Item Analysis for Tests and Surveys

Using Lertap 5

Manual date: 10 December 2001

Larry Richard Nelson

Faculty of Education
Curtin University of Technology
Perth, Western Australia

Item Analysis for Tests and Surveys

Using Lertap 5

This book focuses on the application of the Lertap 5 software system to the analysis of test and survey items.

NOTE: Updated software and documentation are readily available at www.lertap.com. And friendly user assistance is at support@lertap.com.

Electronic copies of the book (also known as "the manual"), and individual chapters, may be downloaded from this link.

Created in Perth, Western Australia, by Curtin University of Technology, December, 2001. This version is for both A4 and American "letter" paper sizes.

Item Analysis for Tests and Surveys Using Lertap 5

Table of Contents

Acknowledgments

It was support from colleagues at the Faculty of Education, Curtin University of Technology, later bolstered by Curtin's Division of Humanities, which enabled this project to get off the ground, freeing me from a year of normal academic responsibilities so that this *opus magnificus* could be completed. Graham Dellar was foremost in providing encouragement for the original idea, and he saw to it that, once things were underway, I remained sheltered behind a virtually unbreachable sabbatical firewall.

It was the efforts of many beta testers which made it possible to release a working version within the time span originally envisaged. Thanks go, first of all, to Todd Rogers of the University of Alberta's Centre for Research in Applied Measurement and Evaluation (CRAME), and to two of his graduate students, Keith Boughton and Tess Dawber. The CRAME team provided a wealth of feedback and suggestions—most of the features found in Lertap 5's support for mastery test analyses link directly to the group at Alberta.

Staff at the Faculty of Education, and the Faculty of Nursing, Burapha University, Thailand, provided a most comfortable home for system development and testing. I am particularly grateful for the debugging assistance and challenges provided by Nanta Palitawanont, and for the warm support of Seree Chadcham and Suchada Kornpetpanee of Burapha's Educational Research and Measurement Department. The hands-on postgraduate lab sessions arranged by Drs Chadcham and Kornpetpanee were invaluable. Dean Chalong Tubsree made this visit possible, and I thank him and his staff.

Carlos Gonzalez of la Universidad Central de Venezuela was the first to process very large data sets with the prototype; Carlos put many thousands of records through the system, using both cognitive and affective instruments. His tests and feedback did much to stabilise the original year 2000 version. Luis Pulido of Proinvesca, Caracas, later continued with large-scale processing. It was the work of Luis which enabled tests of some of the advanced item-weighting schemes.

Ed van den Berg of the Philippines' San Carlos University, working with some of his graduate students, provided very useful feedback on system design and user interfacing. Ed's students were among the first to make good use of the "Spreader", and their experience with Lertap's job specification statements (control "cards") substantiated my hope that new users would be able to master the language without trouble.

Thanks to Piet Abik of Pontianak, Indonesia, for his efforts to help debug the new system in the midst of real political upheaval. Piet has been a stalwart supporter over many years; his work on the third version of Lertap did much to get the system into the hands of many Asian colleagues. I know he would have been on

hand for more testing of the new version had it not been for the turmoil sweeping his city and province in the year 2000.

I express my appreciation for the assistance and encouragement provided by Ian Boyd, a former Curtin colleague now at one of Perth's Technical and Further Education campuses. Ian has been a keen advocate of the use of software in many aspects of testing and assessment, and his persistent comments on the need for criterion-referenced support were directly responsible for Lertap's initial foray into this area.

Thanks to my wife, Khin Khin Than (Angie), for her enthusiastic backing, for tutorial help in our hands-on lab sessions, and, especially, for the many hours spent proofreading this *trabajo de amor*. Should errors remain, they will be found on pages she didn't see.

Finally, thanks to decades of Lertap users. This new version has been a long time in coming. I trust, or at least fervently hope, that it's good enough to justify the wait.

I extend a cordial invitation for readers to contact me via the email address shown below.

Larry Nelson
L.Nelson@curtin.edu.au

Chapter 1

Lertap Comes Alive as Version 5

Contents

Lertap? You'll love it. It's an item, test, and survey analysis system for instructors, teachers, and researchers.

This version of the Laboratory of Educational Research[1] Test Analysis Package is the fifth in a series born in Caracas in 1972. Since then it has travelled the world, finding developmental homes at the University of Colorado, the University of Otago, and Curtin University of Technology. Its growth has been assisted by colleagues working at universities, private and public research agencies, and schools, from Canada through the United States, down to Venezuela and Peru, across the Pacific to the Philippines, Indonesia and Thailand, south to New Zealand and Australia.

Lertap 5 has been created as an application which runs under Microsoft's spreadsheet program, Excel. As such it runs on any platform which supports Excel, including Windows, NT, and Macintosh computers.

This version differs from earlier ones in several ways. The most substantial difference is the move from a stand-alone system to an Excel base, a move which has virtually guaranteed a user interface that will remain current. An Excel base means a near-universal database structure—many other systems can write to, and read from, Excel worksheets. Other major differences include a complete reformatting of former reports pages, the inclusion of many new reports and statistics, and the revival of the Lertap job definition language.

You'll love it.

---

[1] LER was at the University of Colorado, Boulder.

Basic question types (the input)

There are two fundamental types of questions which Lertap users apply in their teaching and research: cognitive questions, and affective questions.

Here's an example of a *cognitive* question[2]:

(10) *MWS control cards are used

a) for achievement items to multiply-weight the item (indicate that more than one response is to receive a non-zero weight), or to indicate that not all response codes are used by an item.
b) for affective items to alter the pattern of forward or reverse weighting established by preceding control cards, or to indicate that not all response codes are used by an item.
c) to override item control information entered on any or all previous control cards.
d) all of the above.

An example of an *affective* question, or item[3]:

		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

Of course, questions such as these are seldom used in isolation—you'll have a series, or set, of questions to process. This set may be referred to as a test, a subtest, or a scale, depending on the situation.

How many questions may be in a single test? The answer is determined by Excel, having to do with the number of columns it will allow in a worksheet. At the present time the limit is 256. How many subtests may there be? In theory there is no limit; in practice the answer to this question will depend on the amount of memory your computer lends Excel to work in.

How many test takers may be processed? Again the answer is determined, in the first instance, by Excel. Excel 97 and Excel 2000 limit the number of cases (rows) to 65,356. Some Lertap sites have processed data sets having fifteen thousand records. There may also be a practical operational limit—a computer with limited memory, for example, may mean that only a few thousand records may be processed.

Our marketing section wants us to mention Lertap strengths in this chapter. That would make for a very long chapter, indeed, so here we will just toss out some *hors d'oeuvres*, a tantalising sample of the multitude of dazzling features which will come to light as you progress through the book: questions may use up to ten response codes, and these may be letters or digits. (Above, the cognitive item is

[2] This question was lifted from the Lertap Quiz, Appendix A.
[3] From Chapter 8. This sample is from the University of Michigan's MSLQ instrument.

using four response codes, all letters, while the affective item is using seven digits.) A cognitive question may have more than one right answer, with different points given to different responses. Affective items may be reverse scored with the flick of the plus and minus keys on your keyboard. Formula scoring may be applied to achievement tests. Missing responses to affective items are not a worry. Responses to affective items may have negative weights. Any number of cognitive and affective subtests may be processed simultaneously.

The responses people give to your questions are entered into a standard Excel worksheet. Lertap comes with a template, and some data entry tools, which make this job easier. Another worthy feature: your items may have almost any handles you want. You can call them, for example, something as imaginative as "1" (for the first item), or "Q1", or "Item 1", "Soal 1", "PreguntaX", "Love-Handle R", and so forth.

Speak Lertap (the intermediary)

A strong feature of this new version is that it allows all those people who learned how to speak Lertap in the 70s to recall their vocabulary. You talk to Lertap using sets of sentences such as these:

```
*col (c2-c11)
*key bccda ddcca
```

These two Lertap utterances would be placed in an Excel worksheet. They tell Excel that you've entered ten item responses in columns 2 through 11 of the data worksheet, and that the keyed-correct answers to the ten items are as indicated on the *key line above.

Here's some more Lertapese, or "Tap":

```
*col (c1-c10)
*sub affective
```

This dynamic duo gets Excel to look for ten item responses in columns 1 through 10 of the data worksheet, and to see to it that Lertap processes the responses as affective questions. What's that? These are Likert-style questions, and every other one is to be reverse scored? Okay, your chance to talk more Tap; you'd say something like this:

```
*col (c1-c10)
*sub affective
*pol +-+-+ -+-+-
```

The *pol line means, of course, scoring "polarities".

Once you get the hang of things, you can be more flowery in your use of the lingo. For example, another way to say the three sentences above would be:

```
*columns (c1-c10)
*subtest affectively_scored
```

*polarities +-+-+ -+-+-

Feeling young again? Get cool, mutter code:

*c (c1-c10)
*s a
*p +-+-+ -+-+-

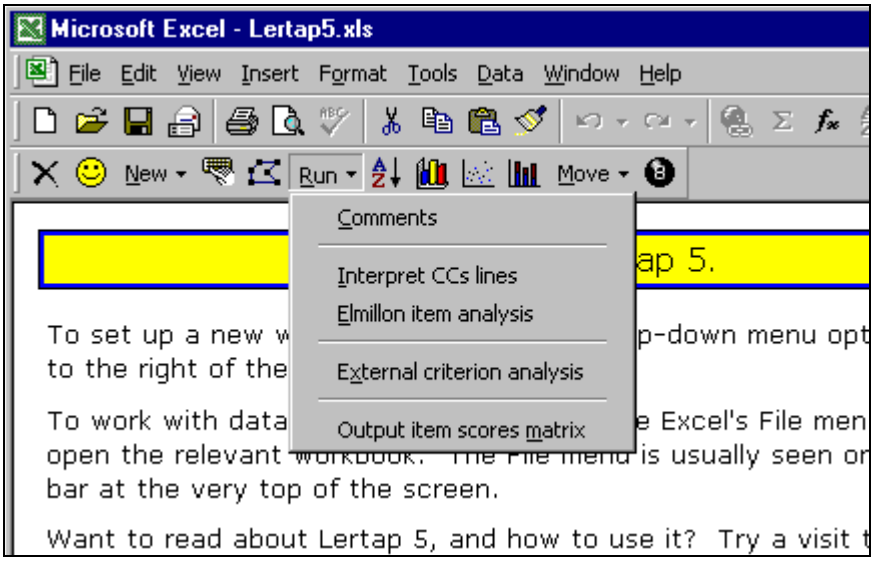
You've heard of rap? With Lertap you speak Tap. Look at some of the rest of this book, and you'll soon be 'tappin and 'rappin. PlayStation 3 look out.

The third and fourth versions of Lertap foolishly abandoned Tap. Instead they used tens of dialog boxes in the grande style which was fashionable for a while. What you could Tap out in just two lines became some fourteen dialog boxes. This was yuk; we've gone back to Tap in this version. You'll love it.

Scores and reports (the output)

So. You have people answer your questions. You put their answers into an Excel worksheet. You talk Tap in another worksheet. Then what? You get the Runs. You use a special Lertap toolbar in Excel, and apply its Run options.

Here's a picture of Excel, Excel's standard toolbar, Lertap's toolbar with the yellow smiley face, and the Run options exposed:



Look at all those icons, just waiting for your mouse to click on them—you're set for a whale of a time. Notice that Lertap will even let you get behind the 8-ball (the last icon on Lertap's toolbar).

The rest of this book spends much of its time showing what happens when you use the Run options—you can get a quick sample by flipping through the next chapter.

Lertap's standard reports for cognitive tests include several classical statistics, such as: standard correlation-based analyses, using point-biserial, biserial, and conventional product-moment coefficients, corrected for part-whole inflation when required; standard upper-lower group breakouts, reporting the "D" index of item discrimination, and, when appropriate, Brennan's generalised "B" index; mastery-test indices, with a Brennan-Kane variance partitioning, and Peng-Subkoviak estimate of classification consistency. Indices of reliability include coefficient alpha, and, in the mastery-test case, the Brennan-Kane index of dependability.

Reports for affective items, that is, for surveys, include item-criterion correlations, corrected for part-whole inflation; item means and standard deviations; coefficient alpha, and a table indicating how alpha would change if items were omitted.

All Lertap analyses may be conducted with an external criterion. Item response frequencies may be charted (graphed using a standard Excel 3-D bar chart).

Test and scale scores may be plotted using Excel's histogram and scatterplot support[4], they may also be rearranged using standard Excel data sorts. Score intercorrelations are standard output, found at the bottom of the Scores report.

An item scores report, or "matrix", may be requested for any test or scale. It has three sections: a conventional matrix of item scores, a summary of item means and variances, and an interitem correlation matrix.

All Lertap 5 reports are Excel worksheets. Its graphs are basically Excel charts of one sort or another. The worksheets and graphs may be printed, copied, pasted, you name it—it's Excel through and through.

Using with other systems

You may have heard of SPSS, the Statistical Package for the Social Sciences[5]?

Next to Lertap (rather arguably), SPSS is (hardly arguably) the most popular data analysis system applied in education, sociology, and psychology. However, after 30+ years, it continues to lack support for scoring and analysing cognitive tests. Its help for users of affective scales is still based on the "Reliability" routine, a subprogram which, after how many years?, continues to make it cumbersome to reverse-score questions, and to make composites. Lertap exists, in part, because of these holes in SPSS[6].

Lertap shines at putting together test and scale scores, but, ah, okay, it might be suggested, by some, that SPSS has maybe just a teeny-weeny bit more support for users who want to do others things with their scores, such as regression analyses, or a means analysis leading to some dandy Boxplots. What about

[4] The histogram requires Excel's Analysis ToolPak; see Chapter 10.
[5] www.spss.com
[6] An SPSS-Lertap powwow was held in Chicago in 1978, with the larger of the two software houses deciding against a merger, or even a plain old lucrative buyout.

poking one or more of Lertap's item scores matrices into some sort of factor analysis? Yes, SPSS could be your man in such cases[7].

Lertap's 8-ball will help you prepare your lovingly-sculpted worksheets and reports for export to SPSS.

The rest of the book

The next chapter introduces you to the Runs, showing what happens when you make some of the clicks suggested above. That's followed, in Chapter 3, by a discussion of how to prepare data for the Runs, including how to begin to talk Tap.

Chapters 4, 5, and 6 delve more extensively into Tap, that is, into job definition statements. Then there are two chapters, 7 and 8, which help interpret Lertap's cognitive and affective test reports. Chapter 9 makes more mention of SPSS, and how to interface to it. Chapter 10 is a bit more technical, getting into computational methods, and summarising some of the time trials completed to date. Finally, Chapter 11 looks at the history of Lertap. You'll love it. Get those fingers and feet Tapping.

[7] Experienced SPSS users will recognise the intended humour in this paragraph. For readers with less experience, we might note that SPSS has *extensive* support for many different types of statistical analyses.

Chapter 2

A Cook's Tour of Lertap 5

Contents

Our objective in this chapter is to get you started as quickly as possibly. In the next few pages we'll have you getting into an actual test drive of the software. However, first a few important pointers.

This version of Lertap has been developed as an application which runs within **Excel**. Excel is part of a suite of programs known as Microsoft **Office**, often simply referred to as "Office", usually with the version number appended, such as "Office 97", "Office 98" (a Macintosh version), or "Office 2000".

Excel is a spreadsheet program. Its basic "pages" are worksheets divided into rows and columns where information is entered, stored, and displayed. At times worksheets include "charts", that is, graphical displays of information.

It is often easy to mistake an Excel worksheet for an ordinary page in a document prepared by a word processor, such as **Word**, another part of the Office suite.

This potential confusion arises when Excel has been directed to hide its fundamental worksheet structure, something which may be done by selecting options which, for example, turn off the display of "gridlines", and "row and column headers". Many of the worksheets used in Lertap have their gridlines and headers hidden. These sheets may look like ordinary pages, but they're not—they have Excel's columns and rows propping them up.

Remembering this can be helpful at times. For example, it's easy to change the appearance of Lertap "pages" by making their columns wider, or their rows higher, as one often wants to do when using tables in a word processor. To find out how to do these things, and how to turn the display of gridlines and headers back on, Excel's extensive on-line Help system is an invaluable aid.

The files used in a word processor, such as Word, generally use the "doc" extension. The name of the file for this chapter, for example, is Chapter2.doc. In Excel, on the other hand, the common extension is "xls". The name of the file which contains Lertap is **Lertap5.xls**.

Ready to get rolling?

Lertap5.xls

Find the Lertap5.xls file on your hard disk, or on your applications server, and open it.

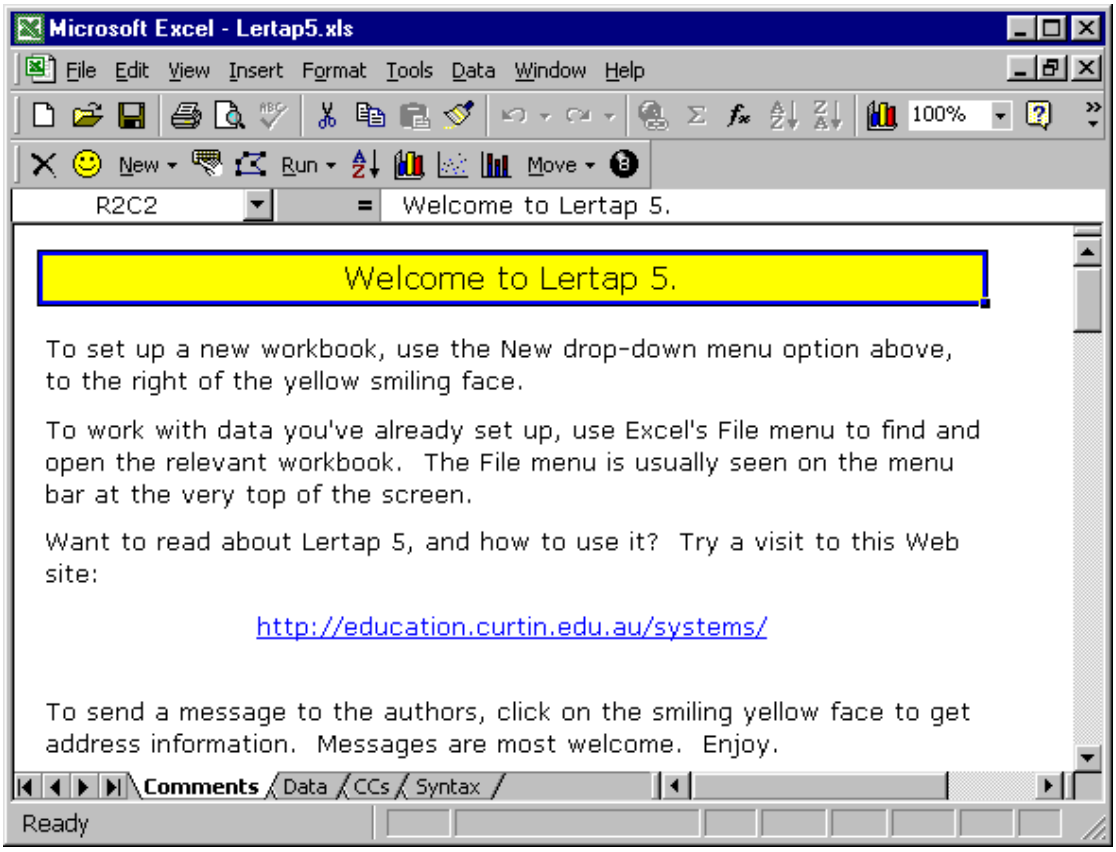
This file contains "macros". Lots of them. The macros contain the computer programming code[8] which forms the basis of Lertap. Lertap will not run until its macros are loaded, or "enabled", as Excel sometimes says.

If Excel was not already running when you opened Lertap5.xls, it will be automatically invoked by the operating system on your computer. If the operating system cannot find Excel, you'll be informed, and will need to seek help.

Starting with Version 5, an Excel xls file came to be referred to as a "workbook". A workbook is a collection of worksheets. You'll see an example of such a collection when you open the Lertap5.xls file.

When the Lertap5.xls file is opened, you should see a screen which resembles the one shown below:

[8] Lertap 5 is written in *Visual Basic for Applications*.



Here are some things to note at this point:

1. One menu bar and two toolbars are showing at the top of the screen. The menu bar begins with options named File, Edit, and View.

2. The upper-most toolbar is from Excel. It's called the Standard toolbar. There are many other toolbars available in Excel. To display, for example, the Formatting toolbar, click on the View option, select Toolbars, then select Formatting.

3. The second toolbar is from Lertap. This is the one with the smiley yellow face towards its left-hand side. If this toolbar is not showing on your computer, it means that Lertap has not loaded correctly. This will happen if macros are not enabled. Macros must be enabled in order for Lertap to be able to do its things.

4. Below the Lertap toolbar the Excel Formula Bar is showing that cell R2C2 of the current worksheet is presently selected. R2C2 means Row 2, Column 2. Rows run horizontally across the screen, while columns run up and down. The worksheet we're looking at above has had its gridlines and row and column headers hidden, and, as a consequence, it's difficult to see where the actual rows and columns are. But they're there, to be sure.

   The Formula Bar also shows the contents of this cell to the right of the =

sign, "Welcome to Lertap 5".

5. The Lertap5.xls workbook is displaying two scroll bars. There's the usual vertical scroll bar at the very right of the screen, and, down towards the bottom, there's a not-so-usual horizontal scroll bar with two distinct parts to it.

   Of these two parts, the right-most one is the conventional horizontal scoller, one which lets you move from left to right on the "page" (or worksheet). The left-most part lets users scroll among the various worksheets which belong to the workbook.

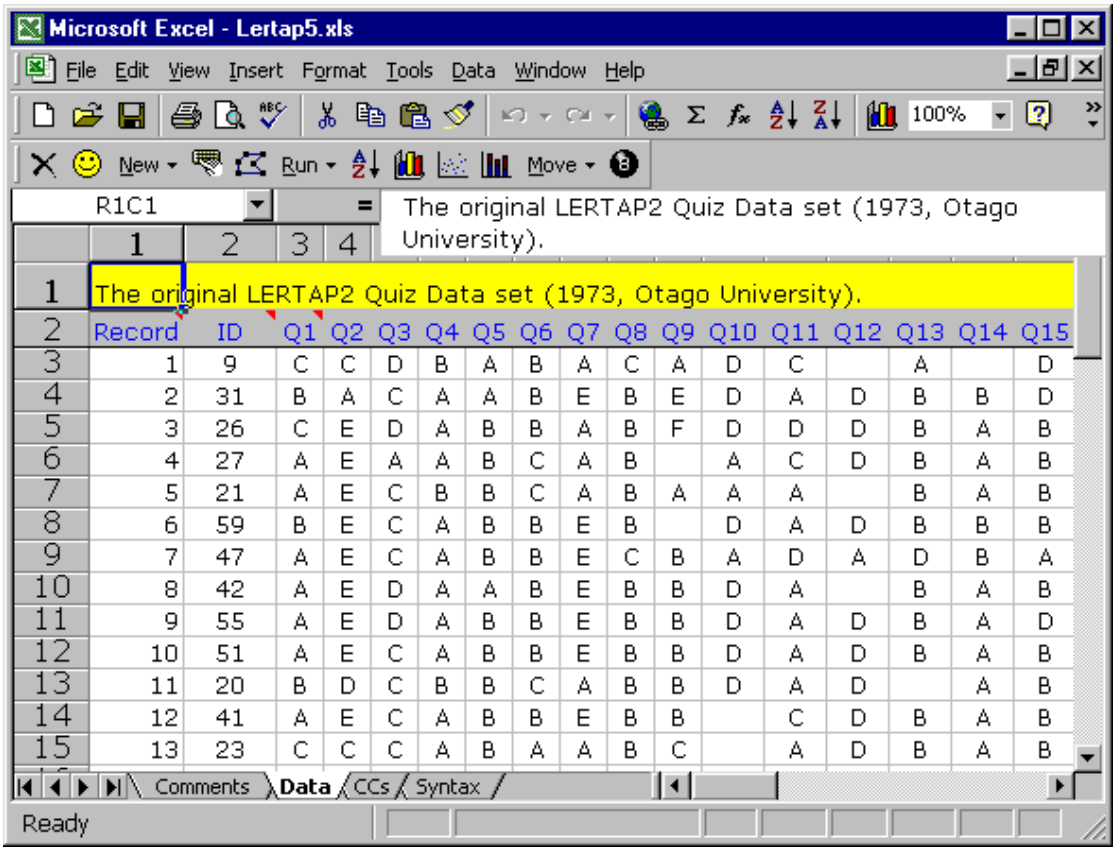
The Lertap5.xls file has four visible worksheets. Their names are shown above as "Comments", "Data", "CCs", and "Syntax" on the respective worksheet tabs seen towards the bottom of the screen.

As Lertap goes about its business it will add more worksheets to the workbook. Each sheet will have a name and a corresponding tab. The number of tabs can come to quickly exceed the space available for them in the horizontal scroll bar, and this is when the little arrow heads to the very left of the scroll bar become useful.

The four worksheets

Now it would be useful to have a brief browse of Lertap's built-in worksheets.

The **Comments** sheet, the one captured above, is the first sheet to show when the Lertap5.xls file is opened. It isn't much, just a welcoming title page. If you're connected to the Internet, a click on the hyperlink shown on the Comments sheet will take you to Lertap's home page (which may also be reached by having your Web browser go to www.lertap.com).



	Record	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
1	The original LERTAP2 Quiz Data set (1973, Otago University).																
3	1	9	C	C	D	B	A	B	A	C	A	D	C		A		D
4	2	31	B	A	C	A	A	B	E	B	E	D	A	D	B	B	D
5	3	26	C	E	D	A	B	B	A	B	F	D	D	D	B	A	B
6	4	27	A	E	A	A	B	C	A	B		A	C	D	B	A	B
7	5	21	A	E	C	B	B	C	A	B	A	A	A		B	A	B
8	6	59	B	E	C	A	B	B	E	B		D	A	D	B	B	B
9	7	47	A	E	C	A	B	B	E	C	B	A	D	A	D	B	A
10	8	42	A	E	D	A	A	B	E	B	B	D	A		B	A	B
11	9	55	A	E	D	A	B	B	E	B	B	D	A	D	B	A	D
12	10	51	A	E	C	A	B	B	E	B	B	D	A	D	B	A	B
13	11	20	B	D	C	B	B	C	A	B	B	D	A	D		A	B
14	12	41	A	E	C	A	B	B	E	B	B		C	D	B	A	B
15	13	23	C	C	C	A	B	A	A	B	C		A	D	B	A	B

The Comments sheet may not amount to much, but the **Data** sheet, shown above, is packed with things to look at. It contains the responses of 60 people to 37 questions.

