

Chapter 8

Interpreting Lertap Results for Affective Instruments

Note inserted 22 November 2022: this [Practical Exhibit](#) is suggested viewing.

Contents

A simple class survey	2
An example of a "scale"	5
What's a good alpha value?	8
Improving reliability	9
Processing a major survey	11
A survey with multiple groups	13
Breaking out subgroups	14
Using an external criterion	23
Making a move for another external criterion	25
More correlations (completing the picture)	27
Further analyses	28

Lertap's original mission in life was to process results from cognitive tests, from measures of student achievement, and/or from instruments designed to assess aptitude and talent. This mission was later expanded so as to encompass testing in what is referred to as the "affective domain". Since then, users have employed Lertap to process results from surveys as often as they have to process cognitive test results.

There are, of course, numerous other systems useful for processing surveys. Of them, the SPSS statistical package¹ would have to be one of the most popular. But we'll point out in this chapter that Lertap can often be a little gem of a survey processor, offering some advantages over SPSS.

Let's get some terminology matters out of the way first. What does Lertap mean when it refers to an affective "test", or "subtest"? What's the difference between an affective "test", and a "scale".

Hopkins (1998, p.273) writes: "Cognitive tests assess *maximum* or *optimum* performance (what a person *can* do); affective measures attempt to reflect *typical* performance (what a person usually *does* or *feels*)."

Linn & Gronlund (1995, p.32) state that the affective domain includes "Attitudes, interests, appreciation, and mode of adjustment".

It is common to refer to an affective "test" as a *scale*. Kerlinger (1973, p.492) assists in drawing a distinction between tests and scales: "... tests are scales, but

¹ <https://www.ibm.com/products/spss-statistics>

scales are not necessarily tests. This can be said because scales do not ordinarily have the meanings of competition and success or failure that tests do.”

In practical terms, the major difference between a cognitive test and an affective test, or scale, is that the questions on a cognitive test have a “correct” answer, a single response to which we attach points, whereas affective test items do not—it is usually the case that scoring an affective test item involves giving different points for different responses.

On a cognitive test item, the right answer usually gets one point, while the other responses usually get none. On an affective item, the first response option may equate to one point, the second to two points, the third to three points, and so on.

We’ll look at some examples. As we do, we’ll often use the term “test”, and “subtest”, even though we’re not dealing with cognitive measures in this chapter. If you’re familiar with the SPSS data analysis package, we will be working in the area which SPSS refers to as “reliability analysis”.

A simple class survey

Fifteen graduate students were asked to answer the following survey. They were not asked to provide their names².

1. *The amount of work I did for this unit was*

very great 1 2 3 4 5 quite small

2. *The quality of my work for this unit was*

excellent 1 2 3 4 5 poor

3. *I learned from this unit*

very much 1 2 3 4 5 very little

4. *The skills learned during the unit will be*

very useful 1 2 3 4 5 useless

5. *The teacher expressed his ideas clearly*

always 1 2 3 4 5 never

6. *The teacher avoided confusing or useless jargon*

always 1 2 3 4 5 never

7. *The teacher covered the material*

too quickly 1 2 3 4 5 too slowly

8. *The class sessions were*

stimulating 1 2 3 4 5 boring

9. *The textbook was (with respect to my work)*

relevant 1 2 3 4 5 irrelevant

² The survey was used at Curtin University. What are called “courses” in North America, and “papers” in New Zealand, are referred to as “units” at Curtin.

10. *The textbook was (in general)*

interesting 1 2 3 4 5 boring

11. *The work required for this unit was*

excessive 1 2 3 4 5 too little

12. *The unit should run again with no major changes*

strongly agree 1 2 3 4 5 strongly disagree

Students indicated their responses by circling their number of choice for each question. When the answer sheets were returned to the instructor, he wrote a sequential number on the top of each sheet in order to have an ID "No." to carry in the data processing.

The responses were then typed into an Excel worksheet, as shown below³:

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Ed 503 class survey, 8 September.												
2	No.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
3	1	3	3	3	4	3	3	3	3	2	3	3	4
4	2	3	2	3	3	2	4	4	4	4	3	3	5
5	3	3	3	2	2	2	2	4	4	3	3	3	5
6	4	1	2	3	4	4	2	1	2	2	2	2	5
7	5	2	2	2	2	3	3	1	3	2	2	2	2
8	6	2	3	2	3	3	3	2	3	4	5	2	3
9	7	2	3	2	3	3	3	1	2	4	3	3	5
10	8	2	4	3	3	3	2	3	2	2	1	2	3
11	9	1	3	3	3	3	2	2	3	2	2	1	5
12	10	2	4	1	1	1	1	3	2	2	2	3	1
13	11	1	3	2	2	2	2	3	2	3	4	3	4
14	12	3	2	2	2	3	2	3	3	2	3	3	4
15	13	3	3	5	1	1	1	1	3	2	3	3	4
16	14	2	2	1	1	3	3	3	2	2	3	3	3
17	15	3	3	3	2	3	2	4	3	3	3	2	4

The initial CCs worksheet had just these two entries:

*Col1 (C2-C13)
*Sub Affective

Lertap's [Run options](#)⁴ were then accessed to "[Interpret CCs lines](#)" and to initiate an "[Elmilon item analysis](#)". This resulted in the creation of the two standard statistical reports for affective tests (or "subtests"), Stats1f, and Stats1b.

³ Excel 2003 was in use when this snapshot was taken.

The [Stats1b](#) report gives a brief summary of item responses, and tosses in a few item statistics as well⁵:

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	20%	40%	40%				+	2.20	0.75	0.34
Q2		33%	53%	13%			+	2.80	0.65	- 0.32
Q3	13%	40%	40%		7%		+	2.47	0.96	0.08
Q4	20%	33%	33%	13%			+	2.40	0.95	0.32
Q5	13%	20%	60%	7%			+	2.60	0.80	- 0.04

What to make of these results? A lot may be gleaned from the percentage figures—for example, the instructor can quickly see that more than half, 13% + 40%, indicated that they learned something from the unit (Q3), with the same total percentage, 53%, saying that they felt the skills learned would be useful (Q4).

The columns to the right of “other” indicate the type of scoring applied to each item (the pol. column⁶), the mean of the responses, their standard deviation, and the degree to which responses on each item correlate with the sum of the responses on the other items⁷. These columns are not always used in interpreting results. In this case, the 12 items were not meant to measure a single construct, or theme. Some of them have to do with the student’s own assessment of her or his work, others with how much they learned, others with the instructor’s delivery, and still others with the text.

In this example, it would not be meaningful to form a subtest “score” by summing the responses over all 12 items. But Lertap does it anyway, producing a Scores worksheet with exactly such scores. Since these have no sound rationale to them, we won’t even look at them. We might just as well delete the Scores worksheet.

⁴ When Excel 2016 is used, the “Run options” are in the [“Run menu”](#).

⁵ The screen snapshot is from Excel 2016.

⁶ “Pol.” is + when the scoring is forward, with one point for the first response, two for the second, and so forth. Pol. is minus (-) when reverse scoring is in effect, in which case the points begin incrementing from the right instead of from the left.

⁷ The correlation is a Pearson product-moment coefficient, corrected for part-whole inflation (see [Chapter 10](#)).

Lertap also produces a “full statistics” report, a [Stats1f](#) worksheet. It contains more detailed information for each item, and then, at the end, details on the overall “subtest”.

The item details in Stats1f look like this:

Lertap5 full item stats for "Test1", created: 23/8/19.

Q1 (c2)

option	wt.	n	%	pb(r)	avg.	z
1	1.00	3	20.0	-0.17	30.3	-0.34
2	2.00	6	40.0	-0.47	29.3	-0.57
3	3.00	6	40.0	0.61	35.0	0.75
4	4.00	0	0.0	0.00	0.0	0.00
5	5.00	0	0.0	0.00	0.0	0.00

Q2 (c3)

option	wt.	n	%	pb(r)	avg.	z
1	1.00	0	0.0	0.00	0.0	0.00
2	2.00	5	22.2	0.10	21.2	0.14

In this simple little class survey, with items measuring different aspects of the unit, only one of the Stats1f columns is likely to be of any use: “n”. It shows the actual number of students selecting each response, something not found in the brief report presented earlier. All the other columns, except the first, have to do with weights, correlations, and scores, and we’re not interested in such matters in this example⁸.

For the same reason, we’re not excited by the summary statistics which appear towards the end of the Stats1f report. For example, the subtest’s reliability figure came out to be 0.56, but, since we’re not interested in the scores which Lertap made by adding together answers to very different questions, we have no use for this figure. It has no meaning—we’re not saying we have a “scale”—we have several unrelated questions and want only to look at responses on an item by item basis.

In short, we’ve started this chapter’s action by looking at a small, but rather typical class survey, something quite a number of instructors will use to get feedback from students at the end of a period of instruction. It was very easy to prepare the data for processing, and Lertap’s two little CCs entries were a cinch. We had results in quick order.

