer to students.

Finally, we come to the last two lines of Lertap's Stats1ul mastery report:

Prop. consistent placings:	0.783
Prop. beyond chance:	0.514

To understand how to use these two figures, let's review what we want our mastery test to do: separate the masters from the others, from the nonmasters. We know that the procedure we have used has error associated with it; we're going to make some mistakes, we're going to erroneously call some people masters when they're not, and vice versa.

How to estimate the degree of classification error associated with mastery testing is something well addressed in an article by Subkoviak (1984). Subkoviak reviewed a number of procedures, and ended up saying (p.284): "All things considered, the Huynh procedure seems worthy of recommendation....". However, Huynh's procedure is computationally complex, and Lertap uses another procedure, the "Peng and Subkoviak approximation" to Huynh's method (Subkoviak, 1984, pp.275-276).

Lertap's "Prop. consistent placings:" figure is Peng and Subkoviaks' approximation to $\hat{p}_0$, an estimate of the proportion of test takers who have been correctly classified as either master or nonmaster. In the example above, the estimate is that some 78% of the students have been correctly classified as either master or "other" (nonmaster). The corollary of this is, of course, that more than 20% of the students, just over one in five, may have been incorrectly classified.

To understand the second figure, the "Prop. beyond chance:", consider this: we could use a coin toss to decide the classification of each student. Is Brenda a master or nonmaster? Toss the coin—heads means master, tails means other. This isn't fair, of course, but nonetheless it's a process which will correctly classify some students, just by chance. With this in mind, the second figure above, the "Prop. beyond chance:", is an estimate of how accurate our classification has been over and above what we might get just by chance.

[9] Dinkum is a word used in Australia to mean genuine.

Lertap's "Prop. beyond chance:" figure is an estimate of "kappa", $\hat{\kappa}$, a statistic of interjudge agreement originally proposed by Cohen (1960). Subkoviak (1984, p.271) writes that "... in many instances, kappa differs little from the familiar Pearson correlation for dichotomous data, i.e., the phi coefficient....".

For a good review of the arguments and methods behind the index of dependability, $\hat{p}_0$, and $\hat{\kappa}$, see Berk (1984), and, again, Subkoviak (1984).

Validity

The reliability of a test is a measure of the test's accuracy, and is an index of how free the test was from error. Lertap provides various indicators of reliability, including coefficient alpha and its corresponding standard error of measurement, and, in the mastery test case, Brennan and Kanes' index of dependability, and its respective error estimate.

These indices allow us to interpret test scores with appropriate caution. They remind us that tests always have error associated with their use—if we use test scores to decide who gets an A, or who can be said to have mastered the material, we're likely to make mistakes, to make false classifications. The calculation and use of confidence intervals, as exemplified above, will help minimise classification errors.

But in all of this we have not asked the most important question which we have to put to our test: Is it *valid*? Is it indeed measuring what we wanted it to? It may be measuring something with good reliability, but is that "something" what we set out to measure?

Here's what Hopkins (1998, p.72) has to say about validity:

> The *validity* of a measure is how well it fulfills the function for which it is being used. Regardless of the other merits of a test, if it lacks validity, the information it provides is useless. The validity of a measure can be viewed as the "correctness" of the inferences made from performance on the measure. These inferences will pertain to (1) performance on a *universe* of test items (content validity), (2) performance on some criterion (criterion-related validity), or (3) the degree to which certain psychological traits or constructs are actually represented by test performance (construct validity).

Here's what Linn & Gronlund (1995, pp.47-48) say:

> Validity refers to the adequacy and appropriateness of the interpretations made from assessments, with regard to a particular use. For example, if an assessment is to be used to describe student achievement, we should like to be able to interpret the scores as a relevant and representative sample of the achievement domain to be measured. If the results are to be used to predict students' success in some future activity, we should like our interpretations to be based on as good an estimate of future success as possible. If the results are to be used as a measure of students' reading comprehension, we should like our interpretations to be based on evidence that the scores actually reflect reading comprehension and are not distorted by irrelevant factors. Basically, then, validity is always concerned with the specific use of assessment results and the soundness of our proposed interpretations of those results....

Lertap's standard variety of reports have essentially nothing to do with test validity. To come back to the example we started the chapter with, Dr Hartog's

EP 412 test on theories of learning, Lertap has no way of telling if the items used on the test had any relationship whatsoever to theories of learning. As far as Lertap knows, the test items could have involved automobile mechanics.

We would presume that Dr Hartog designed his test items to measure how well students had absorbed and mastered the material he gave them on theories of learning. Whether or not the test validly performs in this regard is not something we can expect a computer program to determine. We might want to turn to expert judges, to professionals knowledgeable in the field, and ask them to examine the test's content.

A tool which Lertap has which is sometimes of use in determining validity is its external criterion analysis, one of the options on the Run menu. If we had another test on theories of learning, and if that test had already had its validity confirmed by experts, we could ask Dr Hartog's students to answer this second test as well. We'd then use Lertap's external criterion analysis ability, and tell Lertap to use this second test score as the criterion measure for its analyses. If the test which Dr Hartog created is measuring knowledge of theories of learning, we'd expect the Hartog items to correlate well with scores on the second test.