Of these, the first 25 questions have their responses recorded in columns 3 through 27 of the worksheet. The responses to these 25 questions, Q1 through Q25, consist of letters. These questions corresponded to a cognitive test designed to indicate how well the respondents had mastered an introductory workshop on the use of Lertap 2, a version which appeared in 1973.

The second set of questions, Q26 through Q35, have their responses in columns 28 through 37 of the worksheet. These ten questions were affective in nature, asking respondents to report their attitudes towards using Lertap 2. A 5-point Likert scale was used to record answers.

The last two questions had to do with the number of years which respondents had been using computers and tests in their work.

How many rows are used by the Data worksheet? If you have opened the sheet on your own computer, you'll find that 62 is the answer. The first row has a title, the second has headers for the data columns, and the remaining, beginning in row 3, have the answers given by the respondents, one row for each respondent.

Some of the cells in the worksheet seem to be empty. Cell R3C14, for example, appears to have nothing in it. The respondent whose answers were recorded in row 3 of the Data worksheet did not answer Q12 (nor Q14; see R3C16).

Note how we're saying that these cells "seem" or "appear" to be empty. We say this as it's possible there's a space, or blank, recorded in the cell (in fact, we know this to be the case—unanswered questions were processed by typing a space, which Excel calls a blank). Blank cells are not empty, even though they certainly appear to be.

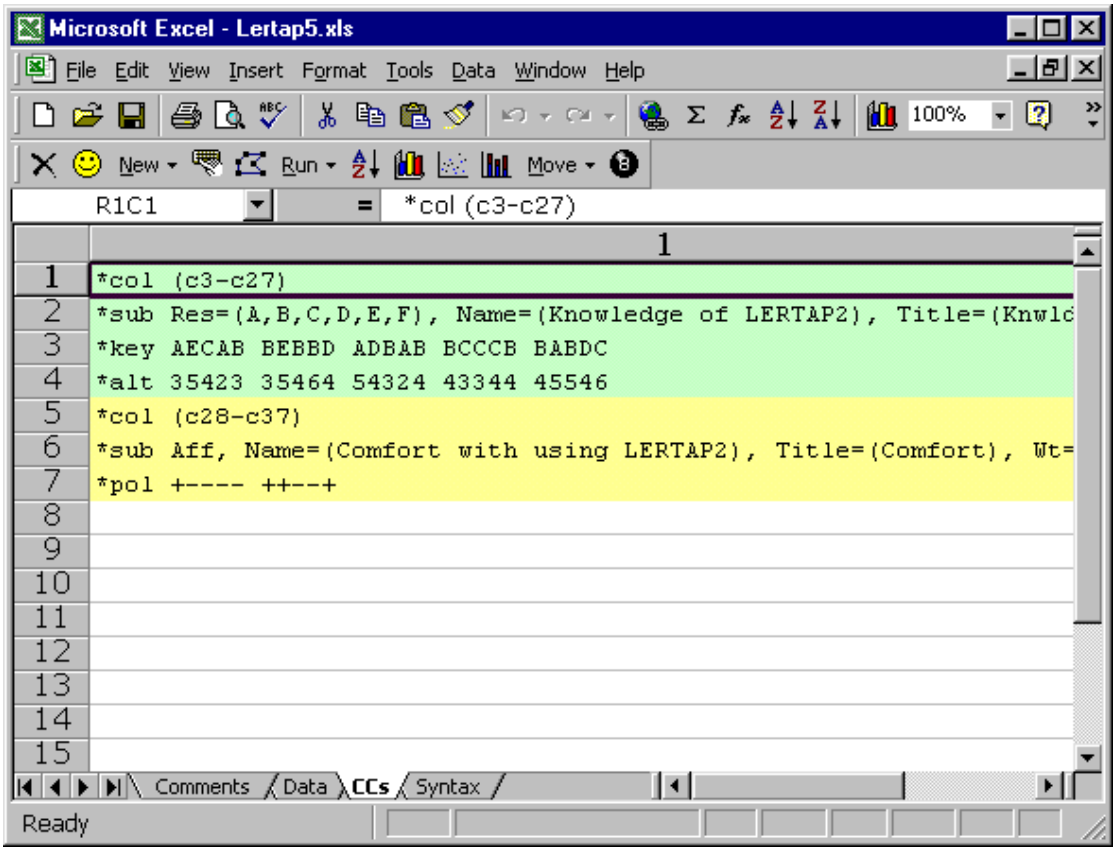
Before looking at the next sheet, CCs (for "Control Cards"), it will be worthwhile to summarise why we'd want to use Lertap to analyse the results seen in the Data worksheet. We could think of some "research questions" which we'd like to have answers to.

The results are from two tests, one cognitive, the other affective. We'd like to have a summary of response-, item-, and test-level data. We'd like to know, for example, how many people got each of the 25 cognitive items correct. Which of these items was the hardest for the 60 respondents? What did overall test scores look like? Do the tests appear to have adequate reliability?

The 10 affective items asked respondents to reveal how they felt about their introduction to Lertap 2. What was it they liked and disliked? How did their attitudes correspond to how well they did on the cognitive test?

We might also like to know how responses to the cognitive and affective questions may have related to the experience levels of the respondents. If they had been using computers for some time, were their attitudes towards the software more positive, or more negative?

In order to have Lertap answer questions such as these, we first need to provide the system with what are referred to as job definition statements. This is done in the CCs worksheet.



The **CCs** sheet shown above has 7 rows of information. The first four have to do with the first test, or "subtest".

The *col (c3-c27) line tells Lertap where the responses to the subtest's items are to be found in the Data worksheet.

The *sub card says that item responses were recorded as upper-case letters, A through F, and gives a name and title to this subtest. Titles are limited to eight characters in length; Lertap (and SPSS) users get to be quite creative in coming up with 8-character titles (called names in SPSS) which are recognisably mnemonic.

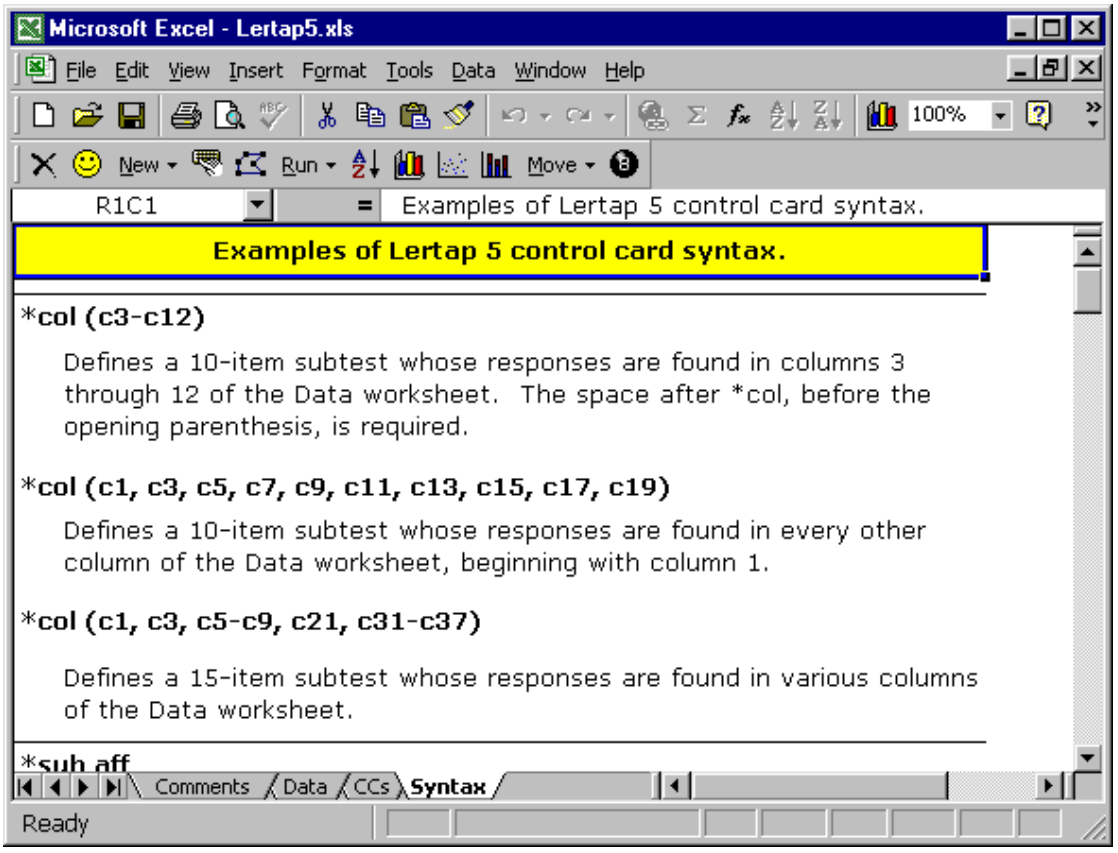
The *key line gives the correct answer to each question, while the *alt card indicates that the items did not all use all of the response letters shown in the *sub card.

There are 25 entries on both the *key and *alt lines, corresponding to the 25 items on the subtest. The correct answer to the first item, or question, was A, and this item used just 3 of the possible response letters. The correct answer to the second item was E, and this item used 5 of the possible response letters.

The *col (c28-c37) line signals to Lertap that there's another subtest to be processed. Answers to its items are found in columns 28 through 37 of the Data worksheet. The *sub card informs Lertap that this subtest is to be processed as

an affective one ("Aff"). The last line, *pol, indicates that some of the questions were positive in nature, while others were negative. An example of a positive "question" would be "This software is terrific!", while the corresponding negative form might be "This software is terrible!".

This is an example of a worksheet where gridlines and row and column headers have not been hidden. They're also visible in the Data sheet shown earlier. Notice how the first column in the CCs sheet has been stretched so as to be extra-wide? Notice how the initial columns in the Data worksheet are all narrow in comparison?



The last of the four visible worksheets in the Lertap5.xls workbook is named **Syntax**. This sheet is used to provide examples of what job definition lines can look like; it's meant to serve as a sort of on-line reference which may obviate the need to refer to this Guide.

Users may add their own examples to the Syntax worksheet. How to do this is discussed in Chapter 10, Computational Methods.

Summary

The Lertap5.xls file is a workbook which contains four visible worksheets, and the collection of macros which effectively define the Lertap software system. The four worksheets are named Comments, Data, CCs, and Syntax.

Of these, the first and the last, Comments and Syntax, are information sheets. The Data and CCs sheets, on the other hand, are much more substantial, content-wise. They exemplify what a dinkum[9] Lertap Excel workbook looks like.

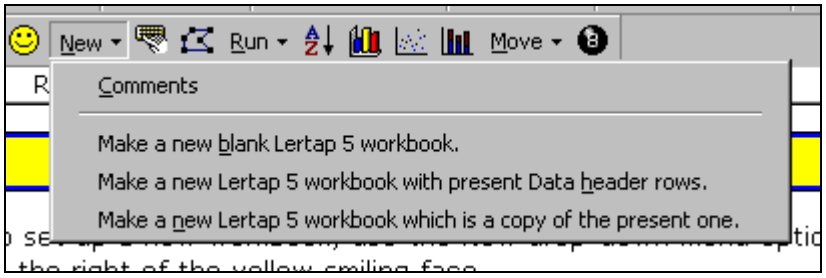
Let's gather some speed. Ready to see some action? Open the Lertap5.xls file on your computer, and read on. (Maybe get a fresh cup of coffee or tea first.)

Setting up a new Lertap 5 workbook

We're going to get you to set up a new Lertap workbook, and then show you how to do some actual data analysis. You must have the Lertap5.xls file open and ready to use.

What we'll do is have you begin by making a copy of the Lertap Quiz data set, or workbook.

Go to the Lertap toolbar, and click on the New option on the Lertap toolbar.



Read the Comments if you'd like. Then click on the last option, the one which says "Make a new Lertap 5 workbook which is a copy of the present one".

You're likely to use this option quite a lot in the future, as your Lertap career gathers momentum. You'll develop your own data sets, and some of them will probably becomes templates, that is, examples which will be used over and over. Actually, it's even more likely you'll use the option which says "Make a new Lertap 5 workbook with present Data header rows". This option creates a copy of your "template", but leaves the data lines empty, ready for you to add new results.

Okay, then, here we are. You've clicked on "Make a new Lertap 5 workbook which is a copy of the present one". What did Lertap have Excel do?

It should have created a new workbook with two worksheets. The worksheets should be copies of the Data and CCs sheets seen in the Lertap5.xls file. Excel probably named the new workbook "Book1", or "Book2", or something like that.

How can you tell the name of the workbook? Look way up at the top of the Excel window. The very top line, the window header, will have an Excel icon, the words "Microsoft Excel", followed by the name of the workbook.

[9] *Dinkum* is a word used in Australia to mean "genuine", or "authentic".

You can have lots of workbooks open at the same time. Switch from one to the other by working through Excel's Window option, available on Excel's Standard toolbar.

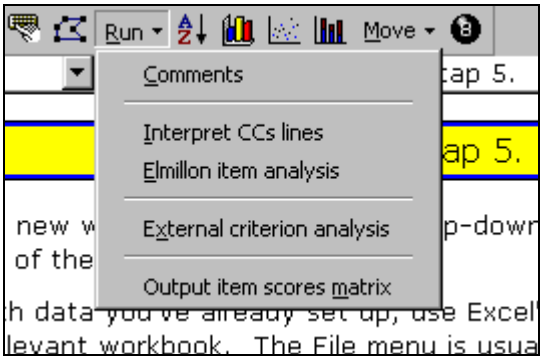
So, what happened? Do you have the new workbook ready? Its Data worksheet has two rows of header information, followed by 60 rows of results? Its CCs sheet has 7 lines of job definition statements?

Good. Green light. Proceed.

Interpret CCs lines

Let's ask Lertap if it can understand the definition statements found in the CCs file.

Click on Lertap's Run option, read the Comments if you'd like, and then click on "Interpret CCs lines".



Lertap goes off and has a look at each line in the CCs sheet. If it finds a syntax error it will gently slap your hand, and stop, asking you to attend to the problem.

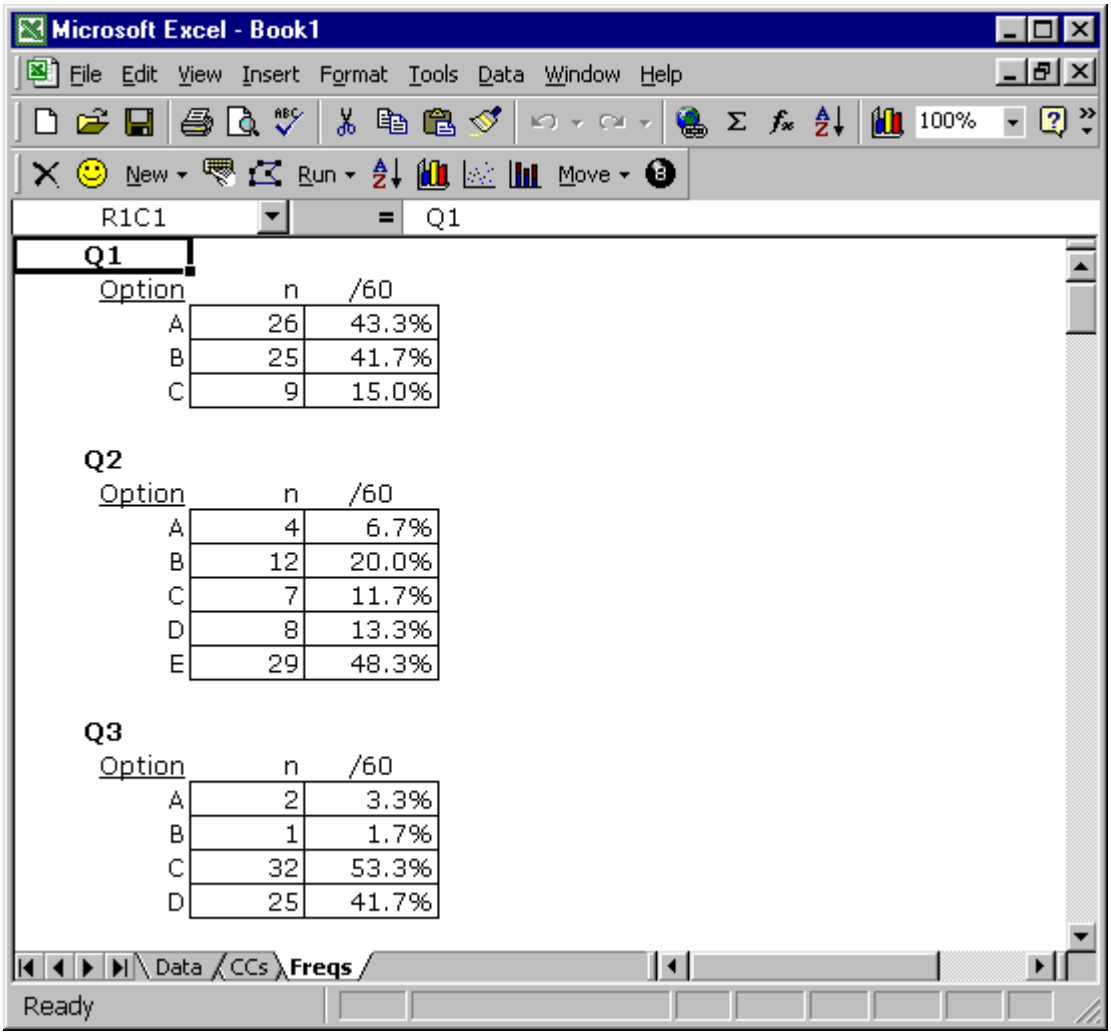
If it finds no errors, Lertap has Excel add some new worksheets to the workbook. For each of the *col lines found in the CCs sheet, Lertap creates a new "**Sub**" sheet. It also creates a "**Freqs**" worksheet.

The Sub worksheets are then hidden from view, and the Freqs worksheet comes to centre stage.

What's this "hidden from view" stuff? Any Excel workbook may have many worksheets, up to 255 in the latest versions of Excel. Worksheets which are not used very much, or which contain data used only by the workbook's macros, may be hidden by using Excel's Format options. The Sub sheets fall into the latter category—they're "system" sheets which Lertap uses to store operational data related to your workbook. You can unhide them if you want, but, if you do, you must promise not to change their contents.

The Freqs worksheet

"Freqs" means "Frequencies". When you ask Lertap to "Interpret CCs lines", it does a bit more—it has a close look at the data columns referenced in *col lines, and makes a tally of response popularities. Look:

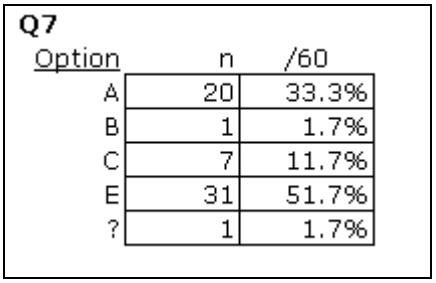


The Freqs sheet contains information which even a rocket scientist might be able to understand, hopefully with little trouble.

For each item, or question, Freqs reports on the number of times letters or digits were found in the respective item column number in the Data worksheet. For Q1 there were 26 As, 25 Bs, and 9 Cs.

The Freqs column headed "/60" indicates that a total of 60 data records were found, and gives the percentage associated with each frequency. For example, there were 26 As for Q1, which is 43.3% of the total number of 60.

As you scroll through your own Freqs sheet, you'll come across some rows which have a ? mark on the left. For example:



Q7

Option	n	/60
A	20	33.3%
B	1	1.7%
C	7	11.7%
E	31	51.7%
?	1	1.7%

What Freqs is saying is that there was one response on Q7 which was not a letter or a digit. You can find out what it was by going back to the Data sheet, and browsing down the column with Q7 responses (column 9). If you do this, you'll find a blank at R17C9, that is, Row 17, Column 9. If you were able to question the perpetrators of the Lertap Quiz, they'd tell you that blanks mean a person did not answer the question.

Why doesn't Freqs show a "D" for Q7? Because nobody chose that option.

While being mindful of Freqs' feelings, we can point out that it's a simple, no-nonsense summary of response frequencies. It has no pretences; it does not claim to be sophisticated.

But it's useful, isn't it? It quickly summarises what went on in the data set. And, very importantly, it's a sure-fire way to see if there are any weird responses in the data columns. For example, if the Q7 tally had included an X, that would be weird as only responses A though F were valid.

What if you do see weird results in Freqs, and want to quickly find the responsible records in the Data worksheet? Excel has a set of Data options offered on its Standard toolbar, and one of them is Filter. This is a powerful little option which will let you rapidly find records which have "weird" results. (If you try this, keep in mind that Excel's on-line Help is there to assist you, should you have questions on how to use the Filter option.)

Moving right along are we. We've seen Freqs, and clamour for more. What else can Lertap show us? Quite a bit.

Back to Lertap's Run options, where we now click on "Elmillon item analysis". You should do the same.

Elmillon item analysis

Why would anyone name their child Elmillon? Talk about weird. Well, it's a label which was applied to the first version of Lertap when it was under incubation at the Venezuelan Ministry of Education in 1972.

Is it ready? Is it ready? Day after day staff in the Data Bank heard that question. When it finally did hatch, the section chief, Rogelio Blanco, looked it in

the eye and said "Un millon!", which in Caracas is a popular way of saying "thanks a million". Since then the main item analysis part of Lertap has been called Elmillon. Naturally.

So. Did you go to the Run options, and click on "Elmillon item analysis"?

After you did this, did you notice Lertap flashing a variety of messages to you? Trained Lertap users have a keen eye for Excel's Status Bar, which is where the system often talks to you. The Status Bar is at the very bottom of the Excel window, where the word "Ready" is frequently seen. When you give Lertap a task to do, it will keep you advised on its progress by displaying little love notes in the Status Bar.

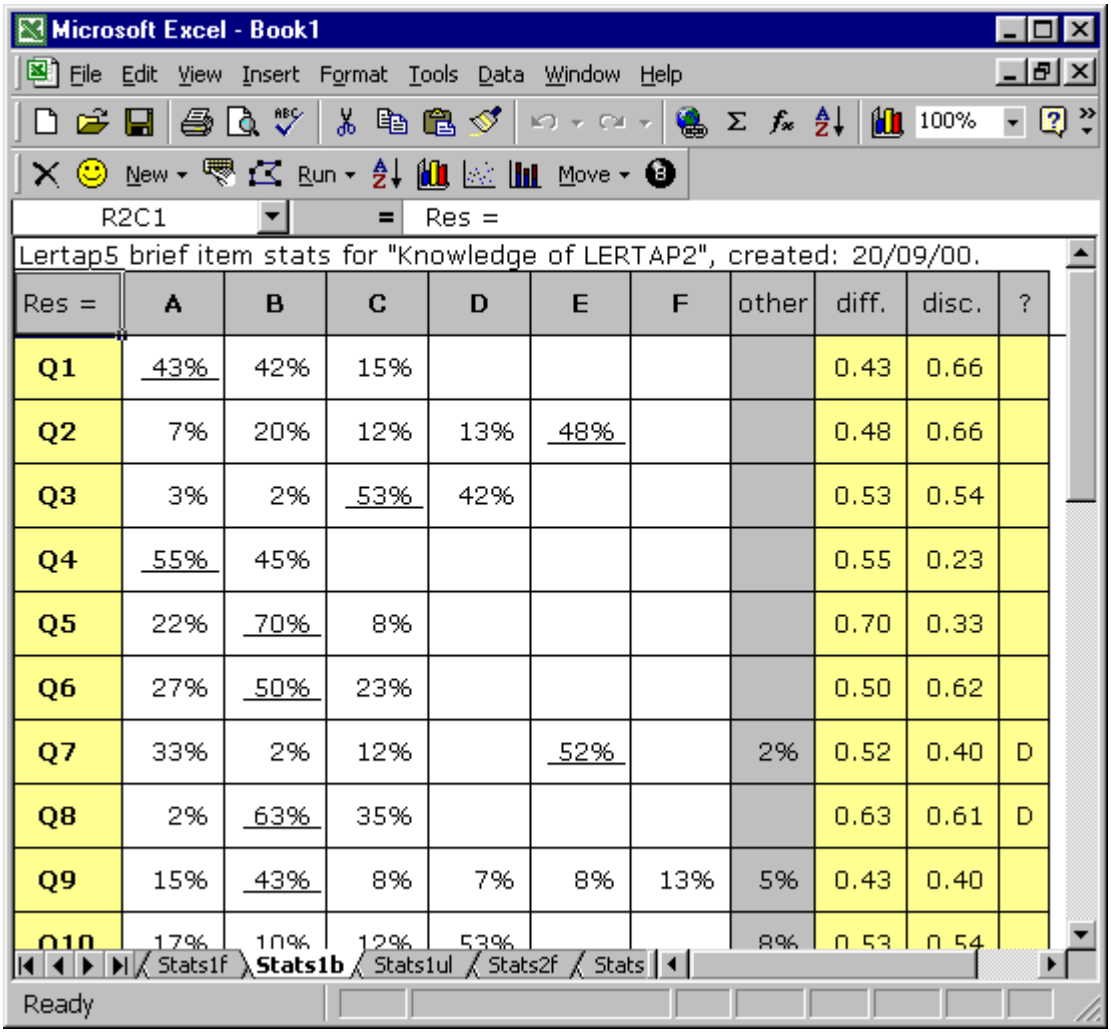
Don't have a Status Bar on your screen? Use Excel's View options, and put a tick next to Status Bar. Highly, highly recommended.

Back to business. After you clicked on "Elmillon item analysis", what happened?

Lertap snuck[10] off to have a read of all the data it stored in those secret (hidden) Sub files which were by-products of the "Interpret CCs lines" stage. Then it read through the Data file two or three times, compiling test scores, writing three summary results worksheets for the cognitive test, and two for the affective test.