An example of a “scale”

Nelson (1974) devised a 10-item survey instrument in an attempt to assess how “comfortable” people felt with the use of Lertap2, a system that appeared in 1973:

⁸ See [Chapter 10](#) for more discussion of the Stats1f reports for affective subtests.

Please indicate your answer to the following questions by checking one of the blanks to the right of each item.

SA = strongly agree.
A = agree
N = neutral or neither agree nor disagree.
D = disagree.
SD = strongly disagree

		SD	D	N	A	SA
(26)	I did well on the quiz above.	—	—	—	—	—
(27)	LERTAP seems very complex.	—	—	—	—	—
(28)	I have used item analysis programs superior to LERTAP.	—	—	—	—	—
(29)	The user's guide to use and interpretation is inadequate.	—	—	—	—	—
(30)	I need clarification on several terms used in the user's guide.	—	—	—	—	—
(31)	I will recommend to others that they use LERTAP.	—	—	—	—	—
(32)	The examples given in the user's guide are good, and instructive.	—	—	—	—	—
(33)	I don't think I could design my own LERTAP analysis.	—	—	—	—	—
(34)	I see areas in which LERTAP could stand improvement.	—	—	—	—	—
(35)	LERTAP control cards seem flexible and easy to use.	—	—	—	—	—

These ten questions are part of the "Lertap Quiz". The entire quiz is given in [Appendix A](#). The actual data resulting from its administration to 60 workshop participants may be found in the Data worksheet, one of the four visible sheets included as part of the Lertap5.xlsm file. If you have Lertap running on your computer, you have a copy of the data.

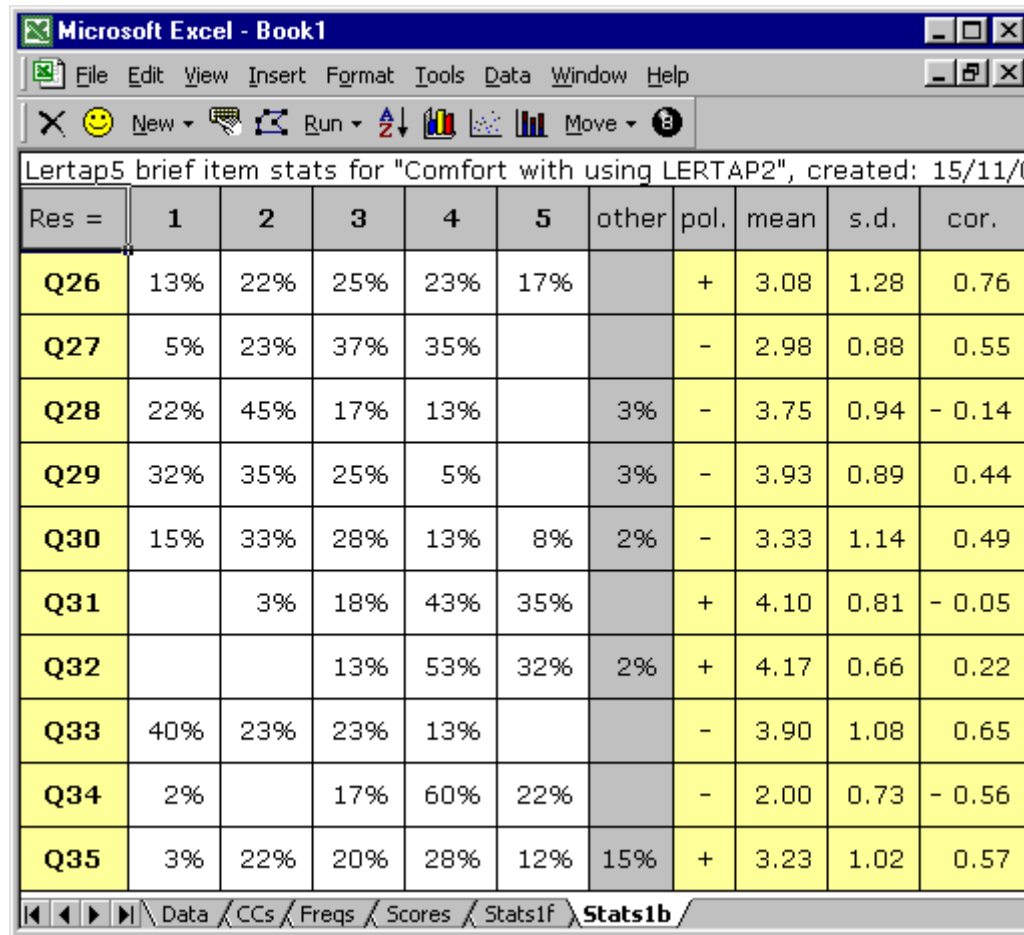
Answers to these 10 items were processed by using five digits: 1 for SD, through to 5 for SA. To get survey results, Nelson used these CCs lines:

```
*col (c28-c37)
*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort)
*pol +----- +-----
```

Nelson used a *pol card to reverse-score questions 27, 28, 29, 30, 33, and 34. People who answered "SD" on these questions got a score of 5 points. He did this

because these questions were negatively worded; their most-favourable response is "SD". He wanted the most-favourable response to each item to get the maximum possible score, or weight, which was 5 points⁹.

Here's the brief stats report for this survey¹⁰:



Lertap5 brief item stats for "Comfort with using LERTAP2", created: 15/11/0

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q26	13%	22%	25%	23%	17%		+	3.08	1.28	0.76
Q27	5%	23%	37%	35%			-	2.98	0.88	0.55
Q28	22%	45%	17%	13%		3%	-	3.75	0.94	- 0.14
Q29	32%	35%	25%	5%		3%	-	3.93	0.89	0.44
Q30	15%	33%	28%	13%	8%	2%	-	3.33	1.14	0.49
Q31		3%	18%	43%	35%		+	4.10	0.81	- 0.05
Q32			13%	53%	32%	2%	+	4.17	0.66	0.22
Q33	40%	23%	23%	13%			-	3.90	1.08	0.65
Q34	2%		17%	60%	22%		-	2.00	0.73	- 0.56
Q35	3%	22%	20%	28%	12%	15%	+	3.23	1.02	0.57

Navigation: Data / CCs / Freqs / Scores / Stats1f / **Stats1b**

In this example, Nelson was interested in all of the report's columns. He wanted to think that his collection of 10 items could be seen as a "scale", a coherent set of questions measuring the same thing, that thing being what he referred to as the apparent "comfort" users reported with his software.

Keep in mind that the maximum possible score on any of the 10 questions was 5. Nelson hoped to see item means close to 5, or at least above 4. In this he would have been disappointed; some of the item means were rather low.

Do the results give him reason to believe he had indeed created a "scale"? No, not exactly. Three of the items had negative correlations, something not expected in a scale of good quality.

He next turned to Lertap's "full" statistics report, Stats1f. There he found that the subtest's, or would-be scale's, reliability (coefficient alpha) came out to be 0.63, a rather low figure, certainly lower than what he wanted.

⁹ The ease with which Lertap allows items to be reverse-scored is a feature not currently found in some other systems, SPSS among them.

¹⁰ As seen in Excel 2003, an older version of Excel.

Of interest were the reliability “bands” reported towards the end of the Stats1f sheet:

<u>without</u>	<u>alpha</u>	<u>change</u>
Q26	0.453	-0.175
Q27	0.550	-0.079
Q28	0.690	0.062
Q29	0.574	-0.055
Q30	0.552	-0.076
Q31	0.664	0.036
Q32	0.618	-0.010
Q33	0.509	-0.120
Q34	0.730	0.102
Q35	0.536	-0.093

These figures confirmed what the correlation results were saying: the subtest’s reliability, as measured by coefficient alpha, would increase if items Q28, Q31, and/or Q34 were to be omitted from the “scale”. Just leaving out one item alone, Q34, would boost the alpha value to 0.73.

These results left Nelson in a contemplative mood. The negative correlations, and low alpha value, led him to give away the idea of considering the 10 items to be a scale—the results indicated that adding up the responses over all items to get a score was not highly defensible—the score had what he felt to be inadequate reliability. Of course, the matter of having or not having a scale was not the only issue which he wanted to investigate. Above all, he wanted to see how people in a four-day workshop reacted to their first experience with his software. In this regard the item results gave him much to mull over—there were some positive outcomes, to be sure, but, as is almost always the case, the respondents seemed to be pointing to areas needing more attention.

What’s a good alpha value?

We’ve looked at just two examples to this point. In the first, there was no thought of using the scores which Lertap produced by summing item responses—scores were not an issue, they were not wanted, the instructor limited his analysis to the individual item level.

In the second example, the user (Nelson) did have an interest in the scores; he hoped Lertap would support the formation of his “Comfort scale”. In this he was disappointed, getting an alpha value of 0.63, and observing several negative item correlations.

What should alpha be? Is there a sort of minimum acceptable figure?