More discussion on using an external criterion can be found in Chapter 2, and also in Chapter 10.

Can I fix my test so that it's better?

This is a common question. We hear it when instructors have used Lertap with their own test, and obtained a low reliability figure, a value for coefficient alpha which is, say, below 0.70.

One of our first responses is to ask the user if s/he realises that low values of alpha do not mean the test was worthless. Alpha values are of interest when we want to have a test which can pull the test takers apart, separate the wheat from the chaff, tell us who appears to know the material, and who does not.

If the test was meant as a formative one, or a diagnostic one, or was designed to work in a criterion-referenced context, then alpha is something which may not be of interest. Lertap's reports of item response frequencies and item difficulties may provide a wealth of information in these situations, as we have discussed above.

Having given this advice, if we then hear that the test was meant to be used to discriminate among test takers, producing scores which would form the basis for grades, or mastery/nonmastery classifications, we ask to see the data set involved.

We look first at the bottom of the Stats1f report, that is, the end of the full statistics worksheet. We do this to first confirm what the user has said about finding a low alpha value. After this we proceed to take in the bands of item difficulties.

The item difficulty bands are scanned to see if any items were very hard, that is, had their difficulty entry in the .00 or .10 band. We review any such items with the instructor—it's not uncommon to find that these items have been incorrectly keyed—a look at the *key card in the CCs sheet will sometimes show that an error has been made. If this happens, the error is fixed, and then the Run menu is used again to "Interpret CCs lines" and apply the "Elmillon item analysis".

After this step, if alpha remains low, we print the brief statistics report, Stats1b. We want to use its ? column in conjunction with the Stats1f and Stats1ul sheets to look at items whose distractors have problems. What we'll often find is a pattern which displays little difference down Stats1f's "avg." column—poor discrimination. Some of the distractors will have avg. values higher than that for the item's correct answer.

We expect to see this confirmed in the Stas1ul report, which is at times easier for the instructor to understand, depending on his/her statistics background. In the Stats1ul sheet, weakly discriminating items will have the upper group spread over all responses, when they should (ideally) be concentrated on the right answer.

How to fix these items is usually something only the instructor can determine. Something is wrong—the strong people are not picking the best answer—there's a need to find out why. Often a very useful procedure is to review the problematic questions with the class, asking them why the various responses to an item could be seen by some as the best answer. The answers which the students give will often reveal ways of interpreting the item's stem (the question), and the responses, which the instructor had never envisaged. Ambiguities will surface—words and phrases which the instructor thought clear will turn out to be questionable, and the need for item revision will become much more obvious.

Actions such as this will clarify what might be done to fix the faulty items, but they leave the instructor with a vexing question: what to do with the results on hand? I've given a test, and it's turned out to have some poor items. I'll work on repairing the bad questions for next time, but: What should I do with the test scores I have now?

Rescoring the test so that bad items are omitted is generally an unattractive proposition as it penalises those who gave correct answers to these items. A more palatable option (perhaps) is to multiply-key items, giving points for more than one answer. In Lertap this is done by using *mws cards, as described in Chapter 5.

In the end the practical consequence of a low alpha value lies in the standard error of measurement (or, for mastery tests, the "68% conf. intrvl." value). If the instructor wants to make use of the test scores, as best possible, decisions on how to interpret a student's test score should be based on a confidence interval. We've given examples of forming and using such intervals above.

Summary

Lertap does not shirk its job. It provides three distinct reports for looking at test results: the full statistics report, Stats1f; the brief stats report, Stats1b; and the upper-lower report, Stats1ul.

If anything could be said as a simple summary, it might be that these reports provide information above and beyond what's needed in many situations.

It's not necessary to use all of the details in these reports—users should pick and choose, deciding what's best for them, and for the needs they have at hand.

We have implied that there are, perhaps, three fundamental uses to which Lertap's various reports may be put: to reflect on the instructional process, to indicate how much students know, and to discriminate among students.

It is only the last of these objectives which begs for a detailed analysis of item discrimination and test reliability. We can look at how students did, and reflect on what their performance means for our instructional strategies, by looking over item response frequencies, and item difficulties. We can ask Lertap to express test scores as percentages, and then use percentage-correct figures as indicators of topic dominance, or as pointers to the need for topic revision.

We can do these things without looking at coefficient alpha, or at the index of dependability. However, when we want to use the test scores to indicate who's the strongest of the students, who's reached mastery level, then, and usually only then, do we start to dig among Lertap's reports, looking for evidence of good item discrimination, and high alpha (or dependability). Being the candid and fair professionals we are, we're up front when it comes to admitting that our testing process is not free from error; in this regard, Lertap provides the figures needed to form confidence intervals, score ranges which reflect the imprecision of our tests.