These new sheets should have their tabs showing at the bottom of the Excel window, as seen here:

[10] *Snuck* is how North Americans say sneaked.



Lertap5 brief item stats for "Knowledge of LERTAP2", created: 20/09/00.

Res =	A	B	C	D	E	F	other	diff.	disc.	?
Q1	43%	42%	15%					0.43	0.66	
Q2	7%	20%	12%	13%	48%			0.48	0.66	
Q3	3%	2%	53%	42%				0.53	0.54	
Q4	55%	45%						0.55	0.23	
Q5	22%	70%	8%					0.70	0.33	
Q6	27%	50%	23%					0.50	0.62	
Q7	33%	2%	12%		52%		2%	0.52	0.40	D
Q8	2%	63%	35%					0.63	0.61	D
Q9	15%	43%	8%	7%	8%	13%	5%	0.43	0.40	
Q10	17%	10%	12%	53%			8%	0.53	0.54	

See the names Stats1f, Stats1b, Stats1ul, and Stats2f? At this point the new workbook has 9 visible (unhidden) worksheets, but only a few tabs fit into the little viewing area at the bottom of the screen. Now you can use those wee arrows at the bottom left and, as you do, different tabs will come into view. Way over to the left, for example, are the tabs for Data, CCs, and Freqs.

And now congratulations are in order: you've done it—set up a new Lertap workbook, obtained a Freqs listing, examined it closely for weirdness, and then gone on to have your new friend, Elmillon, analyse your data and "print" results to a series of new worksheets. Well done. Your Lertap career looks promising.

Next: take a break. When you return we'll have a squiz[11] at what Elmillon has done. Perhaps you should take a long break—there will be quite a bit to look at, and you'll want to be wearing your freshest eyes.

[11] *Squiz* is a word used in Australia to mean "peek".

Scores

Each of the *col lines in the CCs worksheet is said to define a "subtest". Subtests may be cognitive or affective. Cognitive tests measure knowledge or achievement, while their affective counterparts attempt to assess such things as attitudes, opinions, and feelings.

There are two *col cards in the Lertap Quiz's CCs worksheet. The first one points to a total of 25 columns in the Data worksheet, the second points to 10.

There's a *key line shortly after the first *col card, and this tells Lertap that the first subtest, with 25 items, is to be scored as a cognitive test.

The common procedure for scoring responses to cognitive questions is to award one point for each correct answer, and this is what Lertap does. It's possible to award more points, and it's possible to have more than one right answer to any cognitive item. These things are accomplished by putting *wgs and *mws lines in the CCs worksheet—but best we leave these "advanced" matters for a later chapter. And so we shall.

The first *sub line gives a title of "Knwldge" to the first subtest. What will be the possible range of scores for Knwldge? From zero to 25. As we've just said, this is a "common", or standard, cognitive run, there are no *wgs or *mws lines in the CCs worksheet. Consequently respondents get zero points for each incorrect answer, and one point for each right response. There are 25 items, so the maximum possible score for Knwldge is 25.

The second subtest, "Comfort", is affective in nature, something Lertap detects by the presence of the "Aff" control "word" on the *sub card. There are 10 items in this subtest. Lertap will score each, adding results from each item to derive a total Comfort score.

How? Good question. The answer may be obvious to users of earlier Lertaps, but probably not to other readers.

Lertap assumes that each affective item will use 5 possible responses. In fact, it assumes that Res=(1,2,3,4,5). What's this Res=() thing mean? Two things. The number of characters found within the parentheses tells Lertap how many possible responses there may be to any item, while the characters themselves, five digits in this case, are the responses which will be recognised and scored.

How? If someone has an answer of 1 to the first affective item, how many points are awarded? One. Two points for the second response, which is a 2 in this case. Three for the third, four for the fourth, five for the fifth.

This is a common way of scoring affective items. Lertap allows for many other possibilities, scoring-wise. There can be more than five recognised responses—there may be as many as 10. The responses do not have to be digits. The points given to the responses do not have to correspond to the ordinal position of the response—special lines in the CCs worksheet, such as *mws lines, allow other scoring schemes to be effected.

Okay then, the subtest we're talking about, Comfort, has 10 items. The minimum possible score on any single item is 1 (one), while the maximum is 5. For the 10 items as a whole, then, the minimum possible score is 10, the maximum 50.

We're just about ready to look at the scores themselves, but first one final matter. The Comfort subtest has a *pol line associated with it—"pol" means "polarity". The *pol line has ten plus and minus signs, one for each of the subtest's items. The first sign in the *pol card is +, which means that a response of 1 (one) on the first item will get one point, while a response of 5 will get five points. This is referred to as "forward" weighting.

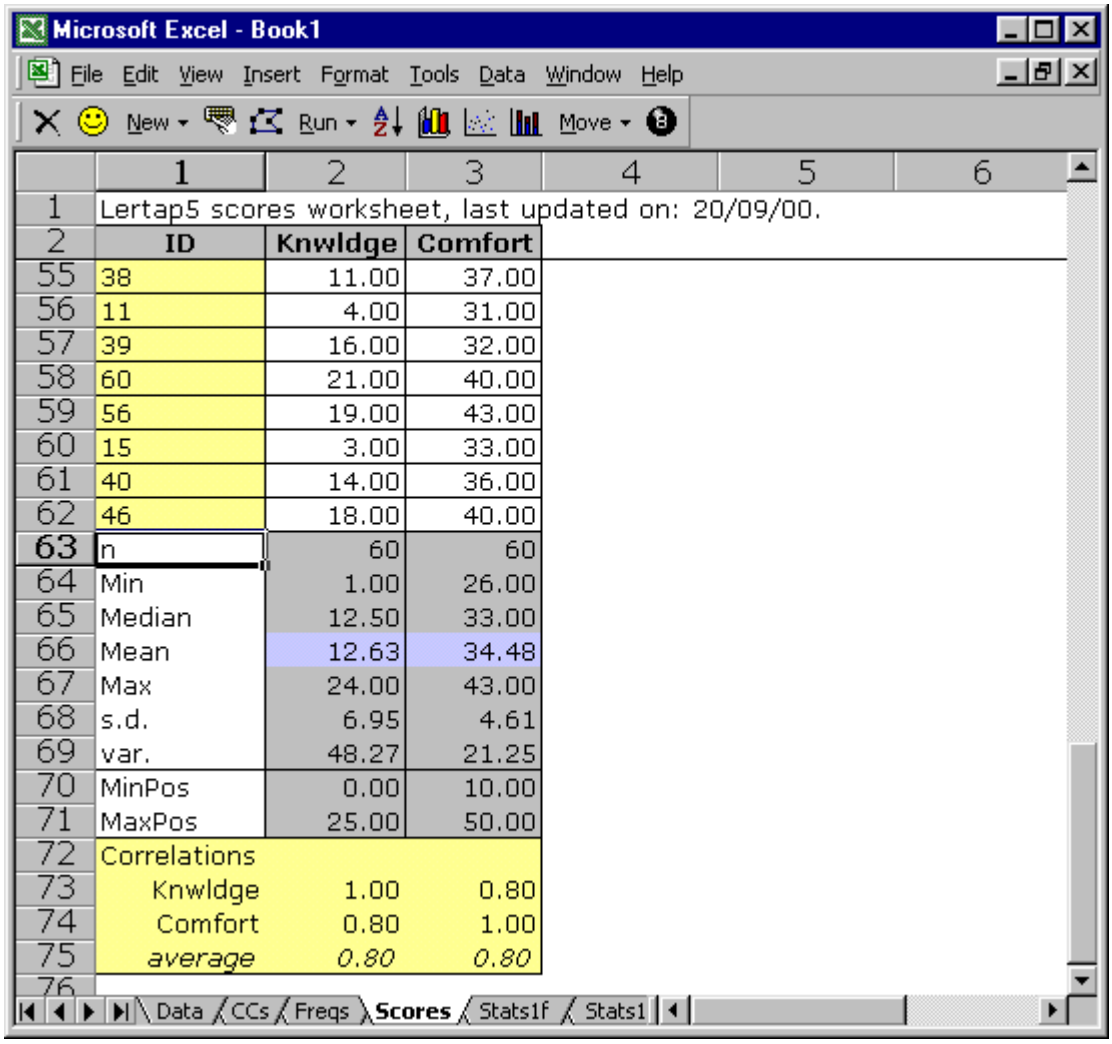
Items whose respective entry in the *pol card is minus will be reverse scored. On these items, a response of 1 (one) will get five points, while a response of 5 will get one point.

Why get into this sort of caper? Because it is not at all unusual for affective tests, or surveys, to consist of a mixture of positive and negative items. People might be asked to state whether they agree or disagree to a series of statements. The first one might be, for example, "Lertap is great!", while the second might be "I would not advise anyone to use Lertap." People who are happy with Lertap, which of course is just about everyone, would be expected to agree with the first statement, and disagree with the second. The use of *pol lines makes it possible to accommodate forward and reverse scoring with some ease.

Let's see them then. Let's have a squiz at these scores, Knwldge and Comfort.

Where are they?

They're in the Scores worksheet. Find its tab at the base of the Excel window. Click on the tab. Look:



	1	2	3	4	5	6
1	Lertap5 scores worksheet, last updated on: 20/09/00.					
2	**ID**	**Knwldge**	**Comfort**			
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	n	60	60			
64	Min	1.00	26.00			
65	Median	12.50	33.00			
66	Mean	12.63	34.48			
67	Max	24.00	43.00			
68	s.d.	6.95	4.61			
69	var.	48.27	21.25			
70	MinPos	0.00	10.00			
71	MaxPos	25.00	50.00			
72	Correlations					
73	Knwldge	1.00	0.80			
74	Comfort	0.80	1.00			
75	*average*	*0.80*	*0.80*			

What do you make of it, this Scores sheet? It uses 75 rows, and 3 columns. Lertap presupposes that the first thing you want to rest your peepers on, as far as the Scores sheet goes, is the summary statistics section at the bottom of the sheet. This is why rows 3 through 54 have scrolled off the display. If you scroll to the top of the sheet, you'll be able to satisfy yourself that there are 60 sets of scores, one pair, Knwldge and Comfort, for each respondent.

Are the summary statistics self-explanatory? Good. You can find out how they're calculated in Chapter 10, Computational Methods.

A couple of quick comments before moving on. The MinPos and MaxPos scores are the minimum and maximum possible scores on each subtest, while Min and Max are the lowest and highest scores actually earned by the 60 respondents. The correlation coefficients are Pearson product-moment coefficients (the most common kind).

Is it possible to get percentage scores? You bet. Use the "Per" control word on the *sub card. It's also possible to get a "scale" score if the subtest is affective.

Such scores divide the original subtest score by the number of items in the subtest, a procedure which is common to quite a number of internationally-known affective instruments[12]. Scale scores are requested by using the "Scale" control word on a *sub line. More about these fancy control words, and others, in a later chapter.

Is it possible to get a total score, one which sums up the scores earned on the individual subtests? Most definitely. In fact, Lertap has to be told not to do this, something which is done by using Wt=0 statements on *sub cards. In the present example, each *sub line has this sort of statement. We didn't want a total score—we didn't think it made much sense to add together results from two very different subtests, one cognitive, one affective.

Doing more with Scores

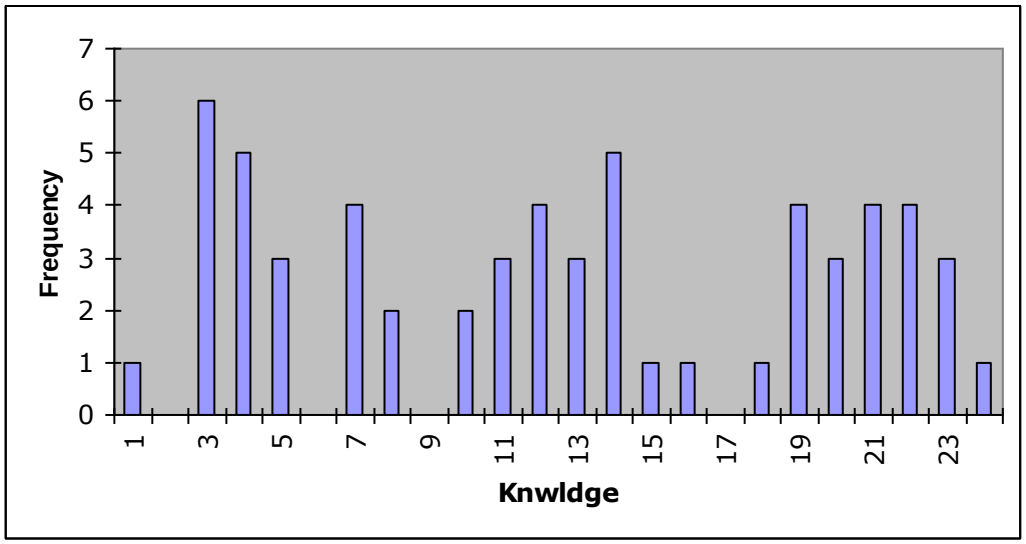
Once a Scores worksheet has been created, there are a few icons on Lertap's toolbar which let you do more with them:

The first of these three icons lets you sort the Scores worksheet according to criteria which you provide. Lertap calls on Excel to do sorts, and Excel has a *very* good sorter (nothing sortid about it, if you will). You can sort on IDs so that the scores are in alphabetical order; you can sort on any particular score so that they're displayed in ascending or descending order.

The Scores worksheet, like all Excel worksheets, may be printed with ease.

Histograms

Here's a pretty picture of the Knwldge scores:

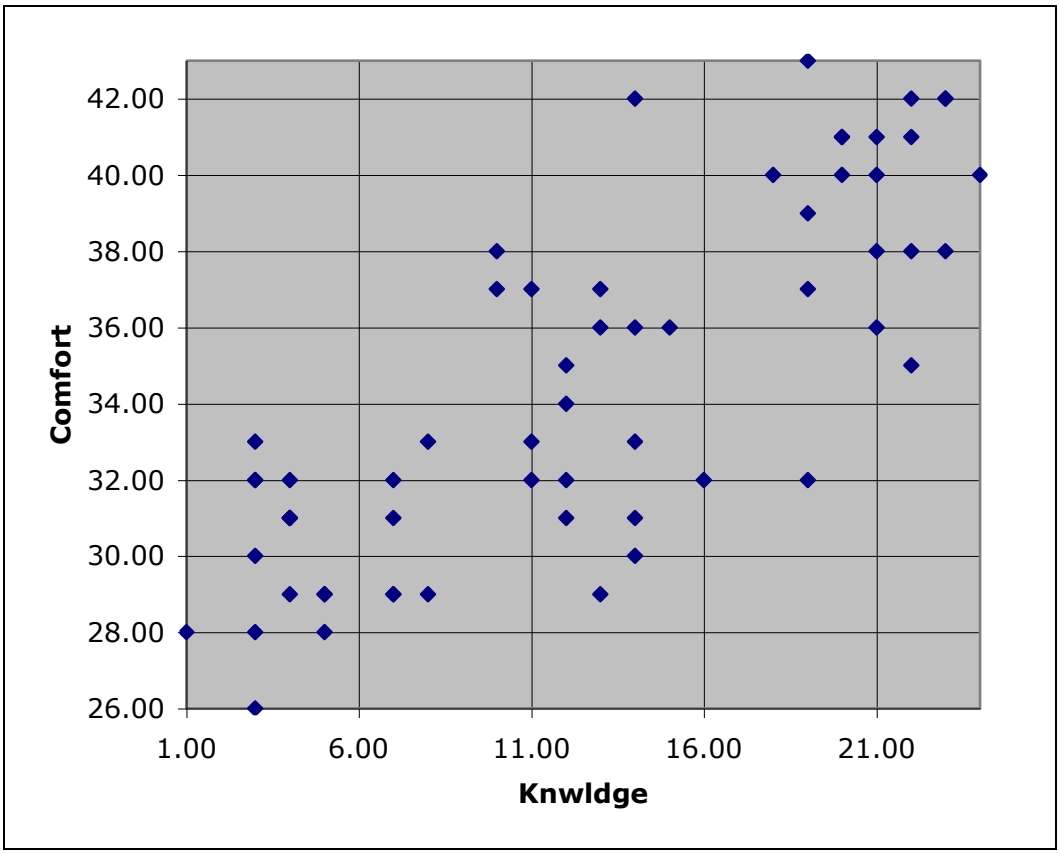


[12] For example, the *Computer Attitudes Questionnaire* from the University of North Texas, and the University of Michigan's MSLQ, *Motivated Strategies for Learning Questionnaire*.

How to? How to make beauties such as this? Use the second of the three icons mentioned above, and please note: the Excel system in use on your computer has to have what's called the "Analysis ToolPak add-in" installed before the histogrammer will give you pictures such as this one. If you don't have it installed, it's easy to do—look in Excel's on-line Help, or refer to Chapter 10, Computational Methods, for more comments.

Scatterplots

You asked for it, you got it. Click on the third icon mentioned above, answer a couple of easy questions, and lo:



Lertap asks Excel to use one of its standard charts for the scatterplot, and, as a consequence, the Analysis ToolPak add-in is not required.

You're doing well, you really are. Ready for more? Let's leave scoring matters, and get into response, item, and subtest statistics.

Statistics for cognitive subtests

Three worksheets of statistics are usually produced for each cognitive subtest. These are referred to as the "full", "brief", and "upper-lower" sheets. The information provided by these sheets often overlaps to a considerable extent, as you're about to see.

In the case of our running example, you should notice that Elmillon added sheets called "Stats1f", "Stats1b", and "Stats1ul". These three sheets correspond to full, brief, and upper-lower statistics for the first subtest. You'll notice that there are two others sheets of this ilk, named "Stats2f" and "Stats2b". These relate to the second subtest, the affective one.

Brief statistics for cognitive subtests

Have a look at the Stats1b worksheet.

Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 brief item stats for "Knowledge of LERTAP2", created: 20/09/00.

Res =	A	B	C	D	E	F	other	diff.	disc.	?
Q1	43%	42%	15%					0.43	0.66	
Q2	7%	20%	12%	13%	48%			0.48	0.66	
Q3	3%	2%	53%	42%				0.53	0.54	
Q4	55%	45%						0.55	0.23	
Q5	22%	70%	8%					0.70	0.33	
Q6	27%	50%	23%					0.50	0.62	
Q7	33%	2%	12%		52%		2%	0.52	0.40	D
Q8	2%	63%	35%					0.63	0.61	D
Q9	15%	43%	8%	7%	8%	13%	5%	0.43	0.40	

ScatP1 ScatP2 Scores Stats1f **Stats1b**

This sheet has its row and column headers hidden, but the contents of the first row are obvious: a header, a title. Note that it includes the name of the subtest, as given in the corresponding *sub line in the CCs worksheet.

The second row in Stats1b displays the responses used by this subtest, from A through F. The "other" column is used to indicate how many weird responses were found for each item. Going to the right, "diff." means item difficulty, and "disc." means item discrimination. The "?" column is used to indicate which, if any, of the item's distractors might be regarded as rendering questionable service.

You'll note that each item's summary statistics are presented in a single row. The percentage figure which is underlined corresponds to the item's correct answer, as taken from the corresponding *key line in the CCs sheet.

The derivation and interpretation of the diff., disc., and ? columns is discussed below, and in subsequent chapters. The information in the brief stats sheet is taken from lines found in another of the three statistical summaries, the "full stats" sheet, Stats1f.

Full statistics for cognitive subtests

The Stats1f worksheet contains a wealth of information, presented in several sections.

The first section gives detailed statistics for each item, as seen here:

Q1

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	1.00	26	0.43	0.66	0.83	18.15	0.79
B	0.00	25	0.42	-0.57	-0.72	7.92	-0.68
C	0.00	9	0.15	-0.17	-0.26	9.78	-0.41

The "wt." column indicates the number of points associated with each possible response option. "p" is "n" as a proportion, and corresponds to the percentage figures seen in the corresponding brief statistics worksheet. The "pb(r)" column indicates the point-biserial correlation of each response with the criterion score, while "b(r)" is the biserial equivalent. If an item has only one correct answer, the pb(r) figure corresponding to it is what is carried over to the corresponding brief statistics sheet, Stats1b, where it is displayed under the "disc." column.

The "avg." column displays the average criterion score for the people who selected each response option. For Q1, 26 people selected option A. Their average criterion score was 18.15. The "z" column converts the "avg." figure to a z-score, using mean and standard deviation values for the criterion score.

Lertap's default criterion score is an internal one, equal to the subtest score. It is possible to set up an external criterion analysis via one of the toolbar's Run options.

Lertap has a studied look at the performance of each item's distractors, that is, their wrong answers. If these options are doing their job, they should, first of all, truly distract people—be selected by someone. If no-one falls for a distractor, Lertap indicates this by listing the distractor under the ? column of the brief statistics worksheet.

The people who are distracted by the distractors should, in theory, be those whose mastery of the test material is below average. Below average is readily signalled by negative z-scores.

An unwanted outcome for a distractor is a positive z-score, which means that the people who took the "distractor" had above-average criterion scores. When this

happens we usually think that the item has perhaps been mis-keyed (that is, the *key line of correct answers in the CCs worksheet may be in error). If it's not mis-keyed, we then tend to think that the option has some intrinsic or extrinsic ambiguity, and requires repair. Distractors such as these, with positive z-scores, are also listed under the ? column of the corresponding brief statistics sheet, Stats1b.

The second section in the Stats1f sheet is the Summary Statistics part:

Summary statistics

number of scores (n):	60	
lowest score found:	1.00	(4.0%)
highest score found:	24.00	(96.0%)
median:	12.50	(50.0%)
mean (or average):	12.63	(50.5%)
standard deviation:	6.95	(27.8%)
standard deviation (as a sample):	7.01	(28.0%)
variance (sample):	49.08	

number of subtest items:	25	
minimum possible score:	0.00	
maximum possible score:	25.00	
reliability (coefficient alpha):	0.91	
index of reliability:	0.96	
standard error of measurement:	2.03	(8.1%)

Much of the information found in this section is also found at the bottom of the Scores worksheet. However, the subtest reliability information is only found here, in this section of the Stats1f sheet.

The Summary Statistics section is followed by two subsections with "bands":

item difficulty bands

.00: Q22
.10:
.20:
.30:
.40: Q1 Q2 Q9 Q11 Q14 Q18 Q19 Q20 Q21 Q25
.50: Q3 Q4 Q6 Q7 Q10 Q12 Q15 Q17 Q24
.60: Q8 Q13 Q16 Q23
.70: Q5
.80:
.90:

item discrimination bands

.00:
.10:
.20: Q4 Q22
.30: Q5 Q14 Q24
.40: Q7 Q9 Q16 Q23
.50: Q3 Q10 Q12 Q15 Q17
.60: Q1 Q2 Q6 Q8 Q11 Q18 Q21 Q25
.70: Q13 Q19 Q20
.80:
.90:

These bands summarise difficulty and discrimination data for the subtest's items; they're meant to make it possible to quickly see which items have performed the best, and which may require further study.

In the present example, Q22 falls into the lowest difficulty band, .00, meaning that it was a very hard item in this group of test takers. Q22 joins Q4 in having the lowest discrimination figure.

These bands are based on the complete item statistics results shown in the first part of the Stats1f output. To find the exact difficulty and discrimination values for any item, scroll up in the Stats1f sheet, or look in the Stats1b sheet. Remember that Excel, like Word, allows for its viewing window to be split, something which makes it easier to see different parts of the same sheet at once.

The bands are followed by the last subsection:

alpha figures (alpha = .9149)

without	*alpha*	*change*
Q1	0.909	-0.006
Q2	0.909	-0.006
Q3	0.911	-0.003
Q4	0.917	0.002
Q5	0.915	0.000
	...	
Q20	0.908	-0.007
Q21	0.910	-0.005
Q22	0.916	0.001
Q23	0.914	-0.001
Q24	0.915	0.000
Q25	0.910	-0.005

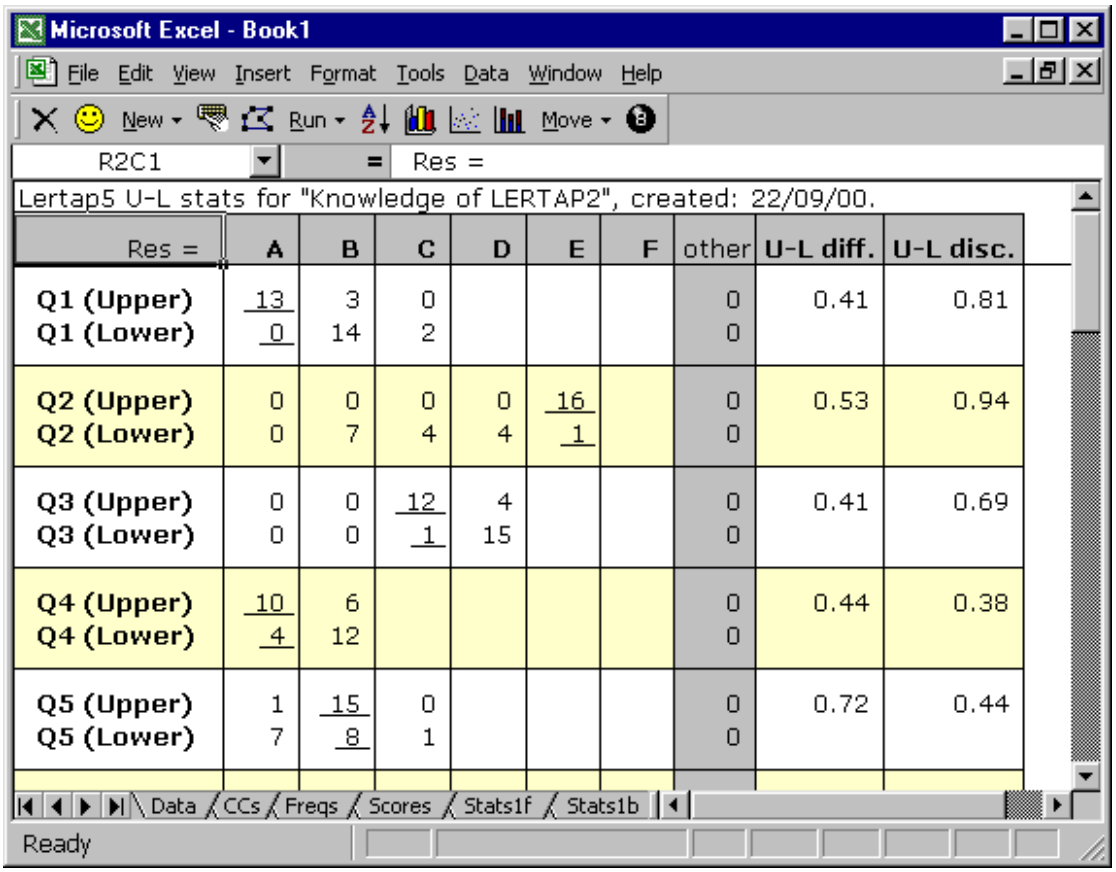
This section (above) makes it possible to see how subtest reliability would be affected if an item were deleted from the subtest. Without Q4, for example, the subtest's reliability index, alpha, would be 0.917, an increase (improvement) of 0.002. (The table above has been shortened in order to save space in this document.)