To a considerable extent, the answer to these questions depends on the uses which will be made of the scores resulting from the test (or survey). It is uncommon to find survey scores used to make decisions about individuals—in the cognitive test examples discussed in the last chapter, a person’s score on a test was sometimes used to decide on a letter grade for the person, such as “A” or “B”, or on a mastery / non-mastery classification placement. Decisions such as these have important consequences for individuals; high reliability figures are *required* in the cognitive realm.

This is not usually the situation in surveys. Rather, survey scores are often used as correlates with other variables. For example, in the second example above, Nelson wanted to see if there was a relationship between participants' affective reactions to use of the Lertap 2 system, and their scores on a cognitive test, "Knwldge", a test designed to index how well they understood some of the inner workings of the same system. (He found a correlation of 0.80, and a scatterplot suggesting a definite relationship; see [Chapter 2](#) for more details.)

When survey scores are used in this manner, their reliability figures don't have to be so high. We might even allow them to dip as low (say) as 0.70 or so. However, it would certainly be the case that we want to avoid having tests, or scales, whose items have negative intercorrelations. When this happens, we have questions which are not hanging together; respondents are demonstrating inconsistency in their responses. In the first example above such inconsistency was anticipated by the instructor. For example, he didn't expect answers to questions dealing with his unit's textbook to have any relationship to a question regarding his use of jargon during class.

In the second example, Nelson did hope for a consistent response pattern over all ten items but didn't get it. The answers respondents gave to some of his items, Q34 in particular, tended to be opposite those given to most of the other items—if someone had a positive response to Q34, in other words, Lertap's results indicated that, by and large, they had a negative response on most of the other items. This is inconsistency, it lowers the alpha figure.

What might the literature say? One of the best references in this area that we know of is Pedhazur & Schmelkin (1991), who devote many pages to these matters. Unfortunately, they dodge making a final recommendation on how high reliability should be, saying "... *it is for the user to determine what amount of error he or she is willing to tolerate, given the specific circumstances of the study* (e.g., what the scores are to be used for, cost of the study)....". Kaplan & Saccuzzo (1993, p.126) state: "It has been suggested that reliability estimates in the range of .70 and .80 are good enough for most purposes in basic research." Mehrens & Lehmann (1991, p.428) write: "Attitude scales, by and large, have reliabilities around 0.75. This is much less than those obtained for cognitive measures, and hence the results obtained from attitude scales should be used primarily for group guidance and discussion."

Improving reliability

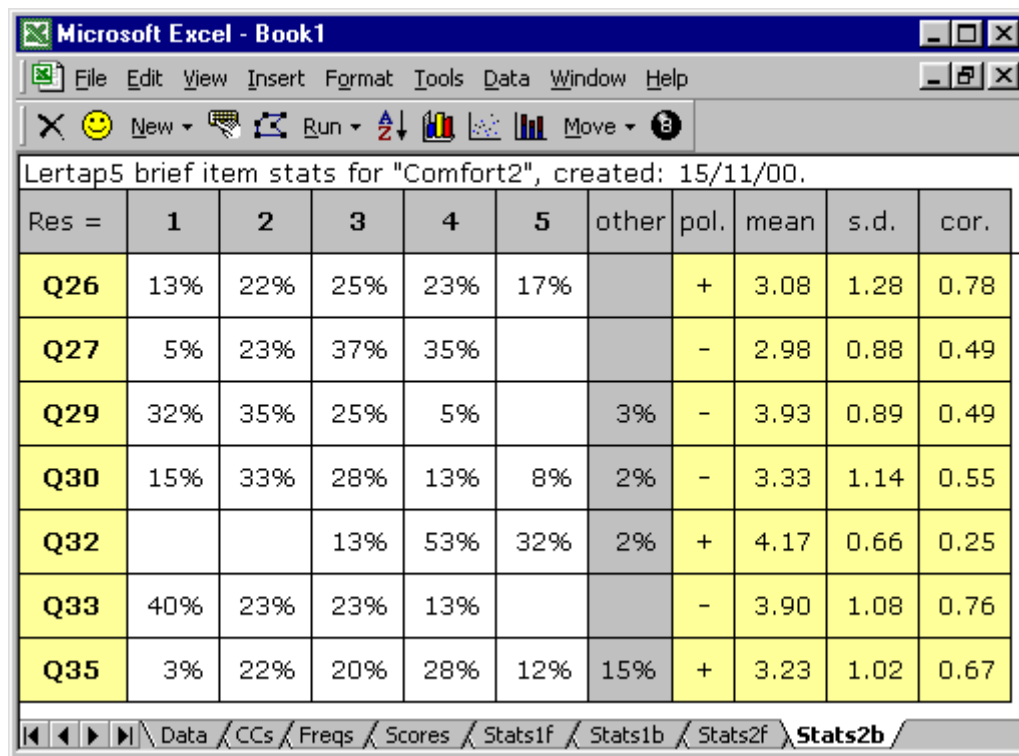
It is generally possible to see an increase in a subtest's alpha estimate of reliability when items with negative correlations are removed from the subtest. As an example, we added six control "cards" to Nelson's original three, ending up with these:

```
*col (c28-c37)
*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort)
*pol +---- +----+
*col (c28-c37)
*sub Aff, Name=(Comfort2), Title=(Comfort2)
*pol +---- +----+
*mws c30, *
*mws c33, *
*mws c36, *
```

There are two *col cards here, defining two groups of items for Lertap to process as subtests. In fact, the *col cards are identical—each subtest is said, initially, to have the same items.

The two *sub cards have obvious minor differences, and the *pol cards are identical. It's the last three cards, the *mws cards, which effectively remove from the second subtest, "Comfort2", the three negatively-correlating items, Q28 (found in column 30), Q31 (column 33), and Q34 (column 36) ¹¹.

Lertap's brief stats report for Comfort2 looked like this:



Lertap5 brief item stats for "Comfort2", created: 15/11/00.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q26	13%	22%	25%	23%	17%		+	3.08	1.28	0.78
Q27	5%	23%	37%	35%			-	2.98	0.88	0.49
Q29	32%	35%	25%	5%		3%	-	3.93	0.89	0.49
Q30	15%	33%	28%	13%	8%	2%	-	3.33	1.14	0.55
Q32			13%	53%	32%	2%	+	4.17	0.66	0.25
Q33	40%	23%	23%	13%			-	3.90	1.08	0.76
Q35	3%	22%	20%	28%	12%	15%	+	3.23	1.02	0.67

Navigation: Data / CCs / Freqs / Scores / Stats1f / Stats1b / Stats2f / **Stats2b**

This is about as good as it gets. The seven items in the new subtest have very good correlations, all comfortably positive. The Stats2f report indicated that coefficient alpha was 0.83, a nice increase over the value of 0.63 found when all ten items were included.

The bottom of the Scores worksheet gives the correlation between the subtests, Comfort and Comfort2, and it was very high: 0.96.

We mentioned that Nelson wanted to look at the correlation between his Comfort subtest, with ten items, and "Knwldge", a 25-item cognitive test. Lertap found it to be 0.80. How would the new subtest, Comfort2, correlate with Knwldge? We put these control cards to Lertap:

¹¹ The [exc "card"](#) could have been used instead: *exc (c30, c33, c36)

```

*col (c3-c27)
*sub Res=(A,B,C,D,E,F), Name=(Knowledge of LERTAP2), Title=(Knwldge)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
*col (c28-c37)
*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort)
*pol +---- +----+
*col (c28-c37)
*sub Aff, Name=(Comfort2), Title=(Comfort2)
*pol +---- +----+
*mws c30, *
*mws c33, *
*mws c36, *

```

The Scores report gave the intercorrelations among the three subtests:

Lertap5 scores worksheet, last updated on: 1/1/2011			
ID	Knwldge	Comfort	Comfort2
n	60	60	60
Min	1.00	26.00	17.00
Median	12.50	33.00	24.00
Mean	12.63	34.48	24.63
Max	24.00	43.00	33.00
s.d.	6.95	4.61	4.95
var.	48.27	21.25	24.53
MinPos	0.00	10.00	7.00
MaxPos	25.00	50.00	35.00
Correlations			
Knwldge	1.00	0.80	0.87
Comfort	0.80	1.00	0.96
Comfort2	0.87	0.96	1.00
average	0.84	0.88	0.92

The correlation between the cognitive test results and the new affective "scale", Comfort2, is 0.87, a worthwhile increase over the 0.80 figure obtained with the original Comfort scale. This outcome reflects a well-known fact: increasing the reliability of a test generally improves its chances of correlating with other measures.

Processing a major survey

Back in [Chapter 3](#) we introduced the University of Michigan's Motivated Strategies of Learning Questionnaire, MSLQ ([Pintrich, et al, 1991](#)). The MSLQ has been popular with a few researchers in our neighbourhood, and most recently Lertap 5 has been used to process results.