Upper-lower statistics for cognitive subtests

We've considered the brief and full statistics sheets for the first subtest, respectively (and respectfully) known as Stats1b and Stats1f. For years Lertap versions provided only the full statistics; the brief statistics sheet has been added in Lertap 5 to make it possible to get a quick idea of how items performed.

These two sheets, the full and brief ones, use point-biserial correlation coefficients to index item discrimination. There's another way of indicating how well an item discriminates between the strong and the weak—divide test results into two groups, one with the strongest performers, one with the weakest—and then look at item-level results in each group.

This is referred to as the "Upper-Lower" method. Lertap 5's Stats1ul sheet contains the U-L statistics corresponding to the first subtest:



Res =	A	B	C	D	E	F	other	U-L diff.	U-L disc.
Q1 (Upper)	13	3	0				0	0.41	0.81
Q1 (Lower)	0	14	2				0		
Q2 (Upper)	0	0	0	0	16		0	0.53	0.94
Q2 (Lower)	0	7	4	4	1		0		
Q3 (Upper)	0	0	12	4			0	0.41	0.69
Q3 (Lower)	0	0	1	15			0		
Q4 (Upper)	10	6					0	0.44	0.38
Q4 (Lower)	4	12					0		
Q5 (Upper)	1	15	0				0	0.72	0.44
Q5 (Lower)	7	8	1				0		

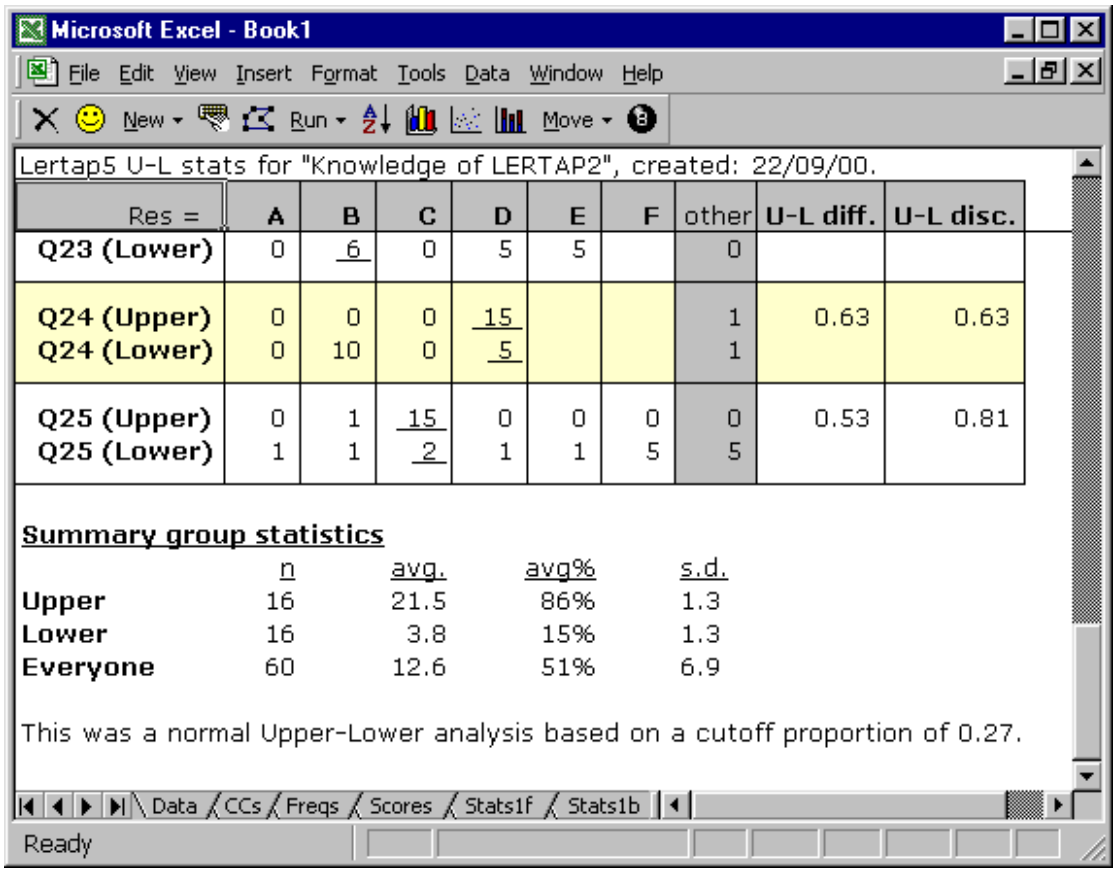
Two rows of results are given for each test item. One row contains response frequencies for the top, or "upper", group; the other has the frequencies for the bottom, or "lower", group.

The two right-most columns of the Stats1ul display summarise item difficulty and discrimination, with the difficulty index formed by dividing the number in both groups who got the item correct (13 in the case of Q1), by the total number of people in the groups (32).

One of the main attractions of the U-L approach lies in its discrimination index. It's simply the proportion of people getting the item right in the upper group, less the same proportion for the lower group. For Q1, this is .81 - .00, or .81.

If the U-L disc. index is 1.00, the maximum possible, it means that everyone in the upper group got the item right, while everyone in the bottom group got it wrong. A result of –1.00, the minimum possible, happens when everyone in the top group got the item wrong, while everyone at the bottom got it right, an unusual and unwanted outcome.

Lertap forms the groups by taking the top 27% and bottom 27% of the test scores, a traditional way of defining the upper and lower groups (see, for example, Hopkins, Stanley, & Hopkins 1990, p.269). It summarises what it's done at the very end of the Stats1ul worksheet, as shown below:



Is it possible to define the groups with different proportions? Yes. The Computation Methods chapter, Chapter 10, explains how.

Mastery test analysis

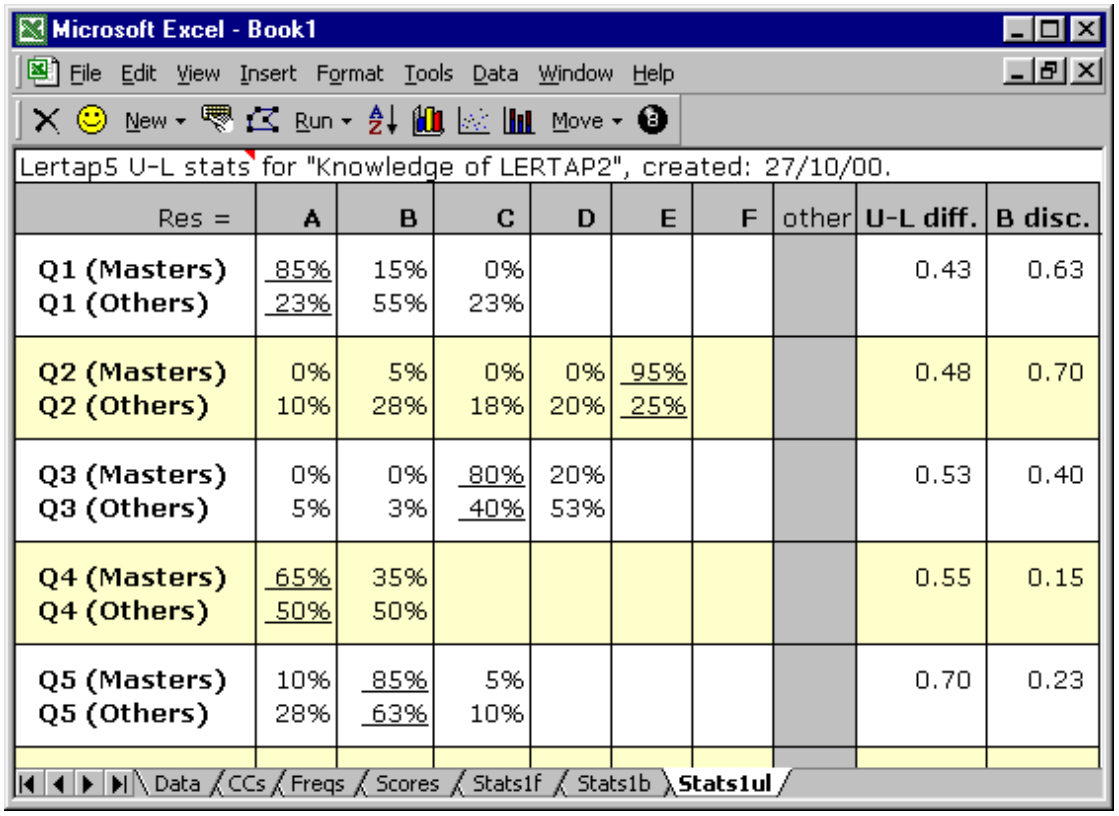
Look what happens if we change the *sub line for the first subtest so that it includes the word "Mastery". Thus far the *sub card has been

*sub Res=(A,B,C,D,E,F), Name=(Knowledge of LERTAP 2), Title=(Knwldge), Wt=0

if we change it to

*sub Mastery, Res=(A,B,C,D,E,F), Name=(Knowledge of LERTAP 2), Title=(Knwldge), Wt=0

and then "Interpret CCs lines", followed by "Elmillon item analysis", the Stats1ul sheet will look like this:



Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

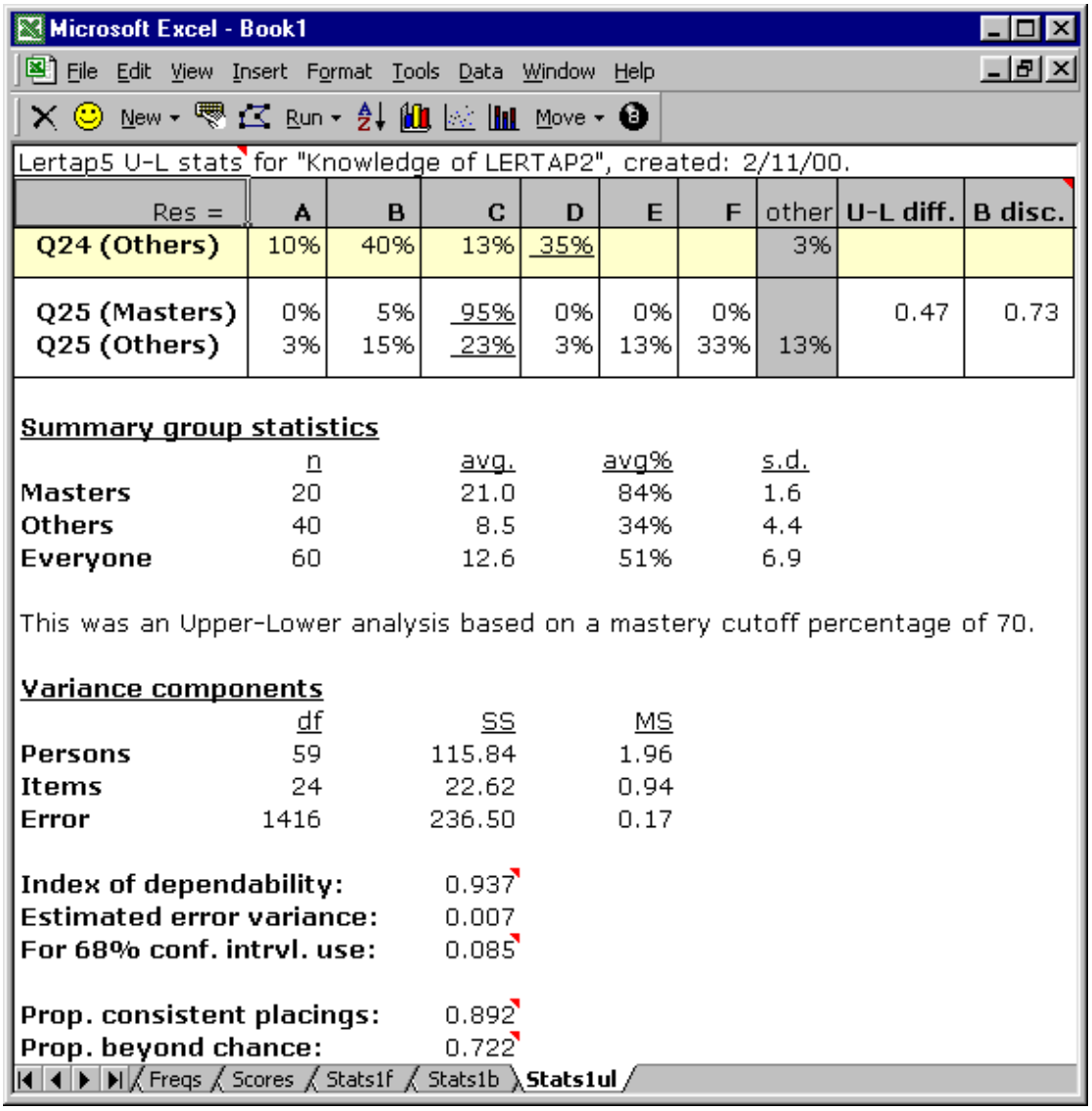
New Run Move

Lertap5 U-L stats for "Knowledge of LERTAP2", created: 27/10/00.

Res =	A	B	C	D	E	F	other	U-L diff.	B disc.
Q1 (Masters)	85%	15%	0%					0.43	0.63
Q1 (Others)	23%	55%	23%						
Q2 (Masters)	0%	5%	0%	0%	95%			0.48	0.70
Q2 (Others)	10%	28%	18%	20%	25%				
Q3 (Masters)	0%	0%	80%	20%				0.53	0.40
Q3 (Others)	5%	3%	40%	53%					
Q4 (Masters)	65%	35%						0.55	0.15
Q4 (Others)	50%	50%							
Q5 (Masters)	10%	85%	5%					0.70	0.23
Q5 (Others)	28%	63%	10%						

Data CCs Freqs Scores Stats1f Stats1b **Stats1ul**

The item-level results are then followed by a variety of statistics which summarise group results, and provide indicators of the reliability of the process:



Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "Knowledge of LERTAP2", created: 2/11/00.

Res =	A	B	C	D	E	F	other	U-L diff.	B disc.
Q24 (Others)	10%	40%	13%	35%			3%		
Q25 (Masters)	0%	5%	95%	0%	0%	0%		0.47	0.73
Q25 (Others)	3%	15%	23%	3%	13%	33%	13%		

Summary group statistics

	n	avg.	avg%	s.d.
Masters	20	21.0	84%	1.6
Others	40	8.5	34%	4.4
Everyone	60	12.6	51%	6.9

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

Variance components

	df	SS	MS
Persons	59	115.84	1.96
Items	24	22.62	0.94
Error	1416	236.50	0.17

Index of dependability:	0.937
Estimated error variance:	0.007
For 68% conf. intrvl. use:	0.085
Prop. consistent placings:	0.892
Prop. beyond chance:	0.722

Freqs Scores Stats1f Stats1b **Stats1ul**

This is a "Mastery" analysis. The top group, the "Masters", has all those who reached "mastery", which, above, was set at 70%. The lower group, the "Others", has everyone else.

The U-L disc. index is now called the "B disc." index after Brennan (1972). The statistics presented in the lower half of this screen snapshot, from "Variance components" down, come from the work of Brennan and Kane (1977), and Subkoviak (1984). Mention of their interpretation is made later, in Chapter 7.

Is it possible to set the mastery level at something other than 70? Yes. The following *sub line sets it at 80:

*sub Mastery=80, Res=(A,B,C,D,E,F), Name=(Know. of LERTAP2), Title=(Knwldge), Wt=0

Is it possible to define the upper and lower groups according to performance on an external criterion? Yes. The Run option on Lertap's toolbar allows for this.

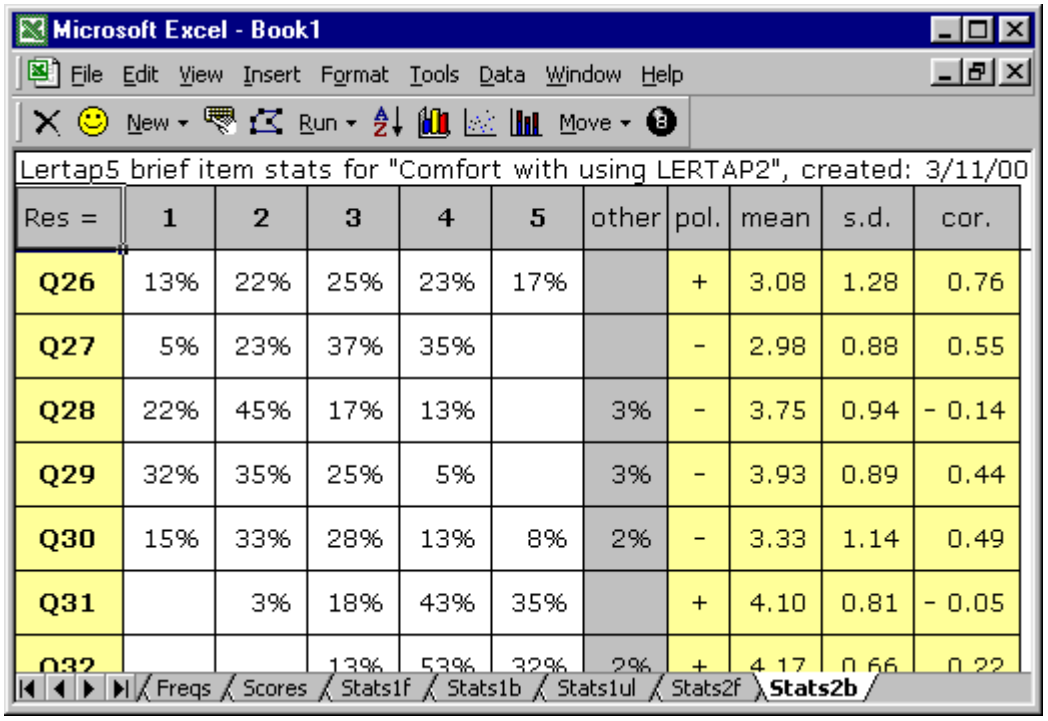
Statistics for affective subtests

Lertap 5 produces two reports sheets for affective subtests, one with "full" statistics, and one with "brief" summaries. These worksheets will have names similar to those for cognitive subtests, such as "Stats1f" and "Stats1b", where the "f" refers to "full", and the "b" to (you guessed it:) "brief".

In the examples below the affective worksheets have names of Stats2f and Stats2b. This is because the CCs worksheet for the example we've been following has two subtests—the first one, a cognitive subtest titled "Knwldge", has been looked at above; because it was the first one in, its sheets have the number 1 (one) in their names, such as Stats1f, Stats1b, and Stats1ul. The affective subtest came next, came second, and as a result its sheets have the number 2 in their names.

Brief statistics for affective subtests

Have a squiz—here are the brief results for the "Comfort" affective subtest (or "scale"):



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

The Stats2b worksheet shown above has its row and column headers hidden at Lertap's request—you can turn them back on, if you want, by using Excel's Tools / Options.

The first row of the Stats2b report includes the subtest's name, "Comfort with using LERTAP2". Where did this name come from? From the subtest's *sub card, or line, in the CCs worksheet.

The response codes used by the subtest's items are shown to the right of the Res = cell, in the second row. This affective subtest used the default response code set of Res=(1,2,3,4,5). If you now were to refer back to the listing of CCs lines, you might wonder why the first subtest, Knwldge, has an explicit Res=() declaration on its *sub card, whereas Comfort does not. The answer has to do with the fact that the Comfort scale used default response codes—Knwldge did not[13].

Looking at the results for Q26, 13% of the respondents opted for the first response, 22% for the second, and so forth. All respondents answered Q26, a fact which is revealed by the absence of an entry under the "other" column. When there's no response to an item, or an invalid response, Lertap lets you know by putting something in the "other" column.

Q26's "pol." is "+". What's this mean? It means that positive scoring applies to Q26, or, in other words, that the scoring for this item has not been reversed. If you're itching to know more about scoring, hang on—it's coming—there's more about it in the next section.

The average of the Q26 responses, and their standard deviation, are found under the "mean" and "s.d." columns, respectively. The "cor." column gives the value of the Pearson product-moment correlation coefficient between the item and the criterion score. At this point the criterion score is the sum of each person's responses to the other items on the subtest, that is, the subtest score less Q26. Why exclude Q26 from the subtest score? So that the effects of part-whole inflation are eliminated—the correlation between an item and the subtest score will be inflated if the item is part of the subtest score, and, to control for this, Lertap applies a correction.

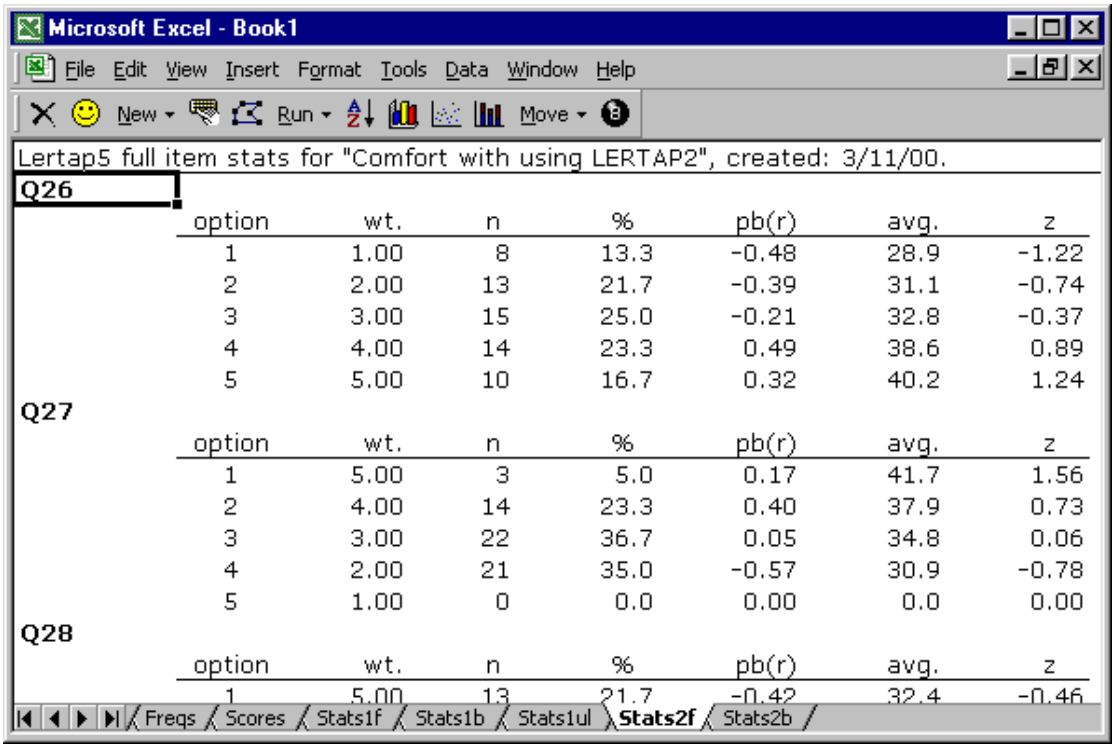
More information about this subtest's items is to be found in the "full" statistics report, Stats2f in this case.

Full statistics for affective subtests

Lertap's "full" report for affective tests has two main areas, starting with item-level results, followed by various subtest summary sections.

The item results look like this:

[13] The default for cognitive subtests is Res=(A,B,C,D).



Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 full item stats for "Comfort with using LERTAP2", created: 3/11/00.

Q26

option	wt.	n	%	pb(r)	avg.	z
1	1.00	8	13.3	-0.48	28.9	-1.22
2	2.00	13	21.7	-0.39	31.1	-0.74
3	3.00	15	25.0	-0.21	32.8	-0.37
4	4.00	14	23.3	0.49	38.6	0.89
5	5.00	10	16.7	0.32	40.2	1.24

Q27

option	wt.	n	%	pb(r)	avg.	z
1	5.00	3	5.0	0.17	41.7	1.56
2	4.00	14	23.3	0.40	37.9	0.73
3	3.00	22	36.7	0.05	34.8	0.06
4	2.00	21	35.0	-0.57	30.9	-0.78
5	1.00	0	0.0	0.00	0.0	0.00

Q28

option	wt.	n	%	pb(r)	avg.	z
1	5.00	13	21.7	-0.42	32.4	-0.46

Freqs Scores Stats1f Stats1b Stats1ul **Stats2f** Stats2b

The full statistics for affective items are quite similar to those provided for cognitive items.

Users can check Lertap's item scoring by looking down the "wt." column. Here, "wt." means weight. Above Q26's weights exhibit "forward" scoring, while Q27's are reversed—do you see why?

On Q26, a response of 1 (the first option) has a corresponding weight of 1.00, and 5 has a weight of 5.00. These "weights" are what people get for their answers. They're item scores. Someone who answers 1 on Q26 will get 1.00 points. However, the scoring has been reversed for Q27. An answer of 1 on Q27 equates to 5.00 points. This forward (+) and reverse (-) scoring is defined on the subtest's *pol card in the CCs worksheet[14].

The pb(r) column gives the point-biserial correlation of each option with the criterion score, that is, the subtest score. At this level, the pb(r) figure is <u>not</u> corrected for part-whole inflation—it is for cognitive items, but not for affective ones, and this subtest, Comfort, is affective.

The "avg." column indicates the average criterion score for the people who selected each option. On Q26, eight (8) people selected the first option, and their average criterion score was 28.9, or, as a z-score, -1.22.

The criterion score in use at the moment is said to be an "internal" criterion: it's the subtest score itself. Thus the z-score for Option 1, Q26, is computed by

[14] If all items are forward scored, the *pol card is <u>not</u> required.

subtracting the subtest mean, 34.5, from 28.9, and dividing the result by the subtest's standard deviation, 4.6.

Weights for missing affective responses

There were no "other" responses for the two items shown above, Q26 and Q27. Every one of the 60 respondents answered these two items by selecting one of the five options. However, on Q28 two people declined to respond—their "answers" were entered as a "blank" (a space) in the Data worksheet. The statistics corresponding to these two are seen below in the "other" row:

option	wt.	n	%	pb(r)	avg.	z
1	5.00	13	21.7	-0.42	32.4	-0.46
2	4.00	27	45.0	0.46	36.8	0.51
3	3.00	10	16.7	-0.16	32.8	-0.37
4	2.00	8	13.3	-0.08	33.5	-0.21
5	1.00	0	0.0	0.00	0.0	0.00
other	3.00	2	3.3	-0.22	29.0	-1.19

It is important to note that Lertap gives a score to "others". In this example, the score for "others" is 3.00 points, which is the middle of the weights for the item.

The reason Lertap assigns a weight, or score, for "others" is to try and keep the response summaries "honest". If "others" got no weight, as is the case for cognitive items, then the item mean would be lowered, and, if users just looked at results in the brief stats sheet, Stats2b, a false impression of item responses might occur—one would have the impression that responses were lower than they actually were.

Assigning a scoring weight to "others" is done automatically, and is referred to in Lertap as the MDO, the missing-data option. This automatic assignment may be turned off by using the MDO control word on the *sub card, as exemplified here:

*sub Aff, MDO, Name=(Comfort with using LERTAP2), Title=(Comfort), Wt=0

Users of previous versions of Lertap will notice that the way MDO works in this version is opposite to what was before. Now MDO is <u>always</u> assumed to be "on", and the MDO control word on the *sub card extinguishes it.

The item-level information in the Stats2f sheet is followed by a series of summaries, as shown here:

Summary statistics

number of scores (n):	60	
lowest score found:	26.00	(52.0%)
highest score found:	43.00	(86.0%)
median:	33.00	(66.0%)
mean (or average):	34.48	(69.0%)
standard deviation:	4.61	(9.2%)
standard deviation (as a sample):	4.65	(9.3%)
variance (sample):	21.61	

number of subtest items:	10	
minimum possible score:	10.00	
maximum possible score:	50.00	
reliability (coefficient alpha):	0.63	
index of reliability:	0.79	
standard error of measurement:	2.81	(5.6%)

mean/max bands

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

correlation bands

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

alpha figures (alpha = .6285)

without	*alpha*	*change*
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

In the Summary Statistics section, the % figures convert the various summary indices to their percent counterparts, where the percentage values are based on the maximum possible score. Thus, for example, the lowest score found in this group of 60 was 26.00, which is 52% of the maximum possible score, 50.00.

The minimum and maximum possible scores are determined by going over the items, one by one, summing the lowest weights to get the minimum possible total score, and summing the highest weights to get the maximum possible.

The reliability figure reported by Lertap is coefficient alpha, an internal consistency figure sometimes called "Cronbach's alpha".

The mean/max bands are based on dividing each item's mean by the highest item weight. For example, the Stats2b sheet indicates that Q26's mean was 3.08. The highest weight for Q26 was 5.00, so the item's mean/max figure was 0.62—this is why Q26 may be seen lying in the 0.60 mean/max band above. These bands make it possible to quickly see where responses were most polarised. In this example, respondents were particularly in agreement on items Q31 and Q32. Of course, a scan down the means column in the Stats2b report will indicate the same (these items have the highest means), but when there are many items the mean/max bands capture the results more efficiently.

The correlation bands simply map the results of the Stats2b "cor." column, making it possible to rapidly identify those items with the greatest correlation with the criterion score.

The alpha figures indicate how subtest reliability would be affected should an item be removed from the subtest. For example, without Q26 alpha would decrease by –0.175. Note that we'd get quite a nice increase in reliability if Q34 were omitted from the subtest. However, whether or not we'd actually want to delete an item from the subtest does not usually depend on reliability alone, a matter further discussed in Chapter 8.

Those research questions

Back on page 22 we posed a few "research questions" which we proposed to set about answering. And so we have, or almost so. We've looked at both the cognitive and affective subtests, finding out which of the cognitive questions were

the most difficult. To determine how the subtest scores looked, we activated the histogrammer.

Did the subtests have adequate reliability? Well, the cognitive subtest, "Knwldge", came through with an alpha reliability of 0.91, standard error of measurement of 8.1% (or 2.03 in raw score terms), which is not bad. The affective subtest, "Comfort", fared less well, having an alpha figure of 0.63, which is weak. We'd have to have more of a look at that subtest, as it is right now we would not feel confident, not too comfortable (as it were), in using the scores from this subtest as reliable indicators—for the moment, we might compromise by proposing to look at Comfort results only at the level of the individual items.

How did respondent attitudes correspond to how well they did on the cognitive test? The correlation between Knwldge and Comfort was found to be 0.80, which is substantial, perhaps even a bit surprising, given the relatively low reliability of the Comfort scale. We looked at a scatterplot of the two scores, and, although we didn't say anything at the time, there is a pattern there—people whose Comfort scores were low had low-ish scores on the cognitive test, Knwldge. The very highest Comfort scores, those at 40 and above, also had very high Knwldge scores, with one exception.

We could use Lertap's support for external-criterion analyses to dig a bit more here, asking for correlations between each of the Comfort items with the overall Knwldge score. We go up to the Lertap toolbar, click on Run, and then on "External criterion analysis". We tell Lertap to use the Knwldge score as the criterion, that is, the score found in column 2 of the Scores worksheet. Then, when presented with the data set's subtests, we say "No", we do not want to "work with" the Knwldge subtest, but then say "Yes" when the Comfort subtest shows up.

Lertap responds by doing its thing, producing item-level summaries of correlations, and then showing its "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:

Items Q28, Q31 and Q34 actually had negative correlations with the Knwldge score. The same three items had the lowest correlations in the ordinary analysis, as shown in the Stats2b and Stats2f reports. There are things to work on here, more questions to answer. Items Q28, Q31, and Q34 are wanting to stand on their own—the results are indicating that these three questions don't follow the response pattern exhibited by the other seven items in the Comfort subtest.

What about a relationship between answers to the Comfort items, and "experience levels"? There are two experience variables in the Data worksheet, recorded in columns 38 and 39. Column 39 indicates the number of years the respondent said s/he'd been using computers. Can we correlate Comfort responses with the data in this column of the Data sheet? Yes, you bet. However, the information in column 39 has to be copied to the Scores worksheet first. Lertap's toolbar has a set of Move options; the first of them allows a column in the Data worksheet to be copied and moved to the Scores worksheet. Once this is done, the Run menu is accessed again, and then an "External criterion analysis" is requested.

When would you want to use Move's second option? When would you want to copy a column from the Scores worksheet to the Data worksheet? Well, a common case arises when Lertap's flexible scoring capabilities are used to form affective test scores, which are then moved back to the Data worksheet, after which Lertap's 8-ball option is used to prepare the data for SPSS analyses.

More reading

There's more for you to read, to be sure. Chapters 7 and 8 talk about making more use of Lertap's various reports for cognitive and affective subtests, respectively.

Chapter 9 looks at several matters, including using the 8-ball to set things up for SPSS. Chapter 10 gets into methodology matters, and further describes how Lertap goes about its computations.

More immediately, the chapters which follow this one have to deal with setting up data sets, and creating the system control "cards" necessary to initiate the analyses you want.

Chapter 3

Setting Up a New Data Set

Contents

The last chapter was a breeze, wasn't it? You didn't have to enter any test results, they were already there. Nor did you have to create any job definition statements in the CCs worksheet as they too were supplied.

We're about to change this scene, and get you into setting up your own job. We'll start by walking you through some examples, real-life data sets from a variety of countries.

As we do this, our assumption will be that you have read through the last chapter. If you're being naughty, and starting out here instead of there, you may run into problems. 'Nuff said.

Workbooks, worksheets, rows & columns

In the last chapter you read how Lertap 5 is based on Excel, a spreadsheet program, one of the four or five applications found in the Microsoft Office suite. Microsoft Word is another Office application, probably the most popular one.

Excel's files are referred to as workbooks. A workbook is a collection, a set, of worksheets. Each worksheet is a spreadsheet. Spreadsheets are based on "pages" of rows and columns. Rows run across the page (or screen), while columns run up and down. The intersection of a row and column defines what is called a "cell".

These rows and columns are *very* much like the rows and columns of the tables which may be used in Word. If you've used tables in Word, you know that you can do all sorts of things with the table's rows and columns, such as hide their gridlines, change column widths, change row heights, and add background colours. You can do the same, and more, with the rows and columns of an Excel worksheet.

One of the big differences between Word's tables and Excel's worksheets is that the rows and columns in an Excel sheet have labels, or "headers", such as 2 for row 2, 3 for row 3, and so forth. Excel's columns may be headed by upper case letters, or by digits, depending on the user's preference. If letters are used, Excel says it's using the "A1" referencing system, and the first column's header will be "A", the second's "B". Under this header system, the cell found in the upper left corner of a worksheet is called cell "A1", denoting the intersection of column "A" with row "1".

On the other hand, if column headers are digits, Excel says it's using the "R1C1" referencing system, and what was cell "A1" becomes known as cell "R1C1", for row 1 with column 1. The cell referencing system, A1 or R1C1, is an option which users set in Excel. Lertap uses the R1C1 system.

When you open a new document in *Word*, what is it usually called? Something like Document1, or Document2, a name which sticks until you save the new document with a different name. Under Windows and NT, Word documents almost always have an extension of "doc".

When you open a new workbook in Excel, it will be called Book1, or Book2, or BookX until you save it with a different name. Under Windows and NT, Excel workbooks usually have an extension of "xls" (the "XL" means Excel, while the "s" means "sheets").

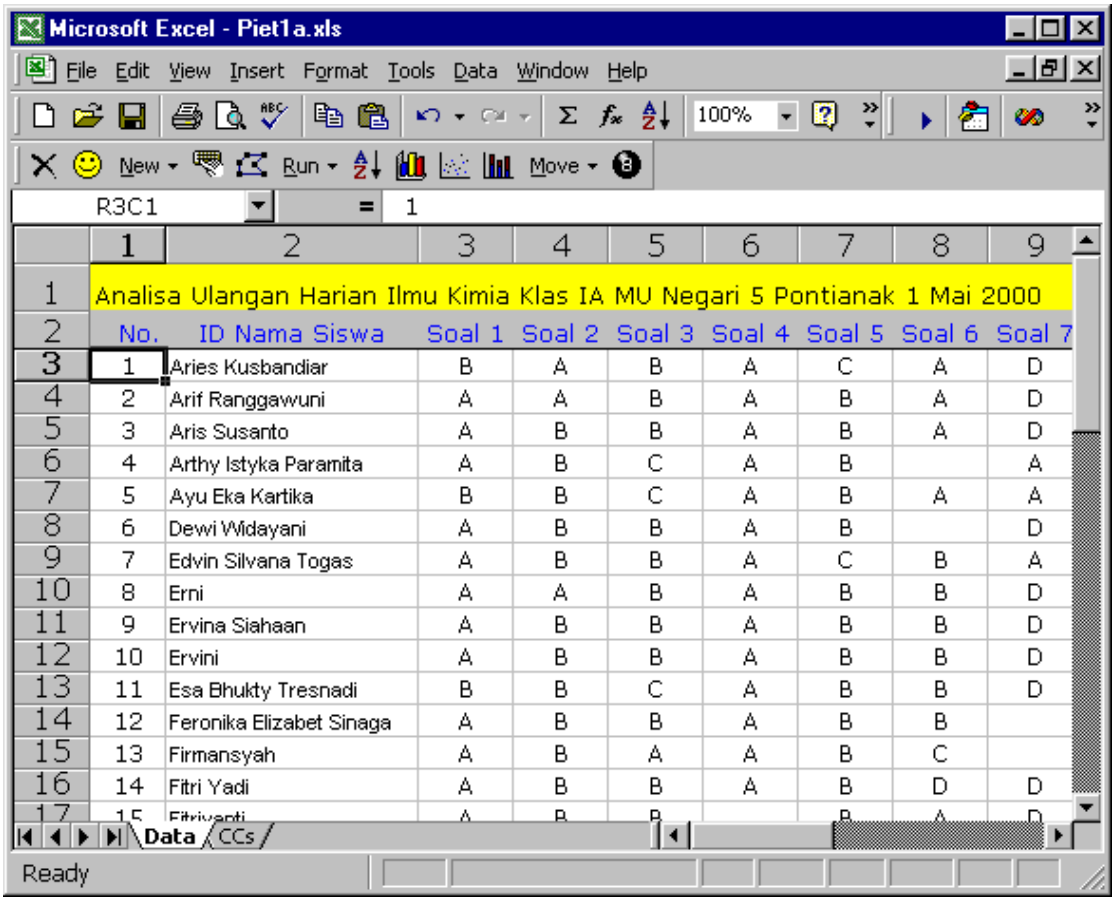
Excel is a very powerful system, and a flexible one. Users of Lertap 5 will find that increasing their familiarity with Excel will be of real benefit. Do you know how to copy a single worksheet from a given workbook to another workbook? It's easy, as is renaming worksheets. How about something as simple as changing the width of an Excel column, or the height of an Excel row? Easy. How about setting the background colour of a cell, or a row, or a column? Easy. How to find out how to do these things? Excel's on-line Help is one way, and a good one—but it can take some practice to master—be patient. How about an introductory book on the use of Excel? There are many excellent ones (for example, see Online Press (1997)).

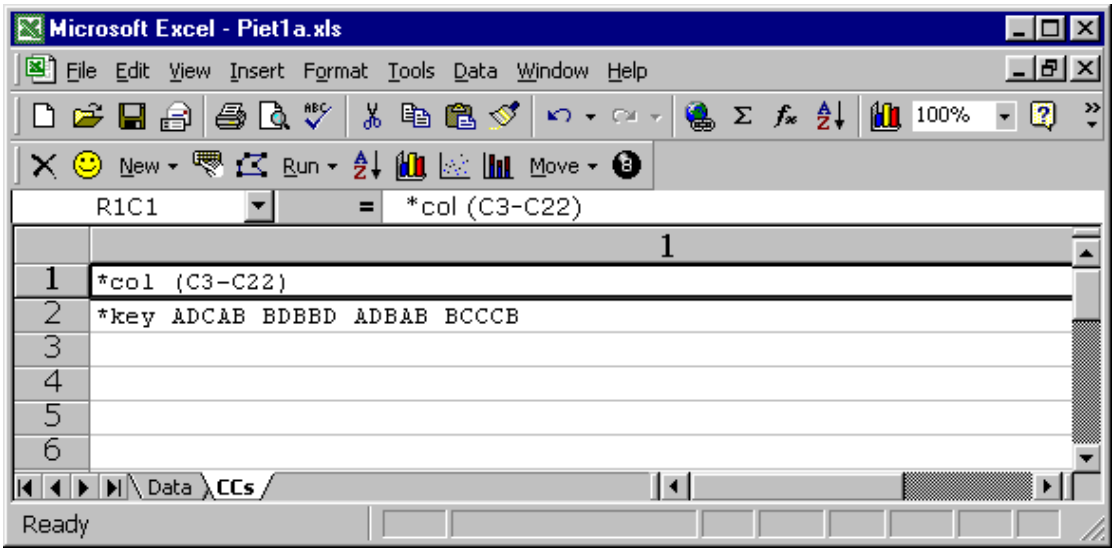
Enough of that, let's get cooking, eh?

Piet Abik's class data set

Piet Abik is a senior teacher of science at a large high school in Pontianak, an Indonesian city on Borneo's west coast. He gave one of his classes a 20-item multiple-choice test on chemistry, and used Lertap to look at results.

Below we show the partial contents of Piet's Data and CCs worksheets:





Every Lertap workbook has worksheets named Data and CCs. Test results go into the Data sheet, while job definition statements go into CCs.

Look at Piet's Data sheet for a moment. If you're seeing it in colour, you will note that the first row has its background colour set to yellow, and contains a title. The background colour of the second row is grey, and this row is used to contain headers for the columns of data.

Piet headed his columns of item responses with the word "Soal" ("Item", in English), followed by a space and then the item's sequential number. This is fine. However, had we set up this data set at Lertap headquarters, we would not have put the space after Soal. Some of the Lertap output is easier to read without that space. For example, the bands of item difficulties and discrimination coefficients found in the Stats1f sheet look better if the items have short names, and no spaces.

Another reason for eliminating spaces from item headers is that this makes it easier to pass the data to other systems, such as SPSS, a popular statistics package. SPSS will call each of the item columns a "variable", and it doesn't like spaces in variable names. (Nor does it like variable names which begin with a digit, or ones which are longer than eight characters.)

Had we been in Pontianak, sitting next to Piet as he set up his Data sheet, we would have urged him to refer to the items as "S1", "S2", "S3" and so on, not "Soal 1", "Soal 2", and "Soal 3". However, this is entirely up to users. Lertap imposes no restrictions on item names. In fact, as we'll see in a later example, you don't even have to have headers above the columns of item responses—this is not recommended, but you can leave the header row empty and Lertap will invent its own item numbers.

This is a good spot to mention that Excel has a superb shortcut for putting item headers into the Data sheet. Above Piet has his first item header in cell R3C3 (row 3, column 3), where he's typed "Soal 1". Once this is done, Excel makes it possible for "Soal 2", "Soal 3", ..., "Soal 20" to be automatically entered in the columns to the right. How? Select cell R3C3 by clicking on it. A box forms around the cell to indicate that it's selected. Grab the lower right corner of this box and drag it across the columns to the right. This is called "extending the series". If it doesn't work for you, use Excel's on-line Help for more instructions; suitable keywords are "extend series".

What else can we note about Piet's Data worksheet? He's used the first two columns as a record "No.", and as an ID field with student names ("ID Nama Siswa"). Since one of these two labels begins with the characters "ID", the names used in the second column will be picked up by Lertap and used to report scores. Had the second column been headed "Nama Siswa", this would not have happened—the scores would have simply been numbered. When Lertap finds a column header beginning with the letters "ID", it uses the information in this column to label the results in its Scores worksheet. This column must be either the first column, or the second column[15].

[15]The two crucial letters, I and D, may be either uppercase or lower case.

Piet's "No." column is just a sequential number. He could have used Excel's "extend series" shortcut to fill this column. There's no need for this column from Lertap's point of view, but Piet obviously found it useful for something.

Now let's get to the CCs sheet shown above—this is where the data analysis was defined. The first line, *col (C3-C22), tells Lertap that item responses start in column 3 and end in column 22, a span of 20 columns. The *key line has 20 letters. This is the string of right answers to the items. The correct answer to the first item, for example, was "A". The correct answer to the sixth item was "B". The correct answer to the 17th item was "C". The spaces which Piet typed in this line are not necessary, but they help read the line, we use them too.

With his results in the Data sheet, and job definition statements in the CCs sheet, what's necessary to get results from Lertap? This is what we demonstrated in the last chapter—Lertap's Run option is used twice, first to "Interpret CCs lines", and then to get the "Elmillon item analysis". This will make all those delectable, luscious worksheets of scores and statistics which you read about in Chapter 2.

Okay, shortly we'll flash up another example. But before we do, we respond to those readers who want to know why Piet's CCs sheet did not include a *sub card.

What's a *sub card? You saw examples of them in Chapter 2. Piet's CCs lines could have been these:

```
*col (C3-C22)
*sub title=(ChemTest), name=(Class test on 20 May)
*key ADCAB BDBBD ADBAB BCCCB
```

Here (above) a *sub line, often called a "card" by veteran Lertap users, has been used to provide a title for the test scores, and a name for the item analysis. The title will be seen in Lertap's Scores worksheet, while the name will be found at the top of all the item statistics sheets, such as Stats1f, Stats1b, and Stats1ul.

If a subtest has no title, Lertap will label its scores as "Test1", and will give it the same name, "Test1", when it creates all its statistical summaries.

Piet's CCs lines could have been these:

```
*col (C3-C22)
*sub res=(A,B,C,D), title=(ChemTest), name=(Class test on 20 May)
*key ADCAB BDBBD ADBAB BCCCB
```

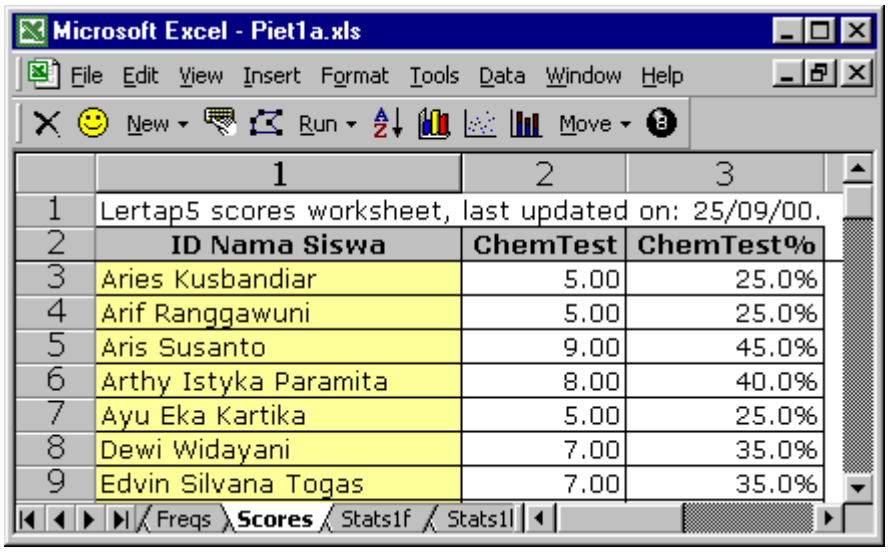
Now Piet is telling Lertap that the items in his multiple-choice test all used A B C and D as possible responses. However, this is Lertap's default assumption for cognitive subtests[16]. The res=() declaration is required on a *sub line only when the test's items use a different set of responses. In the last chapter, for example, there was a *sub card with Res=(A,B,C,D,E,F). This told Lertap that the items used as many as six different responses, from A through F.

[16] For affective subtests, the default is res=(1,2,3,4,5).