We have used a subset of the MSLQ's various scales to collect data from students on their study habits. Our modified version of the MSLQ has a total of 55 items, covering ten scales. The scales have names such as "Test Anxiety", "Critical Thinking", and "Self Regulation". Each scale's items are distributed throughout the survey form; a scale's items are not contiguous. For example, Test Anxiety is defined by answers to questions Q3, Q5, Q9, Q14, and Q20, while Critical Thinking involves Q10, Q21, Q25, Q40, and Q45.

Each question used seven possible responses, and had a format identical to Q14's:

		Not at all true of me						Very true of me
Q14	I have an uneasy, upset feeling when I take an exam.	1	2	3	4	5	6	7

The Lertap control cards for four of the ten subtests, the MSLQ scales, are shown below. Note that ampersands (&s) have been used to separate the subtests—this is not necessary, but it makes it a bit easier to see where each subtest's specifications begin.

MSLQ control card set 1, 4 July 2000.

```
&
*col (c14,c19,c25,c29,c40,c41,c42,c43,c47,c62,c64,c65)
*sub aff, scale, name=(Self-regulation), title=(SelfReg), res=(1,2,3,4,5,6,7)
*pol -++++ ++-++ ++
&
*col (c15,c17,c21,c26,c32)
*sub aff, scale, name=(Test anxiety), title=(TestAnx), res=(1,2,3,4,5,6,7)
&
*col (c16,c30,c36)
*sub aff, scale, name=(Peer learning), title=(PeerLrng), res=(1,2,3,4,5,6,7)
&
*col (c22,c33,c37,c52,c57)
*sub aff, scale, name=(Critical thinking), title=(CritThnk), res=(1,2,3,4,5,6,7)
```

The "scale" control word has been used in each *sub card in order to have Lertap report scores on the same numeric scale, 1 to 7. To understand why this is useful, consider the first MSLQ scale above, SelfReg. It has twelve items and would have a possible score range of 12 to 84; the PeerLrng scale, on the other hand, involves only three items, giving a possible score range of 3 to 21¹². Using "scale" has Lertap divide MSLQ scores by the number of respective subtest items, effectively knocking each MSLQ subtest score to the same 1-to-7 numeric range.

This may be seen in the screen capture below; notice how each MSLQ scale has two scores, the "raw" score, such as SelfReg, and the same score divided by the number of subtest items, SelfReg/. It would be difficult to compare raw-score means because they're based on a differing number of items—for example, the three raw score means are 54.06, 20.30, and 12.08, which might lead some to think that more positive responses were found in the SelfReg scale. But there were many more items in the SelfReg subtest—we need some way to standardise the subtest scores to the same range. Using the "scale" control word does the job: immediate use can be made of the scaled scores--we see, for example, that two of the three scales, TestAnx and PeerLrng, had scaled means very close to "4", the centre of the item scale.

¹² Possible scores on any single item vary from 1 to 7. Multiply each of these figures by the number of subtest items to get the possible score ranges shown.

Lertap5 scores worksheet, last updated on: 15/11/00.

ID_code	SelfReg	SelfReg/	TestAnx	TestAnx/	PeerLrng	PeerLrng/
n	139	139	139	139	139	139
Min	26.00	2.17	5.00	1.00	4.00	1.33
Median	55.00	4.58	21.00	4.20	12.00	4.00
Mean	54.06	4.50	20.30	4.06	12.08	4.03
Max	80.00	6.67	34.00	6.80	20.00	6.67
s.d.	8.92	0.74	6.68	1.34	3.48	1.16
var.	79.58	0.55	44.69	1.79	12.10	1.34
MinPos	12.00	1.00	5.00	1.00	3.00	1.00
MaxPos	84.00	7.00	35.00	7.00	21.00	7.00

Correlations

Navigation tabs: Data, CCs, OrigCCs, Cdbk, Freqs, **Scores**, Stats1f, Stats1b, Stats2f, Stats2b

As a matter of interest, we found alpha figures for five of our ten MSLQ scales to lie in the 0.62 to 0.66 range; four were in the 0.70s; and one was 0.81. Alpha values for these scales found at the University of Michigan are similar, except for one scale, "Help Seeking", which had an alpha figure of 0.52 at Michigan, and 0.64 in our study. de la Harpe (1998) reported finding the same scales to have alphas in essentially the same ranges, from lows around 0.60 to maxima around 0.80. Alpha values of 0.70 and above are often accepted as reasonable for scales – in this case, then, some of the MSLQ scales were not quite adequate¹³.

Look at the last set of control cards again for a moment. Why is there only one *pol card? Of the four subtests, why does only the first have a *pol "card"? Because in this subset of the MSLQ scales, only one of the scales had items which needed to be reverse-scored. If a subtest's items are all scored in the same manner, a *pol card is not required.

In this chapter we have alluded to Lertap's prowess in processing survey results. In the case of the MSLQ, however, we wanted analyses which Lertap could not provide¹⁴. For example, we wanted to have some means comparisons among groups of students, comparing MSLQ scaled averages by student major. For this we turned to SPSS. We first used Lertap to prepare the scaled MSLQ subtest scores, a task it's good at, especially when subtests involve items which must be reverse-scored. Then we used the Move option on Lertap's toolbar to copy selected columns from the Scores worksheet to the Data sheet. After this, we used Lertap's 8-ball icon to prepare a special worksheet which was easily imported by SPSS. There is more on this process, exporting Lertap worksheets, in [Chapter 10](#).

A survey with multiple groups

In October, 2000, the Faculty of Education at Curtin University surveyed a sample of their first-year students, searching for an indication of the extent of their satisfaction with a new outcomes-focused program.

¹³ Scores from scales with low alpha values should be used only with caution, or not at all.

¹⁴ Version 10 will compute results by groups. See "Breaking out groups, 2" on p.18

Twenty questions were answered by students in three groups: early childhood education (ECE), primary education (Pri), and secondary education (Sec). The questions had to do with how effectively the new program had helped them “learn about how to become a competent teacher”; “understand the role of a competent teacher”; and “practise outcomes-focused education”. The questions also asked how extensively academic staff had supported the students; and how well staff had incorporated both outcomes-focused and student-centred learning in their teaching. A third section had to do with overall satisfaction with the program itself.

All 20 questions used a Likert-style format, with five possible responses, from strongly agree to strongly disagree. Forward scoring was used with all questions, with the strongly-agree response getting a weight of 1.00, and strongly disagree a weight of 5.00. *Low scores were best in this case*, indicating the greatest satisfaction with the program.

Responses were anonymous. They were entered into a Data worksheet with the first column used for a sequential ID number, a number pencilled on each individual answer sheet after all sheets had been collected. The second column contained an E for ECE majors, P for Pri majors, and S for Sec majors. Actual question responses were entered in columns 3 through 22. The CCs “cards” were as follows:

*col (c3-c22)
*sub aff

Brief statistics for the first five items, using all 104 student returns, are shown below:

Lertap5 brief item stats for "Test1", created: 22/11/00.										
Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	23%	70%	6%	2%			+	1.87	0.59	0.72
Q2	29%	59%	11%	1%			+	1.83	0.64	0.60
Q3	26%	55%	17%	2%			+	1.94	0.71	0.59
Q4	5%	53%	34%	6%	2%		+	2.47	0.76	0.46
Q5	3%	53%	35%	7%	2%		+	2.52	0.75	0.45

Breaking out groups, 1

The Faculty wanted to see if responses to their survey differed by group, that is, by ECE, Pri, and Sec. Lertap allows for selected responses to be culled and

copied to a new workbook, something which is accomplished by using a *tst "card" as the first line in the CCs worksheet¹⁵.

The following "cards" were used to break out the ECE group:

```
*tst c2=(E)
*col (c3-c22)
*sub aff
```

With the above cards in the CCs worksheet, "Interpret CCs lines" was clicked on from the Run menu. Lertap created a new workbook, making partial copies of both the CCs and Data worksheets. The new CCs sheet excluded the *tst card seen above, while the new Data worksheet had only those records with an E in column 2. The new workbook was saved with a name of Quick20ECE.xls.

After this, the original workbook was returned to, and its CCs cards were modified so that another new workbook would be created for the Pri group:

```
*tst c2=(P)
*col (c3-c22)
*sub aff
```

This new workbook was saved as Quick20Pri.xls.

Finally, the CCs cards in the original CCs sheet were changed once more so as to pull out the Sec group:

```
*tst c2=(S)
*col (c3-c22)
*sub aff
```

The new workbook which resulted was saved as Quick20Sec.xls.

How were results obtained for each group? Each of the new workbooks was selected, and the Run menu used to "Interpret CCs lines" and apply the "Elmillion item analysis".

Did the three groups differ in their responses to the first five items of the survey? Have a look for yourself—compare the brief stats summaries (the groups are identified by the name of the workbook, shown at the top of each of these Excel screen captures):

¹⁵ Using *tst cards as described here is not the only way to breakout group results – see the next section for an alternative which is regarded by some as easier and faster.