Piet's CCs lines could have been these:

```
*col (C3-C22)
*sub title=(ChemTest), name=(Class test on 20 May), per
*key ADCAB BDBBD ADBAB BCCCB
```

Now he's using the "per" control word on the *sub card, see? This asks Lertap to create a percentage score for each person, and causes the Scores worksheet to look like this:

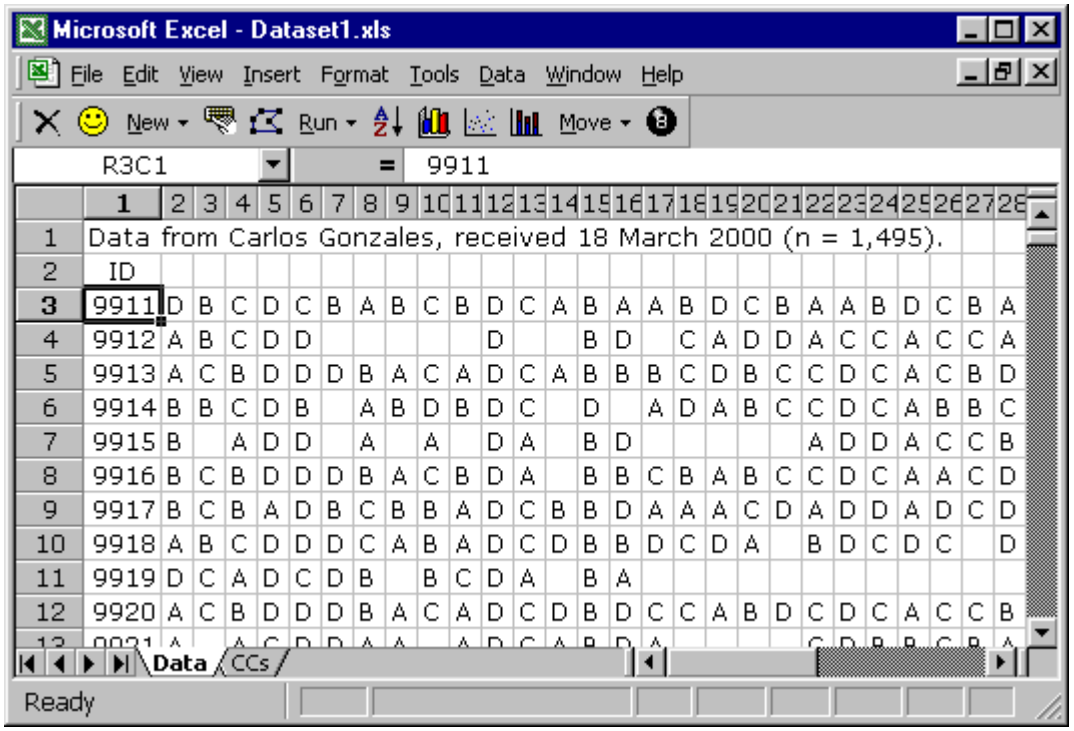


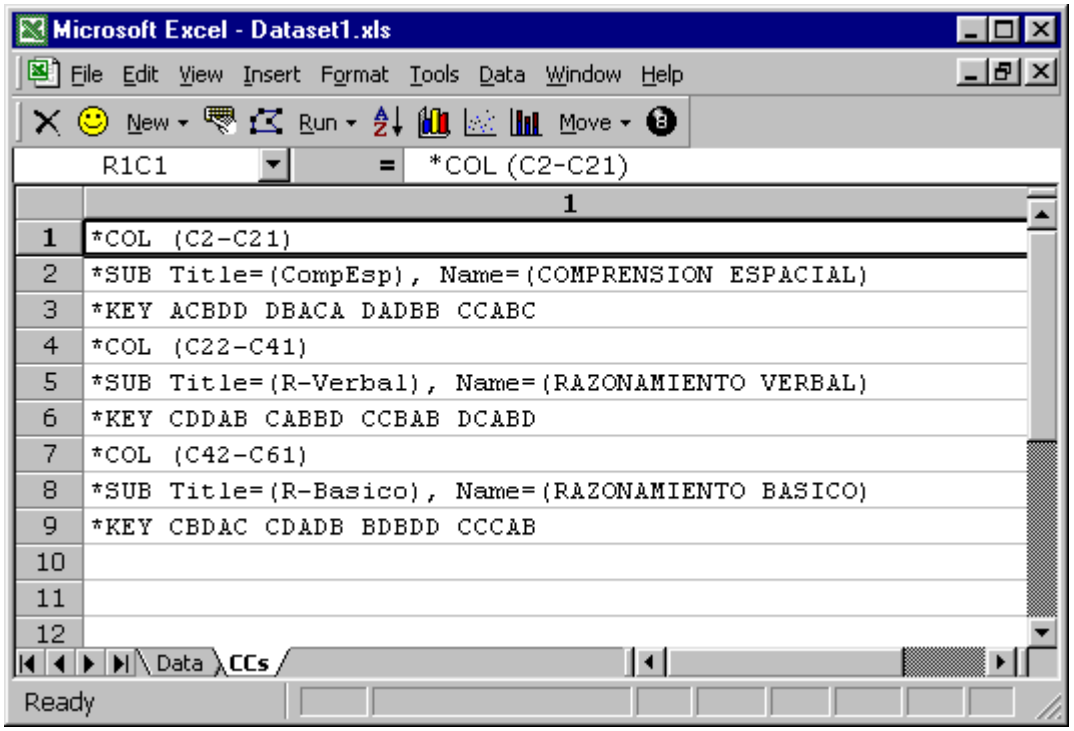
	1	2	3
1	Lertap5 scores worksheet, last updated on: 25/09/00.		
2	**ID Nama Siswa**	**ChemTest**	**ChemTest%**
3	Aries Kusbandiar	5.00	25.0%
4	Arif Ranggawuni	5.00	25.0%
5	Aris Susanto	9.00	45.0%
6	Arthy Istyka Paramita	8.00	40.0%
7	Ayu Eka Kartika	5.00	25.0%
8	Dewi Widayani	7.00	35.0%
9	Edvin Silvana Togas	7.00	35.0%

Freqs | Scores | Stats1f | Stats1l

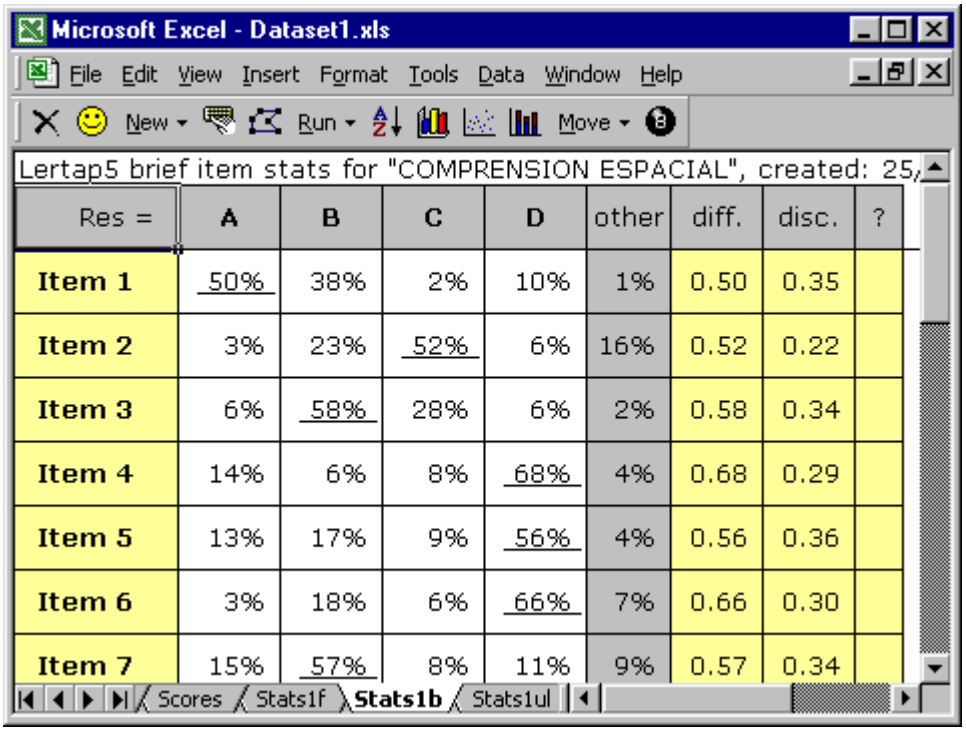
A UCV data set

Lertap was born in Venezuela, and has come to be frequently used at a number of Latin American sites. One of the most active of these is la Universidad Central de Venezuela. Have a look at one of their data sets:





In this example, the Data sheet is not as extensively prepared as Piet Abik's was. For example, there are no headers at the top of the columns which contain the item responses. This is not recommended, but it's certainly okay. In its reports, Lertap will label the items as exemplified below:

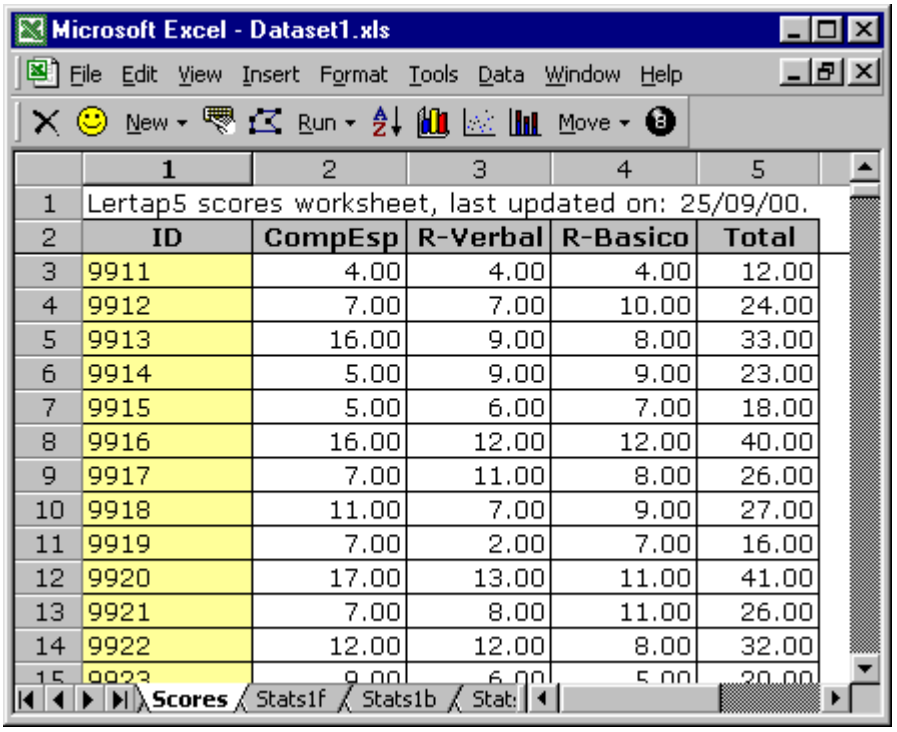


Res =	A	B	C	D	other	diff.	disc.	?
Item 1	50%	38%	2%	10%	1%	0.50	0.35	
Item 2	3%	23%	52%	6%	16%	0.52	0.22	
Item 3	6%	58%	28%	6%	2%	0.58	0.34	
Item 4	14%	6%	8%	68%	4%	0.68	0.29	
Item 5	13%	17%	9%	56%	4%	0.56	0.36	
Item 6	3%	18%	6%	66%	7%	0.66	0.30	
Item 7	15%	57%	8%	11%	9%	0.57	0.34	

Notice how the CCs sheet for this job has quite a number of lines? This is an example of a data set which involved more than one test. There were three subtests here, and the *sub cards give their titles: CompEsp, R-Verbal, and R-Basico.

Each subtest has a *col card, a *sub card, and a *key card. The item responses for the first subtest, CompEsp, start in column 2 of the Data worksheet, and end in column 21.

Lertap's Scores worksheet for this data set looked like this:



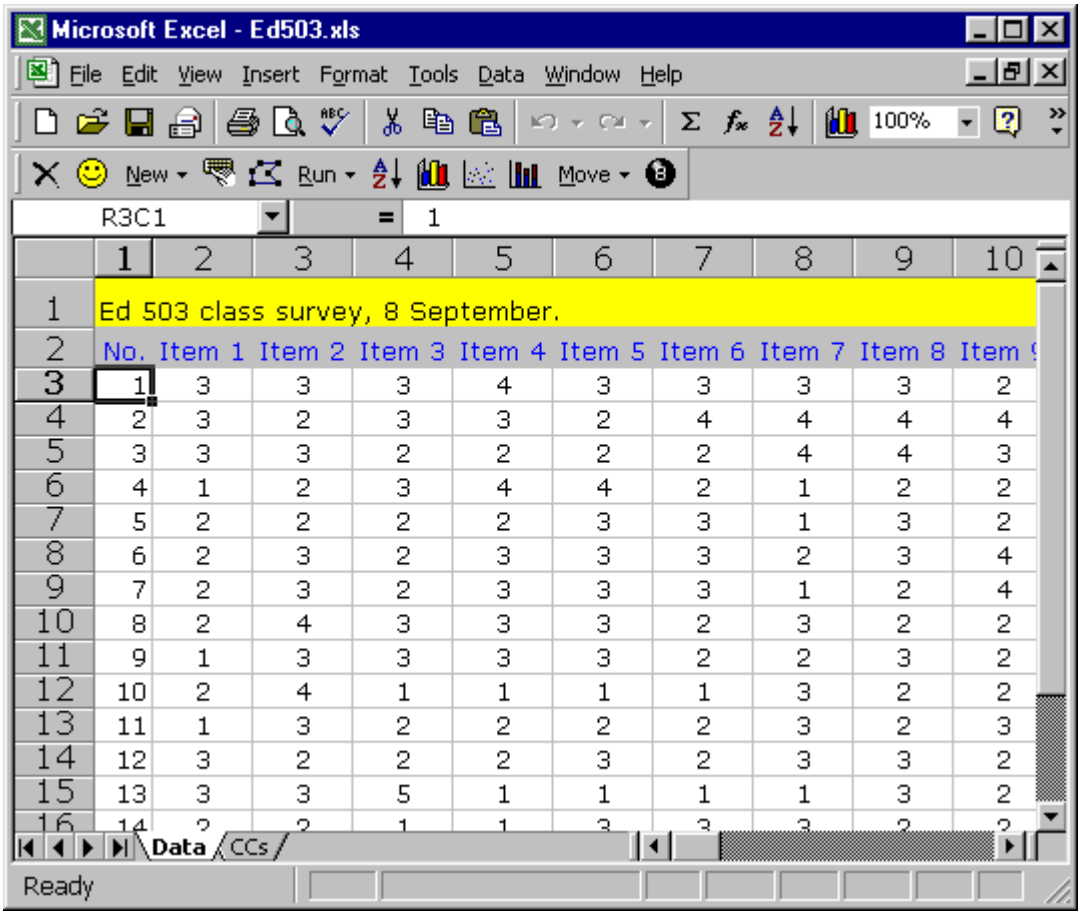
	1	2	3	4	5
1	Lertap5 scores worksheet, last updated on: 25/09/00.				
2	ID	CompEsp	R-Verbal	R-Basico	Total
3	9911	4.00	4.00	4.00	12.00
4	9912	7.00	7.00	10.00	24.00
5	9913	16.00	9.00	8.00	33.00
6	9914	5.00	9.00	9.00	23.00
7	9915	5.00	6.00	7.00	18.00
8	9916	16.00	12.00	12.00	40.00
9	9917	7.00	11.00	8.00	26.00
10	9918	11.00	7.00	9.00	27.00
11	9919	7.00	2.00	7.00	16.00
12	9920	17.00	13.00	11.00	41.00
13	9921	7.00	8.00	11.00	26.00
14	9922	12.00	12.00	8.00	32.00
15	9923	9.00	6.00	5.00	20.00

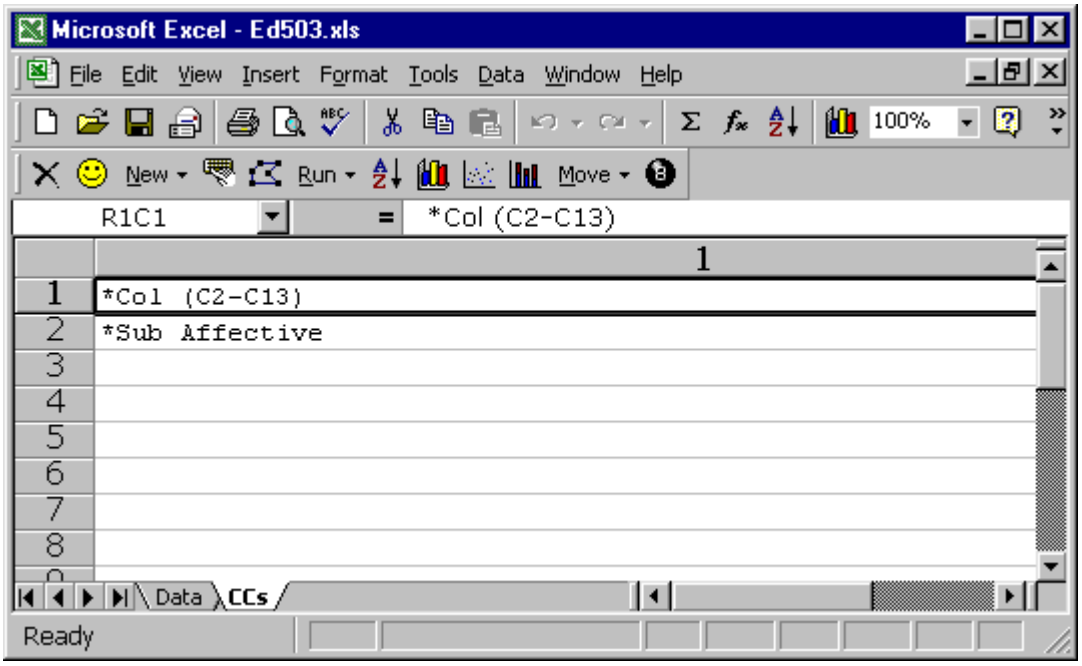
Scores Stats1f Stats1b Stat:

Whenever a job has more than one subtest, Lertap computes and displays a total score, as shown above. This score will be the sum of the individual subtest scores. It is possible to exclude subtests from the total, to combine them with different weights, or to turn off the total score computation altogether. These things are done by using Wt= specifications on *sub cards—see the example shown in Chapter 2.

A class survey

We've presented two examples so far, and both have been based on cognitive tests. Here's a simple example which from an instructor who had used a 10-item survey, an "affective" instrument, to get end-of-semester feedback from his class:





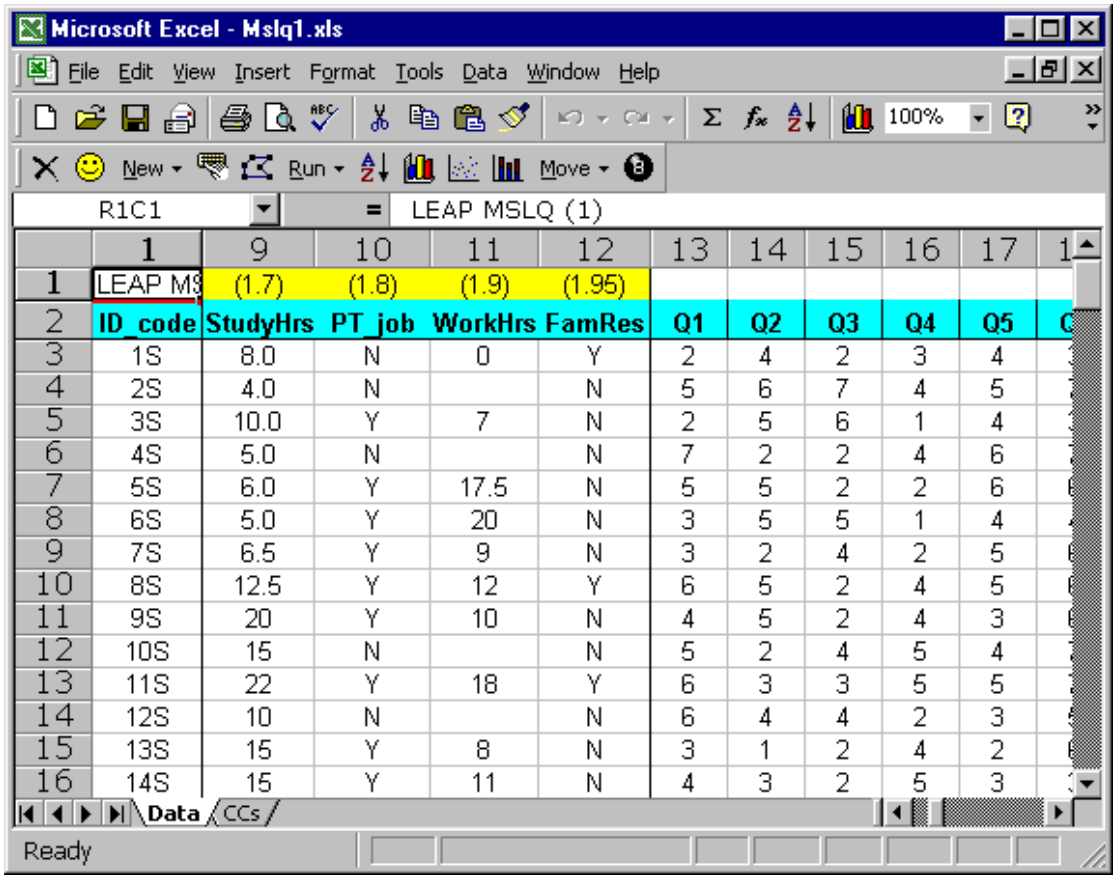
The two CCs lines for the class survey tell Lertap that item responses are found in columns 2 through 13, and that these responses are from an affective "test". The two CCs lines could also have looked like these:

```
*col (c2-c13)
*sub aff, title=(ed503), name=(Ed503 survey, Nov. 1999)
```

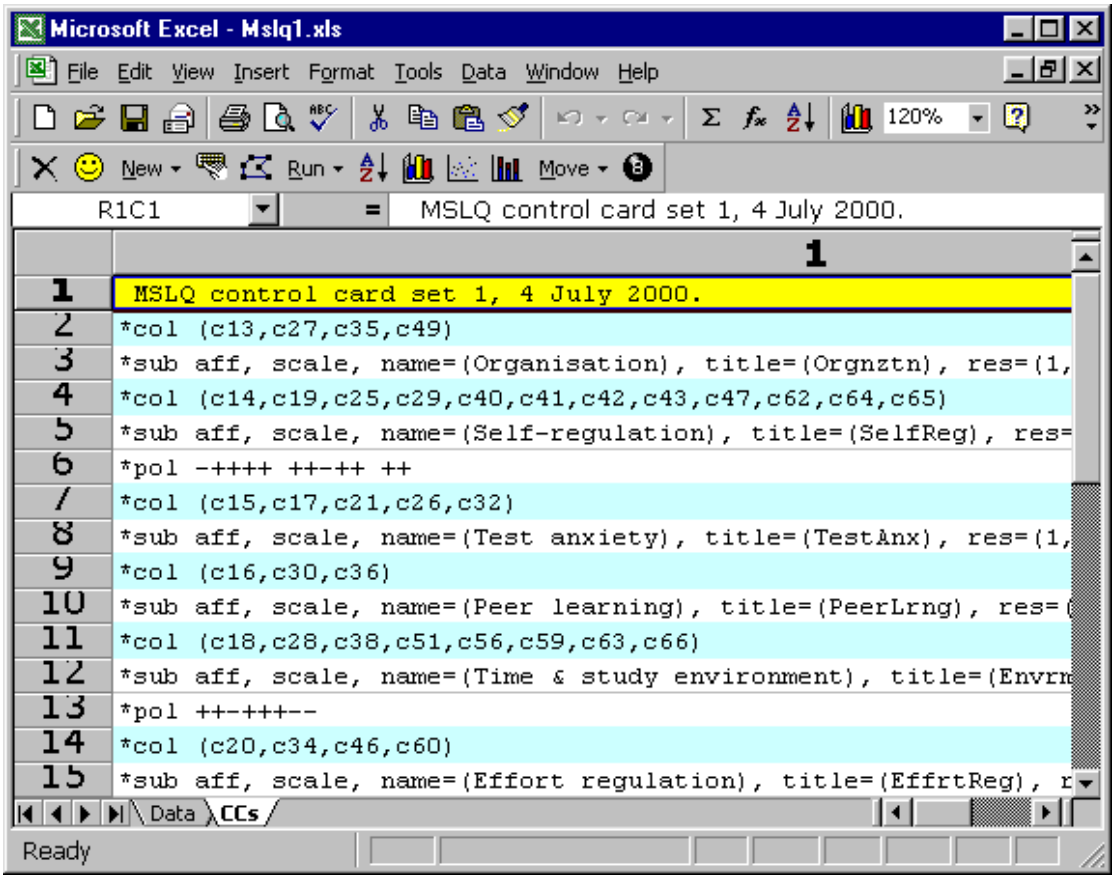
A *sub card must be used with affective subtests so that the "aff" or "affective" control word can be passed to the program. Without this control word, Lertap would assume that the items were cognitive ones, and would then expect a *key card with a string of correct answers.

A large-scale survey

This example is based on the application of the University of Michigan's Motivated Strategies for Learning Questionnaire, MSLQ (Pintrich, *et al*, 1991):



	1	9	10	11	12	13	14	15	16	17	1
1	LEAP MS	(1.7)	(1.8)	(1.9)	(1.95)						
2	ID_code	StudyHrs	PT_job	WorkHrs	FamRes	Q1	Q2	Q3	Q4	Q5	C
3	1S	8.0	N	0	Y	2	4	2	3	4	
4	2S	4.0	N		N	5	6	7	4	5	
5	3S	10.0	Y	7	N	2	5	6	1	4	
6	4S	5.0	N		N	7	2	2	4	6	
7	5S	6.0	Y	17.5	N	5	5	2	2	6	
8	6S	5.0	Y	20	N	3	5	5	1	4	
9	7S	6.5	Y	9	N	3	2	4	2	5	
10	8S	12.5	Y	12	Y	6	5	2	4	5	
11	9S	20	Y	10	N	4	5	2	4	3	
12	10S	15	N		N	5	2	4	5	4	
13	11S	22	Y	18	Y	6	3	3	5	5	
14	12S	10	N		N	6	4	4	2	3	
15	13S	15	Y	8	N	3	1	2	4	2	
16	14S	15	Y	11	N	4	3	2	5	3	



The MSLQ is a good example of an instrument with numerous embedded scales. For example, the CCs lines above indicate that the scale whose title was Orgnztn was comprised of four items, with corresponding responses found in columns 13, 27, 35, and 49 of the Data worksheet. Altogether, there were ten scales in this job—these are "subtests" to Lertap. Each subtest has a *col card and a *sub card; the subtests which correspond to scales with reversed items have *pol cards in addition. (The matter of reversed items was introduced in the last chapter.)

The Data worksheet itself has many columns which contain more than item response information. Column 9, for example, is headed "StudyHrs", while "PT_job" indicates whether or not the respondents had a part-time job. This data set is an example of one which was destined for further processing with the SPSS statistical package. In this case, Lertap was used to form the 10 MSLQ scale scores, which were then copied to the Data worksheet using a Lertap option activated via the Move option on the toolbar. This enhanced Data worksheet was then made ready for SPSS by using the toolbar's 8-ball icon.

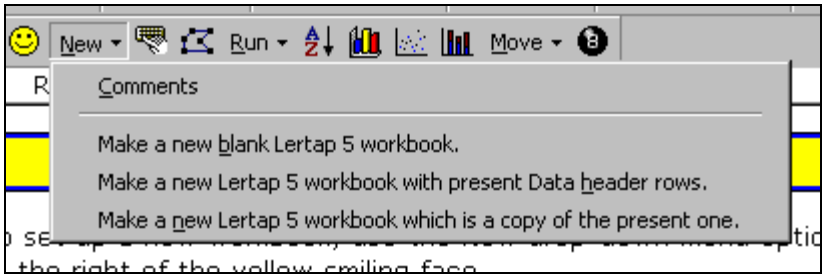
The CCs sheet above is the first we've seen with a comments line in it. It's line 1. There's no asterisk at the start of this line, which means that the line will be ignored by Lertap. Such lines may be used to add comments to the CCs worksheet.

Enough, enough? Let's get you started on your own job. Do you have a sense of what's needed? An Excel workbook with two sheets: Data and CCs. Item responses go into the Data worksheet, with a column for each item.

Setting up a new workbook

If you're an experienced user of Excel, all that's required to set up a new Lertap workbook is to know that Lertap wants a book with two worksheets, one called Data, the other called CCs. The Data sheet must use its first two rows for title and header information, as seen in the examples above.

About the easiest way to establish a new workbook, be you Excel experienced or not, is to use the New option on Lertap's toolbar. Here's a picture of its options:



The first of the three "Make a new ..." options sets up a new Excel workbook with two worksheets. They will be named Data and CCs, and both will be completely blank, or empty.

The second of these three options also creates a new workbook with two worksheets, Data and CCs, but now the Data sheet will not be empty—it'll have its first row with a title in it, and its second row set up with column headers. The title and column headers will be copied from the workbook which was active when the option was taken.

At this point we toss the ball over to you. You're here because you've got data to enter. You first need to open a new workbook, and see that it has a worksheet named Data, and another named CCs. Please do this, using the New option which seems to best suit your needs (we would think it would be the first option, but it could be the second).

Entering data

The first row in the Data worksheet is where you may type a title for your job. Make sure you're in R1C1 of the Data sheet, that is, row 1, column 1, and type whatever you want. Lertap makes no use of the information in this line, so you can type absolutely anything you wish. (You could even leave the line blank, if you want.)

The second row in the Data worksheet is meant for column titles. If you have ID information, such as student names or ID numbers, put this in either the first column, or the second, and make sure the corresponding column title begins with the letters "ID". Of the four examples given above, the first two, and the last,

have such a column. In two of these examples, it's the first column, while in one, Piet Abik's, it's the second.