Microsoft Excel - Quick20ECE.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 brief item stats for "Test1", created: 22/11/00.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	24%	62%	7%	7%			+	1.97	0.76	0.88
Q2	34%	52%	10%	3%			+	1.83	0.75	0.86
Q3	21%	66%	7%	7%			+	2.00	0.74	0.78
Q4	7%	69%	21%	3%			+	2.21	0.61	0.50
Q5		69%	24%	7%			+	2.38	0.61	0.40

Scores Stats1f Stats1b

Microsoft Excel - Quick20Pri.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 brief item stats for "Test1", created: 22/11/00.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	20%	76%	4%				+	1.84	0.47	0.62
Q2	18%	67%	14%				+	1.96	0.57	0.46
Q3	29%	47%	24%				+	1.96	0.73	0.52
Q4	6%	45%	45%	4%			+	2.47	0.67	0.43
Q5	6%	45%	45%	4%			+	2.47	0.67	0.49

Scores Stats1f Stats1b

Microsoft Excel - Quick20Sec.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

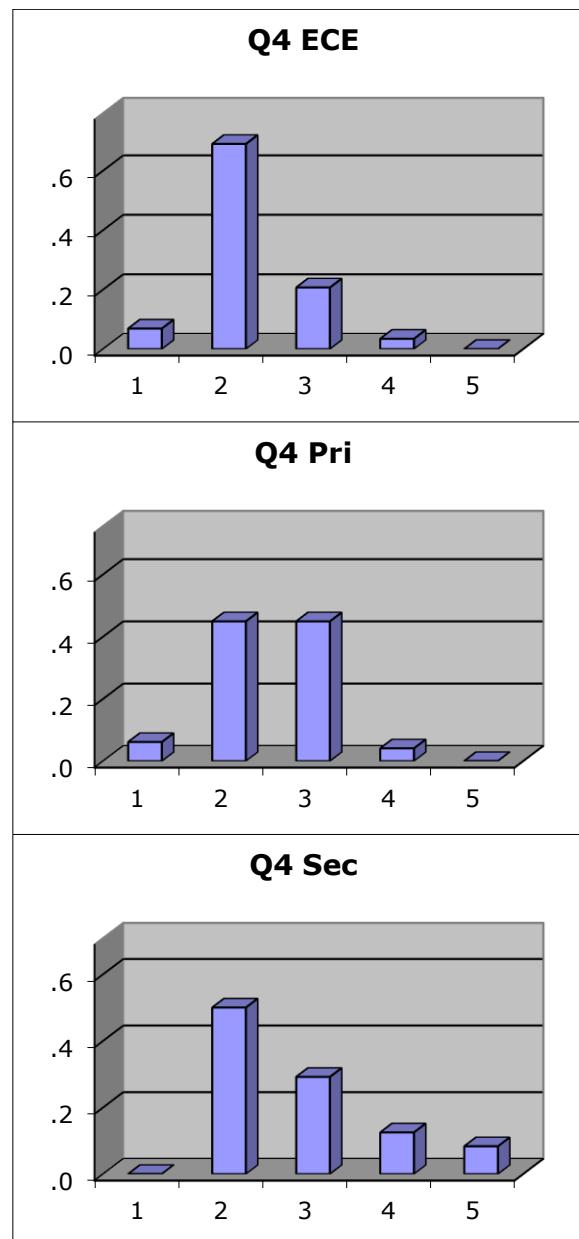
Lertap5 brief item stats for "Test1", created: 22/11/00.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	25%	67%	8%				+	1.83	0.55	0.64
Q2	46%	50%	4%				+	1.58	0.57	0.49
Q3	29%	58%	13%				+	1.83	0.62	0.48
Q4		50%	29%	13%	8%		+	2.79	0.96	0.58
Q5		50%	29%	13%	8%		+	2.79	0.96	0.52

Scores Stats1f Stats1b

The response patterns on the first three items appear to be similar for these groups, but some differences can be noted in Q4 and Q5. We asked Lertap to "Make item response charts from a Stats-b sheet" (there's an icon for this on the

toolbar, to the left of the Move option¹⁶). We then copied the charts for Q4, and present them below:



The response charts quickly capture the action, indicating a negative shift on Q4 responses as we go from ECE to Pri to Sec. (Remember that, in this case, 5 was the most negative response.)

Another way to answer the question about possible group differences might be to look at the overall mean in the three groups, a statistic found in the Stats1f sheet. Before doing this, however, we'd want to look at the subtest's alpha value to see if subtest scores are consistent, and (thus) interpretable as a scale score.

The Stats1f sheet has alpha values. We found them to be 0.92 for the whole group; 0.95 for ECE; 0.90 for Pri; and 0.92 for Sec. These values are high, giving

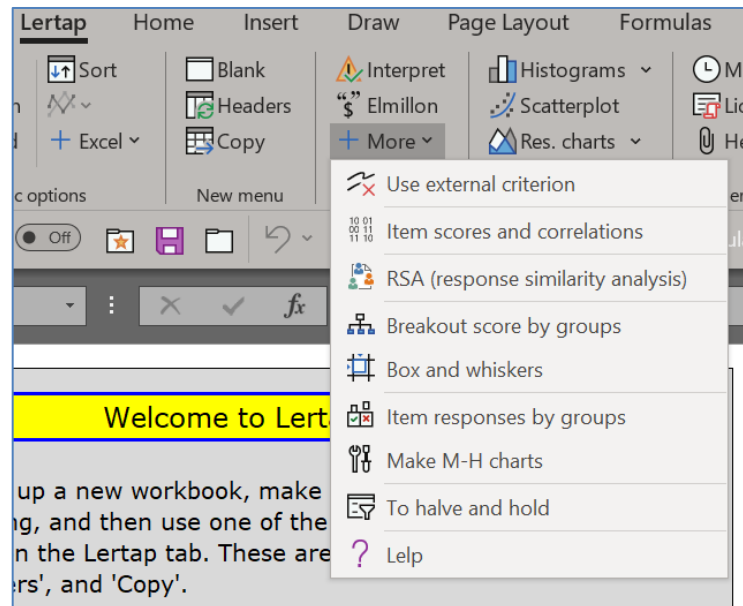
¹⁶ Please see [this topic](#); things have changed since this chapter was originally written.

us a green light for comparing group means: 40.38 for ECE; 40.51 for Pri; and 40.04 for Sec. At the subtest level, differences among the groups are negligible, despite there being some shifts in the response patterns at the item level.

Breaking out groups, 2

The use of *tst control cards described above is just one way to breakout results by groups. The method described in this section is a bit more convenient, and perhaps easier to use.

The “More” option, below the Interpret and Elmillion options, presents what’s called a “drop-down menu” with more options:



The “Breakout score by groups” option, and the “Item responses by groups” option may be used to get score and item response results by groups.

Q20_Leap_Project_Survey_2001.xlsx - Excel Larry Nelson

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Data	from Curtin University's LEAP project												
2	ID	Course	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
3	1e	E	3	3	2	2	2	4	4	2	2	4	4	4
4	2e	E	2	2	2	2	2	2	2	2	2	2	2	2
5	3e	E	2	2	2	2	2	2	2	2	2	2	2	2
6	4e	E	2	2	2	2	2	2	2	2	2	2	2	2
7	5e	E	2	2	2	2	2	2	2	2	3	2	2	2
8	6e	E	1	1	2	2	2	1	1	2	2	2	2	2
9	7e	E	2	2	2	2	2	2	2	2	2	2	2	2
10	8e	E	2	2	3	4	4	2	3	3	2	2	2	2
11	9e	E	1	1	1	1	2	2	2	2	2	1	2	2
12	10e	E	4	3	4	3	4	3	3	3	3			
13	11e	E	1	1	1	2	2	1	1	2	2	1	1	2

AutoSave Off Gridlines Formula Bar

Data CCs

100%

The Data sheet snapshot above displays a few of the results. The grouping variable in this case is "Course" with codes of E, P, and S recorded in column 2.

Item responses begin in the third column and end in column 22.