The columns with item responses may begin anywhere in the Data sheet. Our suggestion (for English speakers), is that the item titles follow a pattern such as "Q1, Q2, Q3, ...", or "I1, I2, I3, ...", or "Itm1, Itm2, Itm3, ...", or "Item1, Item2, Item3, ...". The briefer the title, the better some of Lertap's output will look. If it's likely data will be exported to another program, such as SPSS, it will be very useful to make sure item titles conform to the naming conventions of the other system—SPSS, for example, spits out item titles which are longer than eight characters; which begin with a digit;[17] and which contain a space.

Entering item responses

Let's imagine that you have a 10-item test which uses response codes of A B C and D. Let's suppose that a person named Arkin has answered the 10 items this way:

BCDBAACCDB

To type in Arkin's responses, you'd ordinarily go to the column set aside for the first item, type an A, move to the next column to the right, type a C, move to the next column to the right, type a D, and so on.

How you move from one column to the next column to the right depends on you. Excel lets you use either the Tab key or the right-arrow key for this. If you set Excel up correctly, you can also use the Enter key (in Excel 2000, the "Move selection after Enter" option on the Tools / Options Edit tab defines what happens after the Enter key is pressed).

Lertap has a couple of shortcuts available which make it possible to type in a person's responses as one long string, have the string dissected, and its contents spread over columns to the right. These options are represented by two icons on Lertap's toolbar:  .

It would be worthwhile experimenting with these two options—they can save you quite a bit of typing time. To do so, go to a cell in a row where you'd usually start typing item responses. However, instead of starting to type, click on the first of these two icons, and then type the string of item responses into the data entry box which will appear—do not use spaces in the string. See what happens.

Then do the same in the next row, but use the second of the icons, the "Spreader".

If your item responses are letters, not digits, then both of these icons result in the same action: the string of responses is taken apart, and distributed, one by one, over the columns to the right. If you're using the Spreader, it will spread the responses in the row you start from, and then walk down the rows, one by one, continuing to spread responses until it finds an empty row. (Press the

[17]There is more about SPSS variable names in Chapter 9.

<Esc> key to get the Spreader to stop its walk. Use the toolbar's smiling yellow face to restore the Status Bar so that is says "Ready".)

If, however, your item responses are digits, only the first of the two icons can be guaranteed to work as wanted—the Spreader (the second icon) may fail.

The reason the Spreader fails is that it lets you type directly into the cell on the worksheet. If you do this, and if what you type begins with a digit, then Excel thinks you're entering a number into one of its cells, and things can come undone—your string of item responses will often get converted to a number in "scientific" notation, such as the following: 3E+10 (which is what Excel displayed after we typed a string of 10 responses, 32451245, into a cell on the worksheet). In order to defeat this and have the Spreader work correctly with a string of digits, type a single quotation mark before the first digit ('32451245). This tells Excel to treat your string as textual information, not a number. Another way to defeat this, a more convenient way, is to highlight the column where you start typing item responses, and get Excel to format all cells in the column as text (use Format / Cells). Once you've done this you no longer need to use the single quote before the string of digits.

How many rows and columns?

The Data worksheet may have as many rows and columns as Excel permits. In Excel 2000, there may be 65,536 rows, and 256 columns. To date the largest Data worksheets we have heard of have been processed at la Universidad Central de Venezuela, where data sets with up to fourteen thousand records have been used with Lertap 5 on Pentium-based computers, running with Excel 97, under Windows 98. At Lertap headquarters we have processed data sets with just under ten thousand records on a Pentium III, using Excel 2000 and NT 4.0.

To be noted is that Lertap stops reading information from Data and CCs work-sheets when it encounters a row whose first column is empty. In large data sets, with thousand of rows in the Data sheet, it is sometimes useful to test Lertap by inserting a blank row after, say, row 200. This will let you speedily try the "Interpret CCs lines" and "Elmillon item analysis" options to see if you've got things set up correctly before analysing all data. (For information on Lertap time trials, please refer to Chapter 10, Computational Methods. Data sets with thou-sands of records can take several minutes to process on some computers.)

Missing data

What should be done when someone does not answer a question? We suggest that non-answers be processed as a space, or what Excel refers to as a "blank". For example, let's say that Arkin did not answer item 5. Then his string of item responses would be (using the sample 10-item response string from above):

BCDB ACCDB

When typing a string such as this into Excel, we'd use a space as the fifth response, that is, we'd use the keyboard's Space Bar.

Importing data

Data may be imported to the Data worksheet using Excel's standard options. In Excel 2000 these are found under Data / Get External Data options. If the data are already in an Excel worksheet in some other workbook, go to that workbook, select the sheet with the data (by clicking on its tab, for example), and then copy this worksheet to your new Lertap workbook. Once it's there rename it as Data.

Creating the CCs lines

By far the most laborious part of a Lertap job is getting the Data worksheet set up with records of item responses. Once this is done, job definition "statements" must be entered in the lines of the CCs worksheet.

All Lertap subtests have to have a *col () line in the CCs worksheet.

Cognitive subtests also require a *key card, a line in the CCs worksheet which contains a string with the correct answers. Affective subtests require a *sub card with the control word "Aff" or "Affective" on it.

It is, then, always the case that a minimum of two CCs lines are required for each subtest to be processed—cognitive tests require *col and *key, affective tests require *col and *sub. Many times additional "cards", or lines, are used. At Lertap headquarters, for example, we always use a *sub card for cognitive tests in order to give the test a name and a title.

You've already seen several examples of CCs lines in this chapter, and in the last. There are more in the chapters which follow. Don't forget that "on-line" examples of CCs lines may be found in the Syntax worksheet in the Lertap5.xls file, as mentioned in Chapter 2.

CCs lines must all begin in column 1 of their respective rows. Lines with an asterisk at the front, as the very first character in the line, are parsed by Lertap (examined by the CCs interpreter). Lines which do not have an asterisk as their first character are considered to be comments, and are totally ignored by the program.

It is sometimes useful to type up CCs lines for a subtest, but have Lertap ignore them temporarily. This happens in complex jobs, such as the MSLQ example above, when users want to limit attention, for the moment, to just some of a job's subtests. At Lertap headquarters we take subtests "off line" by putting a space in front of the asterisks of the corresponding CCs lines. Another way to do this is to put a single quote in front of the asterisk. For example, in the following job the second subtest, SelfReg, has been taken "off line":

```
*col (c13,c27,c35,c49)
*sub aff, scale, name=(Organisation), title=(Orgnztn), res=(1,2,3,4,5,6,7), wt=0
'*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), wt=0
'*pol -++++ ++-++ ++
*col (c15,c17,c21,c26,c32)
*sub aff, scale, name=(Test anxiety), title=(TestAnx), res=(1,2,3,4,5,6,7), wt=0
*col (c16,c30,c36)
*sub aff, scale, name=(Peer learning), title=(PeerLrng), res=(1,2,3,4,5,6,7), wt=0
```

Making changes to CCs lines

It is common for users to change their CCs lines. They might do so, for example, only to correct an error as simple as the one seen below, where there's a spelling mistake in the *sub card's name= assignment:

```
*col (C3-C22)
*sub title=(ChemTest), name=(Clas test of 20 May)
*key ADCAB BDBBD ADBAB BCCCB
```

Once changes have been made to CCs lines, both the "Interpret CCs lines" and "Elmillon item analysis" options have to be applied again in order for the changes to be effected.

Just Freqs?

We mentioned above how inserting a blank row in the Data worksheet is sometimes useful. The same holds true for the CCs worksheet. For example, the following four CCs lines

```
*col (C3-C22)

*sub title=(ChemTest), name=(Class test of 20 May)
*key ADCAB BDBBD ADBAB BCCCB
```

would result is a "Freqs only" analysis.

When the "Interpret CCs lines" is taken with CCs cards such as these four, Lertap reads only the first line, creates its Freqs worksheet, and stops without going on to make the Sub worksheet mentioned in the last chapter. It will do this because all processing halts once a blank line is encountered in the CCs worksheet.

Getting results

What's the title of this chapter? "Setting Up a New Data Set".

What is a data set? An Excel workbook tailor-made for use with Lertap. Such workbooks have item response data in a worksheet called Data, and job definition statements in another worksheet called CCs.

Do you need Lertap to set up a data set? No, not really. Anyone with sufficient nous can get Excel to create a new workbook, and then see that it has two worksheets, one named Data, the other CCs.

You may not truly need Lertap to set up a new data set, but it can definitely assist in setting up a new Excel workbook which meets the requirements of a Lertap data set. For example, Lertap's toolbar has its New option which simplifies the creation of an Excel workbook. And Lertap's toolbar has two data entry assistants, one of them called "the Spreader", which can ease the process of typing in all those item responses. But we repeat: Lertap is not really required to set up a data set. Helpful at times, but not critical.

It's when you want to get results that Lertap becomes indispensable. You have to open the Lertap5.xls file in order to have access to the all-conquering Run option, the gateway to "Interpret CCs lines" and "Elmillon item analysis". It's only when these tools are available that you can get those wonderful, information-laden Lertap worksheets with scores and item statistics.

At this point we'll power down, hoping that this chapter, and Chapter 2, have given you all that's required to get set up and running. When your analyses become more complex, or you want to turn on more of Lertap's options, the information in the following three chapters on Lertap control "cards" will be useful.

Chapter 4

An Overview of Lertap 5 Control "Cards"

Contents

You've worked your way through Chapters 2 and 3. Good. You've seen several examples of the job definition statements which go into a data set's CCs worksheet. Very good. You're ready to read more about these statements. Excellent.

Lines, or cards?

The job definition statements which go into the CCs sheet have been referred to as "lines" at times, and as "cards" at other times. The terms are synonymous; whether you hear a Lertap user saying "lines", or "cards", they're talking about the same thing—the term they prefer will be related to their age.

At Lertap headquarters we find it difficult to use any term other than "cards". This is because Lertap came into the world at a time when "punch cards" were the main means of communicating with computers. Such cards were also known as "Hollerith" cards after the man who devised a system for coding information on them (see Sanders, 1981, p.24-25).

The second version of Lertap emerged in 1973. Towards the end of the 70s, and well into the 80s, Lertap 2 was almost a household word. This version was accompanied by an extensive manual (Nelson, 1974), and much of the manual had to do with how to prepare punch cards for Lertap analysis.

At the time these control cards were unique in that they used a free-form, sentence-like syntax which was, at the time, novel. Lertap 2 control cards made a bit of a name for themselves.

The job definition statements seen in Lertap 5, the new version, are very similar to what Lertap 2 users formerly punched onto cards. Hence our inclination to refer to CCs lines as cards.

Lines, or "cards"—in this and the following two chapters the terms will be used interchangeably.

Review (examples from previous chapters)

Let's go over some of the sample control cards which you've already had the chance to admire.

```
*col (C3-C22)
*key ADCAB BDBBD ADBAB BCCCB
```

These two cards, a *col and a *key card, are all that are required to get Lertap to process a cognitive test (or "subtest"). The *col card refers to columns in the workbook's Data worksheet where item responses are to be found. In this example, the responses begin in column 3 of the Data worksheet, and end in column 22, an inclusive span of 20 items.

The *key card gives the keyed-correct answer for each item, starting with the first item. Thus, the right answer to the first item, the one whose responses are in column 3 of the Data worksheet, is A. The right answer to the sixth item is B.

The right answer to the last item is also B.

Let's talk a bit about "syntax", about the "rules", for creating these cards.

They must each have an asterisk as their first character. They must have a space after the card type, col and key. Case is generally not important. These cards are identical to the two above:

```
*COL (c3-c22)
*KEY ADCAB BDBBD ADBAB BCCCB
```

Note that the keyed-correct answers are all upper-case letters. The following *key card is valid, but it's not the same as the *key card above:

```
*KEY adcab bdbbd adbab bcccb
```

The keyed-correct answers must be in the same case used in the Data worksheet to record item responses. In Lertap 5, item responses may be upper- or lower-case letters, or digits. If lower-case letters are used to record answers to cognitive test items, then the letters on the *key card have to be lower-case too.

Let's consider another example:

```
*col (C3-C22)
*sub title=(ChemTest), name=(Class test on 20 May)
*key ADCAB BDBBD ADBAB BCCCB
```

A *sub card has been added to the two control cards. This *sub card is said to use two control "words", or "declarations", or "assignments". They are

title=(ChemTest)
name=(Class test on 20 May)

There are quite a number of control words which may be used on the *sub card; the next two chapters provide more details. For the moment we'll point out that the *sub card's control words must be separated by columns, and their case is not important. The two declarations could have been

Title=(ChemTest)
Name=(Class test on 20 May)

or

TITLE=(ChemTest)
NAME=(Class test on 20 May)

or

T=(ChemTest)
N=(Class test on 20 May)

The last example (above) indicates that only the initial letter of a control word is required. In fact, this also applies to control cards. The following three cards are valid:

```
*c (C3-C22)
*s title=(ChemTest), name=(Class test on 20 May)
*k ADCAB BDBBD ADBAB BCCCB
```

Here's another control card set lifted (and modified) from Chapter 2:

```
*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
```

The set of four cards above define a subtest with 25 items. Item responses are found starting in column 3 of the Data worksheet. The responses used by the items in this subtest involved six upper-case letters, as indicated by the `res=(A,B,C,D,E,F)` declaration on the *sub card. If an res=() declaration is not found, Lertap internally assigns `res=(A,B,C,D)`.

The *alt control card indicates that the first item in this subtest used only the first 3 of the responses found in `res=(A,B,C,D,E,F)`. That is, the first item used just A B and C as possible responses. A scan of the *alt card shows that only two items used all 6 possible responses (the ninth item, and the last).

If a *alt card is not found, Lertap assumes that all items of the subtest use all of the responses in the res=() declaration.

Have you noticed how we changed the font for the four control cards shown above? If a fixed-pitch font is used, such as Courier or Courier New, then the characters in the *key and *alt cards will line up, making them easier to read.

While we're on this matter of things being easy to read, the strings found in the *key and *alt cards do not have to have spaces in them. We think spaces after every five characters make these cards easier to read, but they're not required. The cards above could be:

```
*col (c3-c27)
*sub res=(A,B,C,D,E,F), name=(Knowledge of LERTAP2), title=(Knwldge)
*key AECABBEBBDADBABBCCCBBABDC
*alt 35423354645432443344455 46
```

What about the two cards below?

```
*col (c2-c13)
*sub aff
```

What's happened here? There's no *key card? No, these two cards pertain to an affective test with 12 items. Lertap knows a subtest is to be processed as an affective one whenever it encounters the "aff" control word on a *sub card. These two simple cards are all that would be required to process an affective subtest, providing the items use res=(1,2,3,4,5) as their responses.

Finally, let's look again at the CCs lines which come in the Lertap5.xls workbook:

```
*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 +---- ++--+
```

Any Lertap analysis can have many subtests. This example has two.

The definition of each subtest begins with a *col card. Sometimes users add something to make it easier to see where subtest definitions begin. This may be done in a variety of ways—below, for example, a line with a single quotation mark has been used as a separator:

```
*col (c3-c27)
*sub name=(Knowledge of LERTAP2), res=(A,B,C,D,E,F), title=(Knwldge)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
`
*col (c28-c37)
*sub aff, title=(Comfort), name=(Comfort with using LERTAP2)
*pol +---- ++--+
```

It is not possible to use blank lines as separators. Lertap stops reading the CCs lines whenever it encounters a line with an empty initial column.

Special control cards, *tst included

The most commonly-used control cards are *col, *sub, *key, and *pol, but there are a few others in Lertap's repertoire.

One of these others is the *alt card, mentioned in the examples above. Another is the ***wgs** card. You'll see it in the next chapter—it allows the correct answer to a cognitive item to get more than one point.

There's a particularly-powerful card, ***mws**, used in special situations to give multiple weights to item responses. You'll see this one featured in the next two chapters. The *mws card permits cognitive items to have more than one right answer, and it also finds frequent use in special affective situations.

There is one card which, when used, has to be at the very top of the CCs worksheet. It's the ***tst** card.

The *tst card makes it possible to break out subsets of the records in a data set. Here's an example:

*tst c12=(M)

This card tells Lertap to make a new workbook, one which will have a copy of those records in the active data set which have an M in column 12 of the Data worksheet. The new workbook will have a CCs worksheet too; its contents will be identical to those in the active workbook, missing out the *tst card.

*tst c9=(C,F), c22=(7)

The *tst card above gets Lertap to make a new data set. Those records in the active data set which have either a C or an F in column 9, and a 7 in column 22 of the Data worksheet will be copied to the new data set. All CCs lines found in the active data set, except the *tst card, will be copied to the new data set's CCS sheet.

By "active data set" we mean the workbook whose CCs worksheet has the *tst card.

Examples from the Syntax sheet

After a while Lertap's control cards will be second nature. However, should you forget what they look like, the Syntax worksheet in the Lertp5.xls workbook is useful. For example, here are *some* of the sample cards which come in the standard Syntax sheet:

***col (c3-c12)**
Defines a 10-item subtest whose responses are found in columns 3 through 12 of the Data worksheet. The space after *col, before the opening parenthesis, is required.

***col (c1, c3, c5, c7, c9, c11, c13, c15, c17, c19)**
Defines a 10-item subtest whose responses are found in every other column of the Data worksheet, beginning with column 1.

***col (c1, c3, c5-c9, c21, c31-c37)**
Defines a 15-item subtest whose responses are found in various columns of the Data worksheet.

***sub aff**
Defines an affective subtest. Since there is no explicit **res=**() declaration, Lertap will assume res=(1,2,3,4,5). *The *sub card is required for affective subtests.*

***sub aff, res=(1,2,3,4,5,6), name=(Class survey 1.), title=(Survey1)**
Defines an affective subtest. The **name=()** and **title=()** are optional, but recommended. The title=() can have 8 characters at most. The res=() declaration may have up to 10 response characters, separated by commas.

***sub aff, mdo**
Defines an affective subtest with, **mdo,** the missing-data option, off. Respondents who do not answer an item will get zero points.

***sub name=(Hopkins test, chaps 1 - 5), title=(HopTest1)**
Gives a name and title to a cognitive subtest. *It is the absence of the* **aff** *control word which defines a cognitive test.* Since there is no explicit res=() declaration, Lertap will assume res=(A,B,C,D).

You can modify the contents of the Syntax worksheet, putting in some of your own favourite examples. See Chapter 10, Computational Methods, to find out how.

Copying CCs lines

One of the big advantages of using job definition statements, or, if you will, "control cards", is that you can copy them.

There are several ways you can copy CCs lines from a worksheet in one workbook to a CCs sheet in another. The most obvious is to follow a standard select, copy, and paste procedure: select the lines in one worksheet, copy them, go to the other worksheet, and paste them.

But there are faster ways. One is to use Lertap's New option to "Make a new Lertap 5 workbook with present data header rows". Another is to select the CCs worksheet by right-clicking on its tab, and then Move or Copy it to another workbook (Excel methods such as this are discussed in Chapter 9).

Codebooks for tricky jobs

Some Lertap users, the luckiest (to be sure), get into data sets which have, first of all, many subtests, and, secondly, non-contiguous items.

Look at this set of control cards, for example:

```
MSLQ control card set 1, 4 July 2000.
'
*col (c13,c27,c35,c49)
*sub aff, scale, name=(Organisation), title=(Orgnztn)
'
*col (c14,c19,c25,c29,c40,c41,c42,c43,c47,c62,c64,c65)
*sub aff, scale, name=(Self-regulation), title=(SelfReg)
*pol -++++ ++-++ ++
'
*col (c15,c17,c21,c26,c32)
*sub aff, scale, name=(Test anxiety), title=(TestAnx)
'
*col (c16,c30,c36)
*sub aff, scale, name=(Peer learning), title=(PeerLrng)
'
*col (c18,c28,c38,c51,c56,c59,c63,c66)
*sub aff, scale, name=(Time & study environment), title=(Envrmnt)
*pol ++-+++--
'
*col (c20,c34,c46,c60)
*sub aff, scale, name=(Effort regulation), title=(EffrtReg)
*pol -+-+
'
*col (c22,c33,c37,c52,c57)
*sub aff, scale, name=(Critical thinking), title=(CritThnk)
'
*col (c23,c31,c45,c58)
*sub aff, scale, name=(Rehearsal), title=(Rehearse)
'
*col (c39,c48,c50,c53,c55,c67)
*sub aff, scale, name=(Elaboration), title=(Elaborat)
'
*col (c24,c44,c54,c61)
*sub aff, scale, name=(Help seeking), title=(HelpSkng)
*pol -+++
```

The set of control cards shown above was set up to process a version of the University of Michigan's MSLQ instrument (Pintrich, *et al*, 1991).

Before we get into the main points of this little section, we might point out that this set of cards exemplifies the use of comment and separator lines. Any line in the CCs sheet which does not have an asterisk as its first character is ignored by

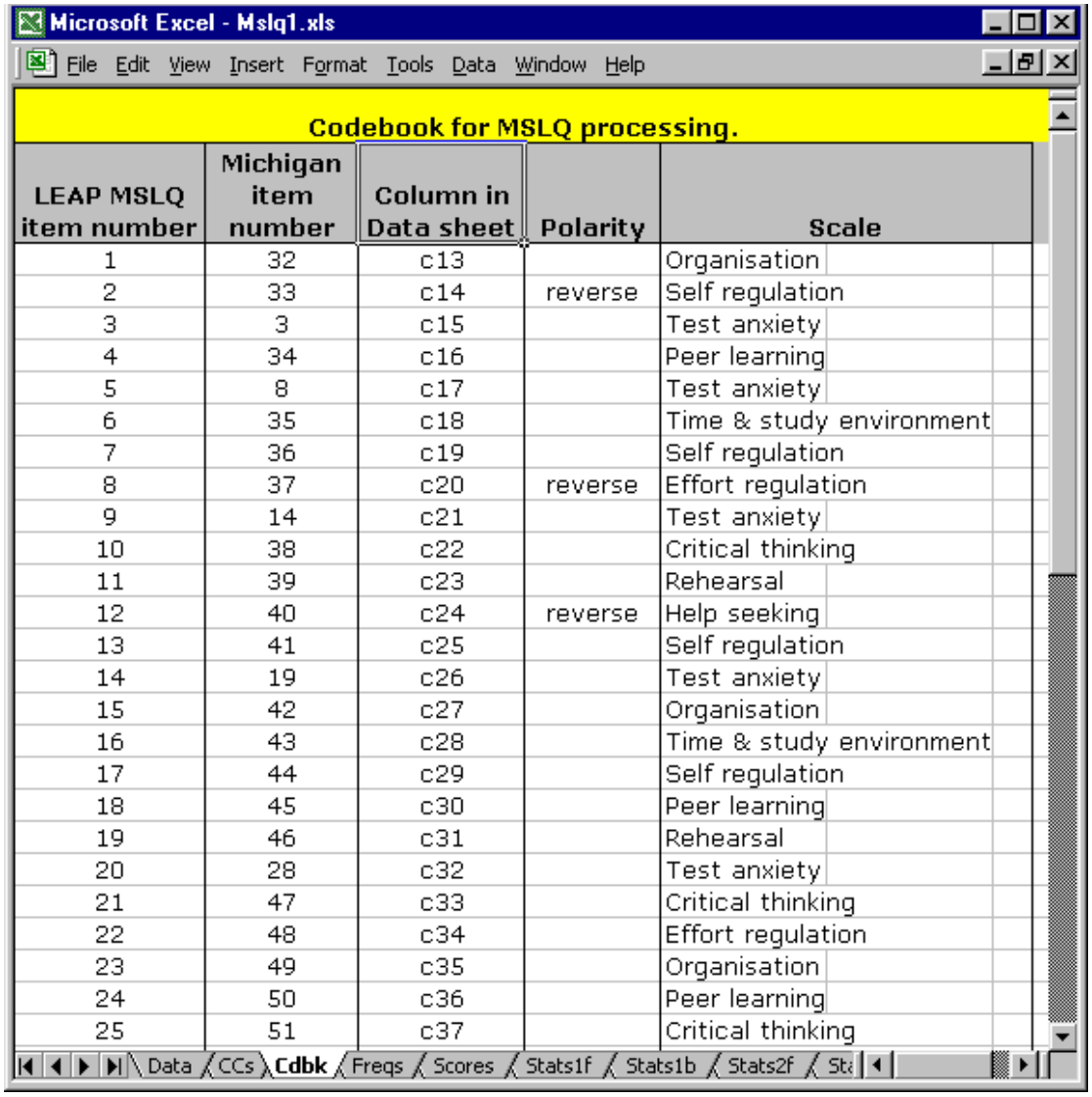
Lertap's CCs parser. Non-asterisk lines may be used for adding comments, as seen in the first line above, and as subtest separators, as seen in the lines which begin with a single quotation mark.

How many subtests are there above? Ten. Count the number of *col cards—these indicate the start of subtest definitions.

How many items are used by the first subtest, the one whose title is Orgnztn? Four. Count the number of columns pointed to in the respective *col card. Where in the Data worksheets are the item responses for this subtest? In columns 13, 27, 35, and 49.

These columns are not contiguous, they're not next to each other.

In jobs like these it is often of real assistance to set up what some experienced data processors refer to as a "codebook". We set one up for this job, using an Excel worksheet for the task:



Microsoft Excel - Mslq1.xls

File Edit View Insert Format Tools Data Window Help

Codebook for MSLQ processing.

LEAP MSLQ item number	Michigan item number	Column in Data sheet	Polarity	Scale
1	32	c13		Organisation
2	33	c14	reverse	Self regulation
3	3	c15		Test anxiety
4	34	c16		Peer learning
5	8	c17		Test anxiety
6	35	c18		Time & study environment
7	36	c19		Self regulation
8	37	c20	reverse	Effort regulation
9	14	c21		Test anxiety
10	38	c22		Critical thinking
11	39	c23		Rehearsal
12	40	c24	reverse	Help seeking
13	41	c25		Self regulation
14	19	c26		Test anxiety
15	42	c27		Organisation
16	43	c28		Time & study environment
17	44	c29		Self regulation
18	45	c30		Peer learning
19	46	c31		Rehearsal
20	28	c32		Test anxiety
21	47	c33		Critical thinking
22	48	c34		Effort regulation
23	49	c35		Organisation
24	50	c36		Peer learning
25	51	c37		Critical thinking

Data / CCs / Cdbk / Freqs / Scores / Stats1f / Stats1b / Stats2f / St

Setting up Lertap control cards is a lot easier when a codebook has been set up, especially when subtest items are not contiguous.