Using the "Item responses by groups" option, as [described here](#), produces the results shown below for the first four of the survey's items:

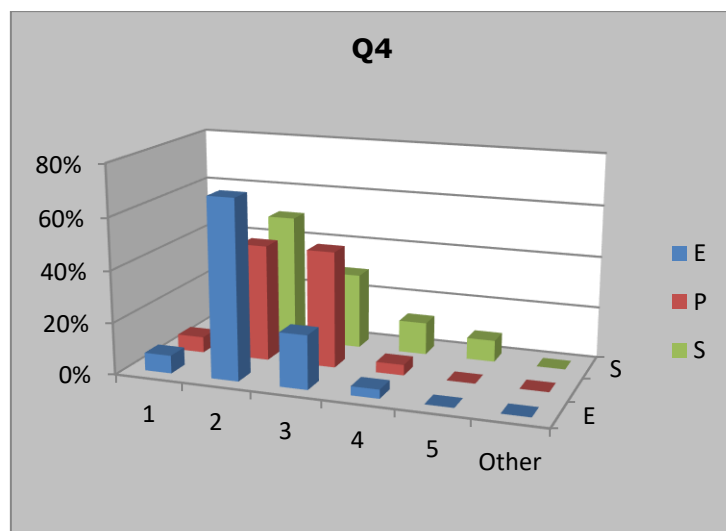
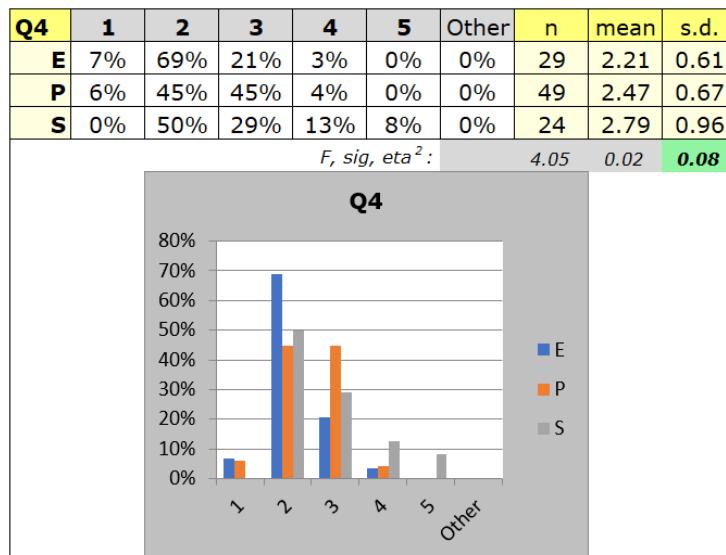
Lertap5 item responses for score Test1, grouped by Course (3 groups)									
Q1	1	2	3	4	5	Other	n	mean	s.d.
E	24%	62%	7%	7%	0%	0%	29	1.97	0.76
P	20%	76%	4%	0%	0%	0%	49	1.84	0.47
S	25%	67%	8%	0%	0%	0%	24	1.83	0.55
<i>F, sig, eta² :</i>							0.50	0.61	0.01
Q2	1	2	3	4	5	Other	n	mean	s.d.
E	34%	52%	10%	3%	0%	0%	29	1.83	0.75
P	18%	67%	14%	0%	0%	0%	49	1.96	0.57
S	46%	50%	4%	0%	0%	0%	24	1.58	0.57
<i>F, sig, eta² :</i>							2.83	0.06	0.05
Q3	1	2	3	4	5	Other	n	mean	s.d.
E	21%	66%	7%	7%	0%	0%	29	2.00	0.74
P	29%	47%	24%	0%	0%	0%	49	1.96	0.73
S	29%	58%	13%	0%	0%	0%	24	1.83	0.62
<i>F, sig, eta² :</i>							0.38	0.68	0.01
Q4	1	2	3	4	5	Other	n	mean	s.d.
E	7%	69%	21%	3%	0%	0%	29	2.21	0.61
P	6%	45%	45%	4%	0%	0%	49	2.47	0.67
S	0%	50%	29%	13%	8%	0%	24	2.79	0.96
<i>F, sig, eta² :</i>							4.05	0.02	0.08

It is sometimes easier to interpret these results if the option to get "response plots" is taken when the "Item responses by groups" analysis begins.

For example, the two following charts provide a visual summary of the group responses for Q4. We can now see that the response pattern for the secondary students differed: they were more negative, more to the right.

Of the two charts below, the first exhibits the [default chart type](#) and style used in Lertap5. The second was readily made by clicking on the first and then choosing from a variety of standard Excel charting options.

Lertap5 has a shortcut, a special 'macro', that makes it possible to copy the changes made to the first chart to all subsequent charts – refer to [this topic](#).

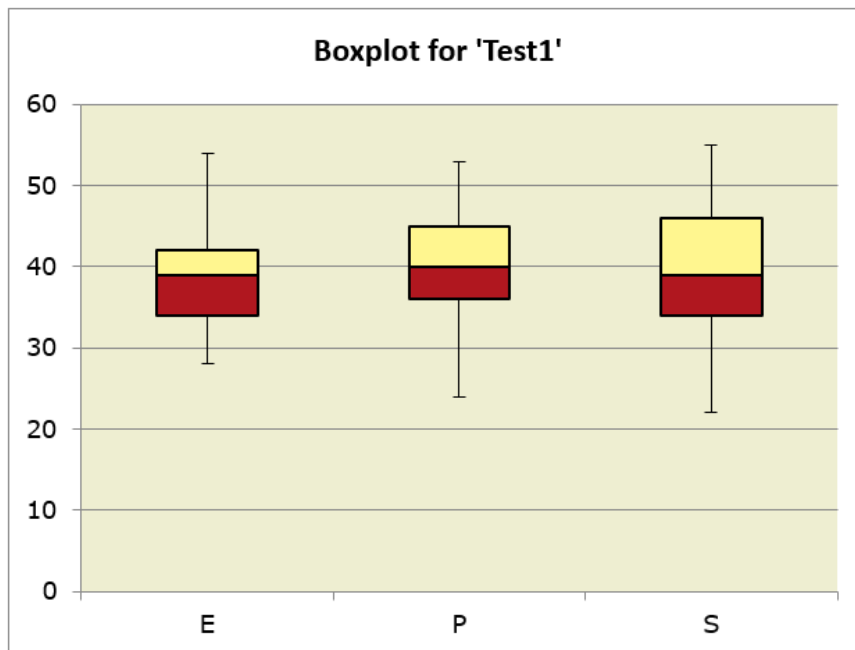


What about breaking out group results for the overall score? Can do? Yes, easy.
The "[Breakout scores by groups](#)" option is used to get tables such as these:

Test1	E	P	S
n	29	49	24
Min	28.00	24.00	22.00
Median	39.00	40.00	39.00
Mean	40.38	40.51	40.04
Max	73.00	61.00	55.00
s.d.	9.52	7.53	9.05
var.	90.65	56.74	81.96
Range	45.00	37.00	33.00
IQRange	8.00	9.00	12.00
Skewness	1.70	0.51	-0.24
Kurtosis	4.02	0.80	-0.58
MinPos	20.00	20.00	20.00
MaxPos	100.00	100.00	100.00

Analysis of variance			
	df	SS	MS
Between	2	4	2
Within	99	7376	75
Total	101	7380	
F ratio:	0.02	.98 (<-sig.)	
eta ² :	0.00		

A boxplot of results is created by using the "[Box and whiskers option](#)":



The bottom of each of the three boxes corresponds to the 25th percentile, the line in the middle to the 50th percentile (the median), and the top to the 75th percentile. We see that the median score was about the same in each group. The height of each box reflects the extent of the middle 50% of scores for each group; the E group's middle scores were less spread than those of the other groups, perhaps suggesting more score homogeneity in that group, at least in the middle.

The lines extending from the top and bottom of the boxes are referred to as "whiskers". The top whiskers extend up to the highest score in each group, excluding "outliers", that is, excluding the very extreme, atypical scores.

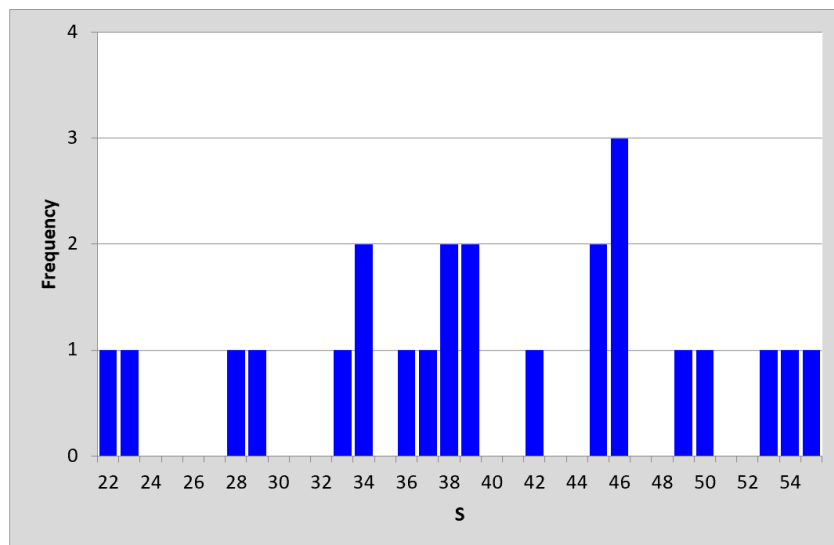
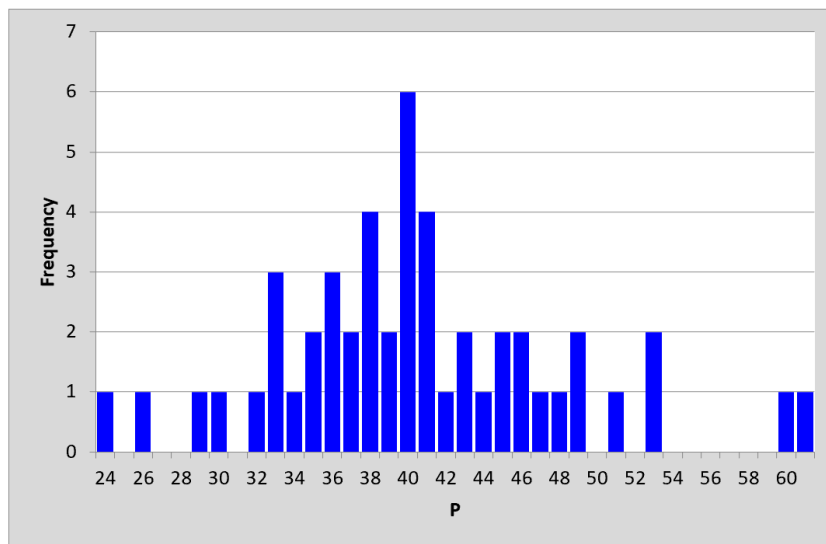
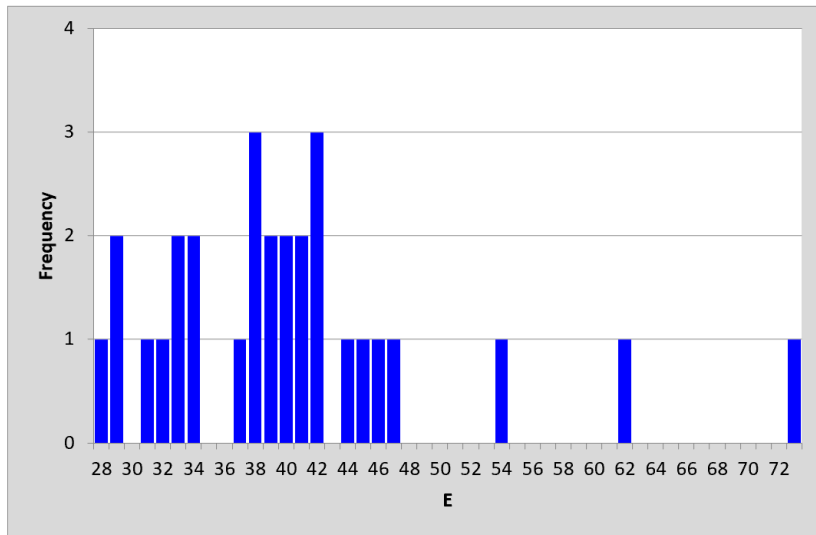
The bottom whiskers extend to the lowest score in each group, excluding outliers.

When looking at these results, keep in mind that in this case high scores correspond to negative responses.

Score histograms¹⁷ for each group are shown on the next page. All scores are reflected in the histograms, including outliers. The E group now shows as having two outliers, students with extreme negative scores (62 and 72). The P group also has two outliers on the negative score side (60 and 61).

More examples of using Lertap5 to breakout results are in [this paper](#).

¹⁷ Histograms for each group are made from "Breaks" reports [as described here](#).



Using an external criterion

It is possible to correlate the responses given to each item of any subtest with what's called an "external" score. In Lertap, an "external" score is any score found in a data set's Scores worksheet.

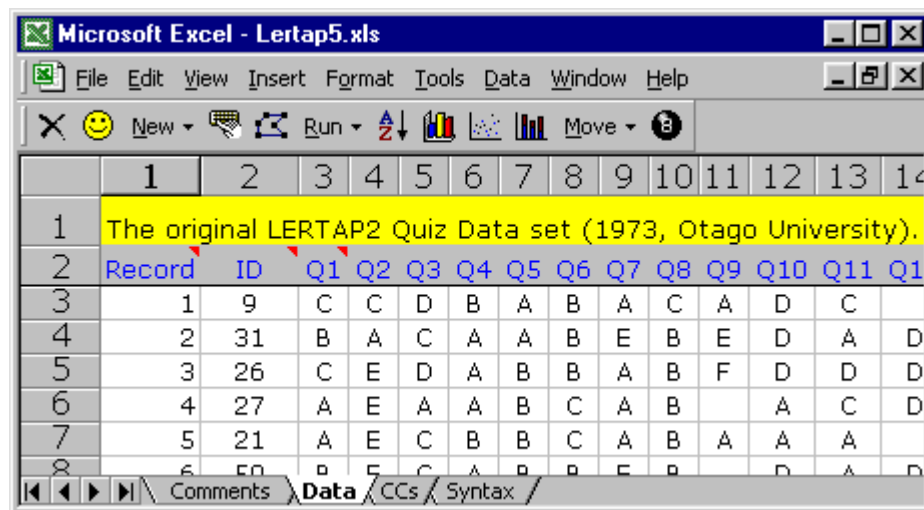
We'll walk you through two examples.

The Lertap Quiz data set has been mentioned above. Its 37 questions may be seen in [Appendix A](#). The actual data from this quiz are found in the Data sheet which comes as part of the Lertap5.xlsx workbook.

The Lertap quiz consists of 25 cognitive items, numbered Q1 to Q25, 10 affective items, numbered Q26 to Q35, and two free-response, or open-ended, questions, Q36 and Q37.

We want to do two things: (1) look at how responses to each of the ten affective questions correlated with a person's score on the 25-item cognitive test, and then (2), how responses to the same ten affective items correlated with Q37, denoted as "YrsTest" in the data set. Q37 asked respondents to indicate how long they had been using tests in their job.

If you want to follow this example on your own computer, you'd begin by looking at the data set which comes with Lertap:



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	The original LERTAP2 Quiz Data set (1973, Otago University).													
2	Record	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
3	1	9	C	C	D	B	A	B	A	C	A	D	C	
4	2	31	B	A	C	A	A	B	E	B	E	D	A	D
5	3	26	C	E	D	A	B	B	A	B	F	D	D	D
6	4	27	A	E	A	A	B	C	A	B		A	C	D
7	5	21	A	E	C	B	B	C	A	B	A	A	A	
8	6	50	B	E	C	A	B	B	E	B	D	A	D	

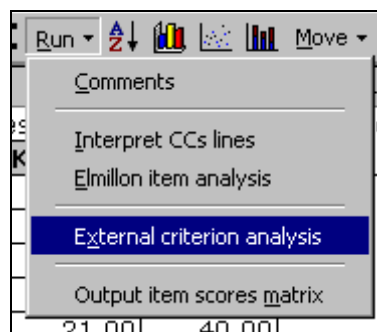
You'd use Lertap's [New menu](#) to "Make a new Lertap workbook which is a copy of the present one". Then you'd access the [Run options](#) to "Interpret the CCs lines", and to get an "Elmillion item analysis".

If you've followed these steps, your new workbook should look like this:

	1	2	3	4	5	6
1	Lertap5 scores worksheet, last updated on: 21/11/00.					
2	ID	Knwldge	Comfort			
54	37	8.00	33.00			
55	38	11.00	37.00			
56	11	4.00	31.00			
57	39	16.00	32.00			
58	60	21.00	40.00			
59	56	19.00	43.00			
60	15	3.00	33.00			
61	40	14.00	36.00			
62	46	18.00	40.00			
63						

The workbook above is displaying nine tabs, with the Scores sheet selected. Note that there are two scores, Knwldge and Comfort.

To look at how responses to each of the ten affective questions correlated with a person's score on the 25-item cognitive test, go to Lertap's toolbar, open the [Run menu](#), and then click on "Use external criterion"¹⁸.



Lertap will ask for the "column number of the score which will serve as the external criterion". It's referring to the columns in the Scores worksheet. In this case, the column with the scores to use as the external criterion is the second column—we want to use the Knwldge score as the external criterion.

The next bit of information which Lertap requests is the identification of the subtest whose items are to be correlated with the external criterion. Lertap will cycle through the subtests, one by one, pausing to ask if "...this is the subtest you want to work with". In this example, the first subtest is Knwldge, which is not the one we want. The second subtest is Comfort, and this is the one.

¹⁸ The entries on the Run menu have changed since this chapter was first written.

With this information in hand, the program is able to create a new worksheet for the second subtest. It will be called "ECStats2f", with contents as exemplified below:

Lertap5 external criterion stats for "Comfort with using LERTAP2", created: 21/11/00.

Q26

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
1	1.00	8	0.13	-0.42	-0.66	5.25	-1.06
2	2.00	13	0.22	-0.25	-0.35	9.38	-0.47
3	3.00	15	0.25	-0.21	-0.29	10.07	-0.37
4	4.00	14	0.23	0.35	0.48	17.00	0.63
5	5.00	10	0.17	0.51	0.76	20.50	1.13
r/ec:				0.71			

Q27

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
1	5.00	3	0.05	0.14	0.30	17.00	0.63
2	4.00	14	0.23	0.27	0.37	16.00	0.48

The correlation of item Q26 with the external criterion score, Knwldge, is **r/ec: 0.71**. This is a [product-moment correlation](#) coefficient. The pb/ec and b/ec columns give the [point-biserial and biserial coefficient](#) of each option with the external criterion; the avg/ec column indicates the average external criterion score for those respondents who chose an option—for example, the 8 people who selected option 1 on Q26 had an average external criterion score of 5.25. The last column in the report, z, expresses the avg/ec figure as a z-score, using the external criterion's mean and standard deviation to compute it.

A summary of the r/ec figures is provided at the end of ECStats2f report. For the ten items of the Comfort subtest, the summary turned out like this:

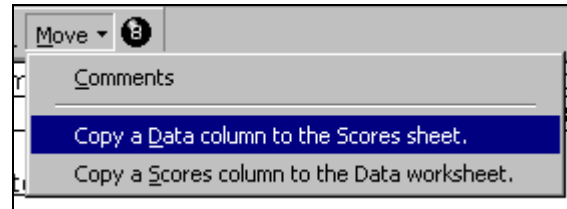
correlation bands (with external criterion)	
.00:	Q28 Q31 Q34
.10:	
.20:	
.30:	Q27
.40:	Q32
.50:	
.60:	Q29 Q30
.70:	Q26 Q33 Q35
.80:	
.90:	

Making a move for another external criterion

We posed two questions above. We've found the correlations of the affective items with Knwldge, the cognitive subtest, finding five items with high correlations: Q26, Q29, Q30, Q33, and Q35. But we also wanted the same correlations with another criterion, Q37, a question which had to do with the number of years respondents had been using tests in their work.

If we use the Run menu to request another external criterion analysis, Lertap will ask us to point out the column in the Scores sheet which has the score to use as the external criterion. Q37 is not there; it's in the Data worksheet, not Scores.

We need to use Lertap's [Move options](#) to copy Q37's column from the Data worksheet to Scores:



The use of this option is straightforward. We tell Lertap that column 39 of the Data sheet has the scores to be copied and pasted. Lertap checks to make sure that the column seems to have valid data, and then makes the copy, giving us an amended Scores worksheet:

	1	2	3	4	5
1	Lertap5 scores worksheet, last updated on: 21/11/00				
2	ID	Knwldge	Comfort	YrsTest	
3	9	3.00	32.00	3.00	
4	31	12.00	32.00	4.00	
5	26	13.00	37.00	4.00	
6	27	11.00	32.00	4.00	
7	21	14.00	33.00	2.50	
8	59	19.00	37.00	12.00	
9	47	14.00	42.00	6.00	
10	42	20.00	41.00	6.50	
11	55	20.00	41.00	5.50	
12	51	24.00	40.00	5.50	
13	20	12.00	34.00	3.00	
14	41	21.00	36.00	4.50	

Now we're free to go for another external criterion analysis, this time using the 4th column in the Scores sheet, and again indicating that the second subtest, Comfort, is the one of interest. Lertap goes off, but doesn't proceed quite as rapidly as before. It feels a need to know the maximum possible value for YrsTest, and we enter 60 (which seems a reasonable maximum value for someone to have worked with tests).

The next hurdle: Lertap announces that this subtest already has an external criterion report. And it's right. It does. The ECStats2f sheet from our first external criterion analysis is still there. We need to rename this worksheet, or delete it, so that Lertap can create a new one. Once we've done one of these things, Lertap creates a new ECStats2f sheet, again calling it ECStats2f. The lower part of the new sheet is shown below:

Lertap5 external criterion stats for "Comfort with using LERTA correlation bands (with external criterion)

.00:	Q28 Q31 Q34
.10:	
.20:	Q27 Q32
.30:	Q29 Q30
.40:	Q33
.50:	Q26
.60:	
.70:	Q35
.80:	
.90:	

Stats2f Stats2b 1stECStats2f **ECStats2f**

In the screen capture above, only two of the ten Comfort items, Q26 and Q35, have correlations with YrsTest in excess of **r/ec: 0.50**. Once again we see the trio of Q28, Q31, and Q34 having low correlations.

More correlations (completing the picture)

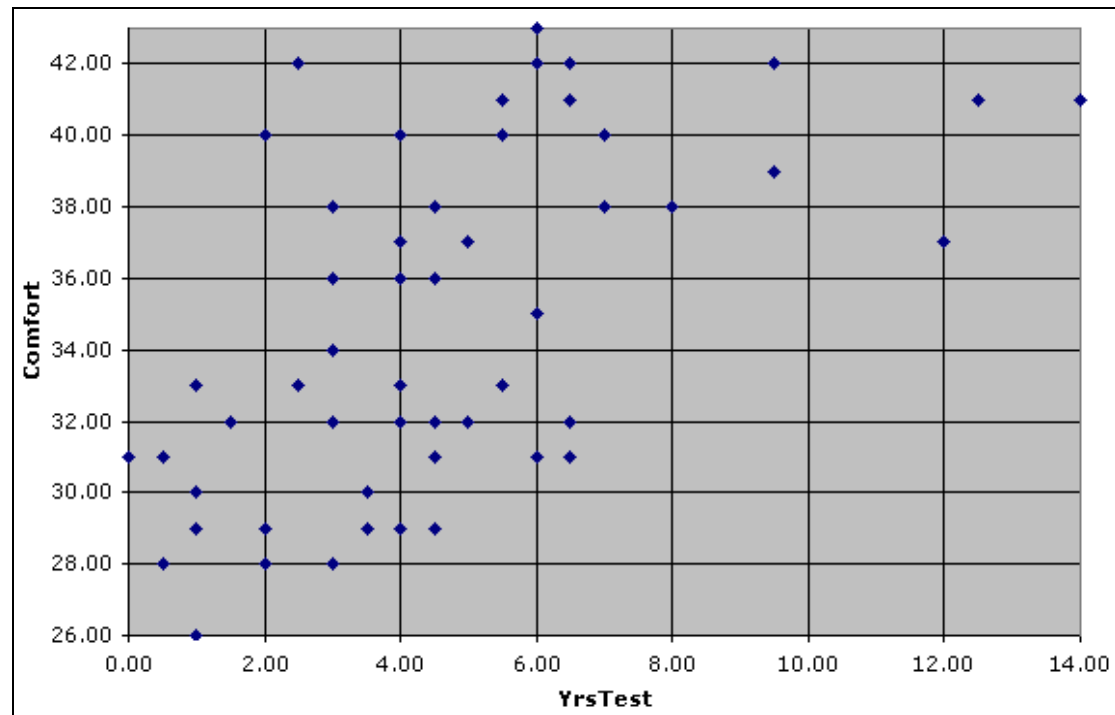
We could easily find out more about the correlation between YrsTest and the Comfort scale. For example, the Scores worksheet gives these values for the product-moment correlations between the three scores:

Lertap5 scores worksheet, last updated on:

ID	Knwldge	Comfort	YrsTest
MinPos	0.00	10.00	(unknown)
MaxPos	25.00	50.00	60.00
Correlations			
Knwldge	1.00	0.80	0.62
Comfort	0.80	1.00	0.59
YrsTest	0.62	0.59	1.00
average	0.71	0.70	0.61

Data CCs Freqs ScatP1 ScatP2 **Sc**

The correlation between the Comfort and YrsTest scores is 0.59. We asked for a [scatterplot](#) of these two variables, and this is what we got:



The scatterplot shows that the six respondents with the most test experience, 8.00 years or more, had high Comfort scores. If we take as a low Comfort score any score below 35.00, the plot reveals that all of the people with low scores had less than seven years of experience with tests. We would have some justification for suggesting that the veteran test users in the group of 60 felt “comfortable” with the use of Lertap 2. This isn’t an entirely satisfactory outcome—ideally, everyone would be tickled pink with the program, no matter how much experience they’d had with the use of tests. Perhaps a mitigating factor was computer usage; could it be that those with low Comfort scores were also those with little computer experience? This was back in 1973. Personal computers had not yet appeared; users had to go to the computer centre to punch cards, and sit around waiting for results. Maybe we should investigate the correlation between Comfort and the YrsComp score, column 38 of the Data worksheet? We’ll leave that as an exercise for you to pursue.

Remember the Comfort2 subtest defined above? It was composed of just seven of the ten Comfort items—a subtest free of the trio of Q28, Q31, and Q34. We have already seen how this Comfort2 scale had a higher correlation with Knwldge than did Comfort. Would we note a similar improvement when correlating Comfort2 with YrsTest? No. With some surprise, we noted that the correlation of Comfort2 with YrsTest came out to be 0.61, not much of a difference. We also looked at the scatterplot of Comfort2 and YrsTest, and found it to look very similar to the original one shown above.

Further analyses

We have mentioned, on several occasions, that the SPSS data analysis system provides support for more extensive analyses. The [next chapter](#) discusses this in more detail.