Summary

There aren't all that many control cards in Lertap—less than 10. They have a particular form to them which has to be followed, a definite "syntax", but it's not a complex "language" at all. With practice and reference to examples you will, we truly trust, get up to speed in short order.

We've gotten into some non-standard data sets in this chapter, and have at times used a fair number of control cards to set up our jobs. It is possible that some readers will be wondering how long it might take them to become CCs experts, and to them we say *relax, go slow*. Most users do not need to become CCs whizzes. Many users have jobs with are very straightforward. Teachers with

their own classroom tests will often be able to use just *col and *key cards, and these are easy to set up. Classroom surveys can often be processed with just two cards too, *col and *sub.

The next two chapters go into control cards in more detail, first for cognitive subtests, then for affective ones. Among other things, you'll see how to "double-key" cognitive items (allow for more than one right answer), and how easy it is to use affective scales, or subtests, having items which need to be reverse scored.

Chapter 5

Control Cards for Cognitive Subtests

Contents

The purpose of this chapter is to discuss the various Lertap control "cards" used for processing results from cognitive tests and quizzes. The term cognitive refers to an instrument which is meant to measure knowledge, or achievement. For Lertap to work effectively, the items (or questions) used by the instrument must use fixed-choice responses—multiple-choice and true-false questions are examples of items use fixed-choice responses.

Here are some sample cognitive items:

(2) What control word is used on which control card to activate the correction-for-chance scoring option?

A) MDO on *ALT
B) CFC on *TST
C) WT on *SUB
D) MDO on *FMT
E) CFC on *SUB

(3) The minimum number of control cards used for any subtest is two.

A) true
B) false

Additional comments on Lertap 5 control cards, of a more introductory nature, may be found in Chapter 4.

List of Control Cards for Cognitive Subtests:

Here is an overview of the control cards which are used for processing results from cognitive subtests:

Card	Required?	Comments (cognitive subtests)
*col	yes	This card does several things; above all, it tells Lertap how many items there are in the subtest, and the columns where these items are found in the Data worksheet.
*sub	no	Not required if the items use A B C D as their response set. Otherwise it's required. Although not strictly required, this card is commonly used to provide a name and title for the subtest.
*key	yes	Tells Lertap the correct answer to each item.
*alt	no	Used when the items do not all have the same number of options. If all items use the same number of options, this card is not required.
*wts	no	Use this card when the correct answer to the subtest's items have different weights, that is, are worth different points.
*mws	no	This is a very special card which is used, for example, when it's necessary to give points for more than one answer to an item.

Now for more detailed comments about these cards, with some examples:

<table>
<tr><td>*col</td><td colspan="2">The definition of every subtest must begin with a *col card. This card tells Lertap where the item responses are in the Data worksheet. It does this by using a format exemplified in these sample *col cards:

*col (c3-c12)
*col (c5, c7, c9, c11-c20)

Here the first example says that item responses start in column 3 and end in column 12. The second example says pick up the first item response from column 5, the second from column 7, the third from column 9, and the rest starting in column 11 and ending in column 20.</td></tr>
<tr><td rowspan="3">*sub</td><td colspan="2">The *sub card is not required with cognitive subtests, but it's recommended, mostly because it provides the chance to add a name and title to the work you're doing. There are several control words which may be used on this card; here's a list:</td></tr>
<tr><td>CFC</td><td>Means "correct for chance". This control word is not used very often. When it's found on a *sub card, subtest scores will be corrected for the chance benefits which might result from informed guessing.</td></tr>
<tr><td>Mastery</td><td>The presence of this word causes Lertap's U-L (upper-lower) item analysis to be based on a cut-off percentage, usually referred to as the "mastery" level. The default level is 70%.</td></tr>
</table>

	Mastery=	Use of the = sign allows the mastery level to be set to a specified value. For example, Mastery=80 would set the mastery level to 80%.
	Name=()	Allows a name to be given to the subtest. The name may be of any length, and may contain any characters except an opening or closing parenthesis. Lertap's subtest name is equivalent to SPSS' variable label. Optional.
	PER	Means "percentage" scoring. Original subtest scores will be reported for each test taker, along with a percentage-of-maximum-possible score. Optional.
	Res=()	This is an important control word. It tells Lertap both the number and nature of response codes used by the items of the subtest. Examples: Res=(A,B,C,D) Res=(A,B,C,D,E,F) Res=(1,2,3,4,5) Res=(u,v,w,x,y,z) If the items of your cognitive subtest use res=(A,B,C,D), you don't have to have an res=() declaration on the *sub card—this response code set is the default for cognitive items. The maximum number of response codes which may be used is 10.
	SCALE	Means "scaled" score. Original scores will be reported for each test taker, along with a scaled score, which, for cognitive subtests, is a z-score. Optional.
	Title=()	Gives a short name, or title, to the subtest. There may be up to 8 characters between the parentheses. Whilst any characters may be used, it is suggested that only letters and digits be employed. For compatibility with SPSS, the title should begin with a letter, and should not contain a space or full stop (period). Lertap's subtest title is the same as SPSS' variable name. Optional.
	Wt=	Assigns a compositing weight to the subtest. By default, Lertap assigns Wt=1 for all subtests. If there is more than one subtest with Wt=1, Lertap forms a Total test score by adding together all subtest scores. To exclude a subtest from the Total, use Wt=0 (zero).
	Examples: *sub res=(1,2,3,4,5,6,7), name=(Hopkins chap 5), title=(Hopkins5) *sub title=(Ed503), name=(Ed 503 quiz), wt=0 Here the second example does not have an res=() declaration, and Lertap will use its default assignment for cognitive subtests, which is res=(A,B,C,D).	
*key	This is a required card—every cognitive subtest ***must*** have a *key card	

	which indicates the keyed-correct answer for each item. Example: *key BCBDD AADCA There must be one keyed-correct answer for every item. There must be a space before the first keyed-correct answer, but after that spaces are optional.
*alt	This control card is used when not all items use all of the response codes found in a subtest's res=() declaration. For example, if res=(A,B,C,D,E), and the following card is used *alt 44444 55555 then Lertap will know that the first five items of the subtest use only the first 4 response codes, while the last five items use all 5. Optional.
*wgs	The keyed-correct answer to a cognitive item usually gets 1 (one) point. To give more points a *wgs card may be used. For example, the following card indicates that three items, the second, sixth, and tenth are to have 2 points given for their keyed-correct answer. *wgs 12111 21112 This card is optional. If only one or two item are to have scoring changes of this type, *mws cards may be easier to use.
*mws	The "multiple-weights specification" card is used to change the response weights for a designated item. Its use is optional. As an example, if a subtest is using response weights of res=(A,B,C,D), and the following *mws card is used *mws c3, 0, 2, 0, 0 then the weights for the item whose responses are found in column 3 of the Data worksheet will be zero (0) for all but the second response, which, in this case, is "B", as defined by the res=() declaration. The following card will give 1 (one) point if a student selects either the first or third answer for the item whose responses are found in column 30 of the Data worksheet: *mws c30, 1, 0, 1, 0 The weights found on the *mws card do not have to be integers: *mws c17, 0.00, 0.25, 0.50, 0.75, 1.00 this card applies to the item whose responses are found in column 17 of the Data worksheet. For this item, the first response is to have a weight

	of 0.00, the second a weight of 0.25, the third a weight of 0.50, and so forth. *mws Call, 1, 0, 0, .5 this card's "Call" means all columns, that is, all items which belong to the respective subtest. There are many countries which use a decimal separator different to the full stop (or period). Users in these countries are required to express decimal values as shown here, with the full stop, but Lertap will convert them correctly.

Example sets

Below we've included some real-life examples of sets of control cards for cognitive subtests.

Set 1:

*col (c28-c37)
*key ABBDC DDACA

There are 10 cognitive items in this subtest. As there is no *sub card with an res=() declaration, Lertap will assign res=(A,B,C,D), the default for cognitive subtests. Since there is no *alt card, all 10 items will be assumed to use all four response codes. And, since there is no *wgs card, the correct answers for the items will get 1 (one) point each.

What will be the Name and Title of this subtest? Again, there is no *sub card, so Lertap will set Name=(Test1), and Title=(Test1).

What would be the minimum possible score on this 10-item subtest? Zero; if a student gets all items wrong, zilch is the resultant subtest score. On the other hand, the maximum possible score will be 10, a "perfect" score on this 10-item subtest.

Set 2:

*col (c28-c37)
*sub name=(Class quiz of 25 July), title=(Quiz25J)
*key ABBDC DDACA

A *sub card has been added here in order to give a name and title to the output produced by Lertap. The name will appear at the top of various item statistics pages, such as Stats1f and Stats1b while the title will be used to label subtest scores.

Set 3:

*col (c28-c37)
*sub res=(A,B,C,D,E,F), title=(Quiz25J)
*key AEBDC DDACF
*alt 35444 55356

This example includes an res=() declaration on the *sub card, indicating that items use as many as six response codes. The *alt card adds some precision to the scene, telling Lertap that only one item, the last, uses all 6 response codes. Four items, the second, sixth, seventh, and ninth use (A,B,C,D,E) as response codes. Two items, the first and the eighth, use just the first 3 responses codes, (A,B,C).

In this example, the subtest has been given a title, but not a name. In such cases Lertap will assign a name which is identical to the title.

Set 4:

```
*col (c28-c37)
*sub r=(A,B,C,D,E,F), n=(Class quiz, 25 July), t=(Quiz25J), per
*key AEBDC DDACF
*alt 35444 55356
*wgs 21111 21113
```

The addition of a *wgs card tells Lertap to give 2 points for the correct answers to the first and sixth items, while a whopping 3 points will go to those who get the last item correct. The maximum possible score on this 10-item subtest is 14.

The "per" on the *sub card tells Lertap to add a percent-correct score for this subtest. This score will appear next to the original subtest score (often called the "raw" score) on the Scores worksheet produced by the program.

Notice how some of the control words on the *sub card have been abbreviated? This is permitted, as mentioned in Chapter 4.

Set 5:

```
*col (c28-c37)
*sub res=(A,B,C,D,E,F), name=(Class quiz of 25 July), title=(Quiz25J)
*key AEBDC DDACF
*alt 35444 55356
*mws c28, 2, 0, 0
*mws c37, 0, 0.50, 0, 0.50, 0, 0
```

A couple of *mws cards are included in this example. The item whose responses are found in column 28 of the Data worksheet is to be scored by giving 2 points to the first answer (or response), which is "A", and zero points to the other two permitted responses.

Why are there only another two permitted answers to this item? Because the *alt card indicates that the first item, which corresponds to that in column 28, uses just the first 3 response codes.

Meanwhile, the item whose responses are found in column 37 of the Data worksheet now has two keyed-correct answers, "B", and "D". A student will get half a point if s/he selects either of these answers.

Set 6:

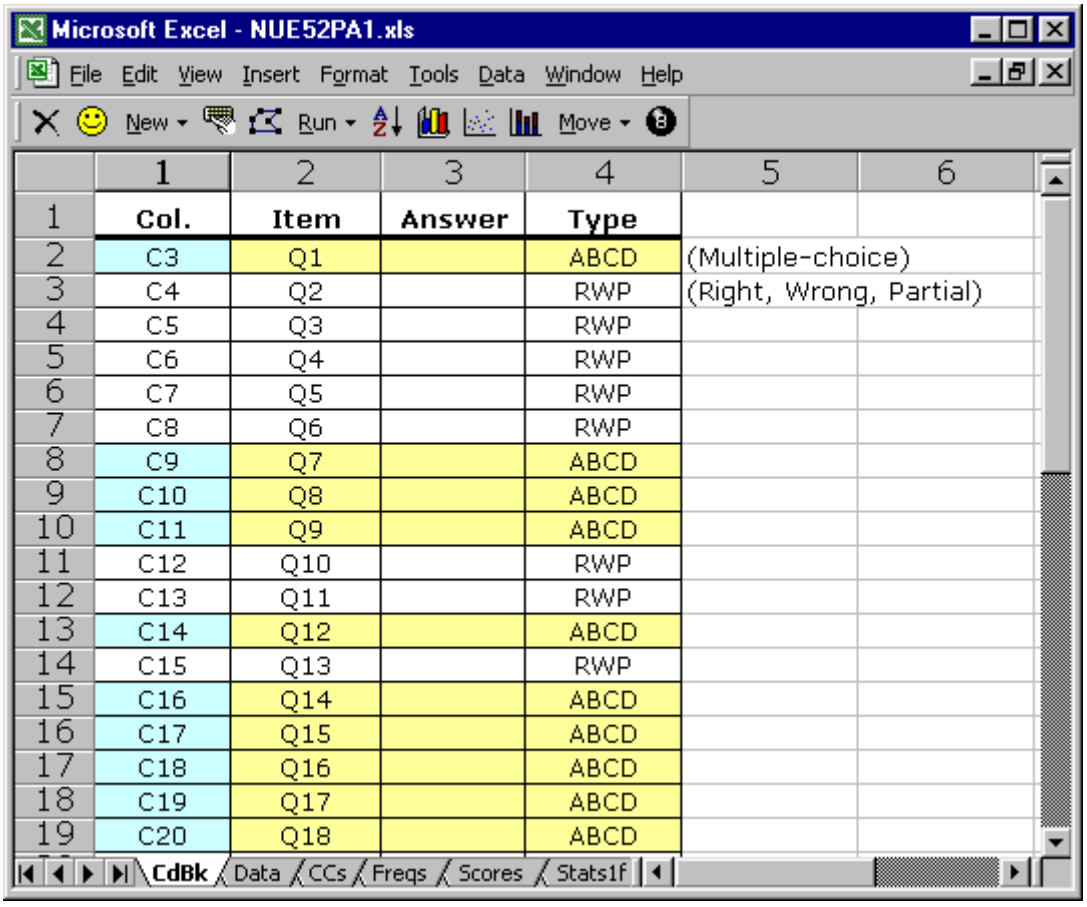
```
*col (c28-c37)
*sub name=(Class quiz of 25 July), title=(Quiz25J)
*key ABBDC DDACA
*col (c28-c37)
*sub cfc, name=(CFC class quiz of 25 July), title=(CFCQuiz)
*key ABBDC DDACA
```

Two subtests are defined by these six control cards. Notice that the *col cards point to the same columns—here a subtest of 10 items is to be scored twice, once in "normal" fashion, and once with the CFC scoring option applied.

Set 7:

```
Data from one of TAFE's applied diploma classes (Sept 2000).
&
*col (c3,c9-c11,c14,c16-c20,c24,c25,c28,c30-c32,c37,c38,c41)
*sub mastery=60, title=(NUE52mc), per
*key DBCDD DCBAD CADAC ADCC
*mws c16, 0, 1, 0, 1
&
*col (c4-c8,c12-c13,c15,c21-c23,c26-c27,c29,c33-c36,c39-c40,c42-c44)
*sub mastery=60, res=(R,P,W), title=(NUE52sa), per
*key RRRRR RRRRR RRRRR RRRRR RRR
*mws Call, 1.0, 0.5, 0.0
&
*col (c3-c44)
*sub mastery=60, title=(total), res=(A,B,C,D,R,P,W), wt=0
*key DRRRR RBCDR RDRDC BADRR RCARR DRACA RRRRD CRRCR RR
```

This is <u>not</u> a straightforward, easy-to-understand job. An instructor has used a test with 42 items, of which 19 were multiple choice, and 23 were short answer. The item types were mixed—as shown in the codebook below, the first item was multiple-choice, the next five were short answer, the next three were multiple-choice, and so on.



Microsoft Excel - NUE52PA1.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

	1	2	3	4	5	6
1	**Col.**	**Item**	**Answer**	**Type**		
2	C3	Q1		ABCD	(Multiple-choice)	
3	C4	Q2		RWP	(Right, Wrong, Partial)	
4	C5	Q3		RWP		
5	C6	Q4		RWP		
6	C7	Q5		RWP		
7	C8	Q6		RWP		
8	C9	Q7		ABCD		
9	C10	Q8		ABCD		
10	C11	Q9		ABCD		
11	C12	Q10		RWP		
12	C13	Q11		RWP		
13	C14	Q12		ABCD		
14	C15	Q13		RWP		
15	C16	Q14		ABCD		
16	C17	Q15		ABCD		
17	C18	Q16		ABCD		
18	C19	Q17		ABCD		
19	C20	Q18		ABCD		

CdBk Data CCs Freqs Scores Stats1f

The first control card in this example has no asterisk the beginning, and there are three other cards which have the & character at the start. These lines will be ignored by Lertap. Chapter 4 pointed out that lines with no asterisks at the very beginning may be used as comments, and as separators between subtests.

The multiple-choice items all used res=(A,B,C,D), that is, each of the multiple-choice items presented options A, B, C, and D to students. This is called Lertap's default res=() assignment for cognitive tests, and, whenever this is the case, there is no need to use res=() on the *sub card.

The short answer items, on the other hand, were marked Right, Wrong, or Partial Credit, with letters of R, W, and P entered in Lertap's Data worksheet to denote each possible mark.

Now—note the two *mws cards. The first one tells Lertap that the item whose answers were coded in column 16 of the Data sheet is to be scored so that both the second and fourth options, which would be B and D, get 1 (one) point, while the first and third options, A and C, get zero points.

The second *mws card uses "Call", which tells Lertap that it's referring to all the items used by the subtest (Call means "columns all"). For these items, a

response of R will get 1 (one point), while P will get half a point (0.5), and W will get zero points.

Finally, the third subtest will not work well as seen above. It's meant to be a total test, one comprised of all 42 items.

The reason this subtest scoring will not work well is because are no *mws cards following the subtest's *key card, meaning that item Q14 (which belongs to column 16, as indicated in the codebook) will not be double-keyed as it was in the first subtest. The lack of *mws cards also means that marks of P (Partial Credit) on the short-answer items will not get the half-point they got in the second subtest.

When we asked the teacher who developed this test why she had not used *mws cards in the third subtest, she said she was just experimenting, and then proceeded to ask us if it was true she'd have to enter 24 *mws cards for the third subtest if she wanted to do the job right.

And yes, this would be the case. She has one item in the multiple-choice set which is to be double-keyed, and 23 items in the short-answer set which have one of their responses, P, getting half a point. If she wanted to have a third subtest with correct item scoring, she'd end up with something like the next example.

Set 8:

```
Data from one of TAFE's applied diploma classes (Sept 2000).
&
*col (c3,c9-c11,c14,c16-c20,c24,c25,c28,c30-c32,c37,c38,c41)
*sub mastery=60, title=(NUE52mc), per
*key DBCDD DCBAD CADAC ADCC
*mws c16, 0, 1, 0, 1
&
*col (c4-c8,c12-c13,c15,c21-c23,c26-c27,c29,c33-c36,c39-c40,c42-c44)
*sub mastery=60, res=(R,P,W), title=(NUE52sa), per
*key RRRRR RRRRR RRRRR RRRRR RRR
*mws Call, 1.0, 0.5, 0.0
&
*col (c3-c44)
*sub mastery=60, title=(total), res=(A,B,C,D,R,P,W), per
*key DRRRR RBCDR RDRDC BADRR RCARR DRACA RRRRD CRRCR RR
*mws c16, 0, 1, 0, 1
*mws c4, 1, 0.5, 0
*mws c5, 1, 0.5, 0
*mws c6, 1, 0.5, 0
*mws c7, 1, 0.5, 0
*mws c8, 1, 0.5, 0
*mws c12, 1, 0.5, 0
*mws c13, 1, 0.5, 0
*mws c15, 1, 0.5, 0
 ... ... ...
*mws c44, 1, 0.5, 0
```

This example is in answer to the instructor whose control cards were shown in Set 7 above. Now there are multiple *mws cards—in fact, we haven't shown all of them (there would be 24 for the third subtest, but we've shown only 10).

We hear some readers asking why an *mws card with "Call" could not have been used for the third subtest, as it was in the second. One reason, the most compelling one, is that this subtest has two different sets of items, multiple-choice and short-answer, and they use different response codes. Another reason has to do with the fact that the scoring pattern for the items in the first subtest is not uniform over all subtest items—sometimes the first response, A, gets one point, sometimes it's the third response, C (and so on). Note that this is not the case in the second subtest, where R always gets one point, P always gets half a point, and W is always a loser.

If you find it difficult to follow the examples shown in the last two control card sets above, worry not. These are complex examples. Lertap is capable of scoring test items in just about any manner imaginable, but things can get a bit hairy in special cases.

Peeking at Sub worksheets

If you're not sure how Lertap will interpret your control cards, use the Run option on the toolbar to "Interpret CCs lines", and then look at the subtest's Sub worksheet.

Sub worksheets are normally hidden from view. To unhide them, use Excel's Format / Sheet option. Sub sheets are not spectacularly formatted, but you will probably be able to understand most of their contents.

The *tst card

There is another card which may be used with any subtest, including cognitive ones. The *tst card is used to break out certain data records from the Data worksheet, after which Lertap's Run options are used to get results for these records only.

For example,

```
*tst c12=(3)
```

will have Lertap make a new Excel workbook containing only those cases in the Data worksheet which have a 3 in column 12. Once this workbook is created, all options on Lertap's toolbar are available for use, including, of course, the Run options.

There is more information on the use of the *tst card in Chapter 4.

Chapter 6

Control Cards for Affective Subtests

Contents

In this chapter we discuss the various Lertap control "cards" used for processing results from affective tests and/or surveys. The term affective refers to an instrument which is meant to measure attitudes and opinions. For Lertap to work effectively, the items (or questions) used by the instrument must use fixed-choice responses. Here are some sample affective items:

(7) The city of Perth is:

cold 1 2 3 4 5 6 7 hot

(31) I will recommend to others that they use LERTAP.

1) strongly disagree
2) disagree
3) undecided
4) agree
5) strongly agree

Additional comments on Lertap 5 control cards of a more introductory nature may be found in Chapter 4.

List of Control Cards for Affective Subtests:

Here is an overview of the control cards which are used for processing results from affective subtests:

Card	Required?	Comments (affective subtests)
*col	yes	This card does several things; above all, it tells Lertap how many items there are in the subtest, and the columns where these items are found in the Data worksheet.
*sub	yes	This card is required of affective subtests, and must have the "Aff" control word. Otherwise Lertap will assume a cognitive subtest. This card is also used to give the subtest a name, and a title. If the subtest's items do not use 1 2 3 4 5 as their responses, then this card must carry an Res=(….) specification.
*alt	no	Used when the items do not all have the same number of options. If all items use the same number of options, this card is not required.
*****pol	no	If the "best" answer to an affective item is not always the first answer, or is not always the last answer, this card is used to indicate where the best answer is. Use a + sign if the best answer is last; use a – sign otherwise
*mws	no	This card is not used as much as it is for cognitive subtests; it's used to specify unique response weights.

More detailed comments about these cards, with some examples:

<table>
<tr><td>*col</td><td colspan="2">The definition of every subtest must begin with a *col card. This card tells Lertap where the item responses are in the Data worksheet. It does this by using a format exemplified in these sample *col cards:

*col (c3-c12)
*col (c5, c7, c9, c11-c20)

Here the first example says that item responses start in column 3 and end in column 12. The second example says pick up the first item response from column 5, the second from column 7, the third from column 9, and the rest starting in column 11 and ending in column 20.</td></tr>
<tr><td>*sub</td><td colspan="2">Affective subtests must make use of the *sub control card. There are several control words which may be used on this card, and one of them, “aff”, has to be present in order for Lertap to know that the subtest is an affective one. Here’s a list of the control words which may be used on the *sub card:</td></tr>
<tr><td></td><td>Aff</td><td>Means “affective”. This control word must be present.</td></tr>
<tr><td></td><td>MDO</td><td>Means “missing data assignment off”. When this control word is present, Lertap’s default missing data weights are turned off. If for example Res=(1,2,3,4,5), a weight of 3, the centre of the weights, is ordinarily assigned to missing data. If MDO is present this automatic assignment is extinguished, and missing data will receive a weight of zero.</td></tr>
<tr><td></td><td>Name=()</td><td>Allows a name to be given to the subtest. The name may be of any length, and may contain any characters except an opening or closing parenthesis. Lertap’s subtest name is equivalent to SPSS’ variable label. Optional.</td></tr>
<tr><td></td><td>PER</td><td>Means “percentage” scoring. Original subtest scores will be reported for each test taker, along with a percentage-of-maximum-possible score. Optional.</td></tr>
<tr><td></td><td>Res=()</td><td>This is an important control word. It tells Lertap both the number and nature of response codes used by the items of the subtest. Examples:

Res=(1,2,3,4,5)
Res=(A,B,C,D,E)
Res=(1,2,3,4,5,6,7)
Res=(u,v,w,x,y,z)

If the items of your subtest use res=(1,2,3,4,5), you don’t have to have an res=() declaration on the *sub card—this response code set is the default for affective items. The maximum number of response codes which may be used is 10.</td></tr>
</table>

<table>
<tr><td></td><td>SCALE</td><td>Means "scaled" score. Original scores will be reported for each test taker, along with a scaled score equal to the original score divided by the number of items in the subtest. Such a score is sometimes known as a "normalised" score. Optional.</td></tr>
<tr><td></td><td>Title=()</td><td>Gives a short name, or title, to the subtest. There may be up to 8 characters between the parentheses. Whilst any characters may be used, it is suggested that only letters and digits be employed. For compatibility with SPSS, the title should begin with a letter, and should not contain a space or full stop (period). Lertap's subtest title is the same as SPSS' variable name. Optional.</td></tr>
<tr><td></td><td>Wt=</td><td>Assigns a compositing weight to the subtest. By default, Lertap assigns Wt=1 for all subtests. If there is more than one subtest with Wt=1, Lertap forms a Total test score by adding together all subtest scores. To exclude a subtest from the Total, use Wt=0 (zero).</td></tr>
<tr><td></td><td colspan="2">Examples:

*sub aff, res=(1,2,3,4,5,6,7), name=(Perth Q1), title=(Perth1)
*sub aff, mdo, title=(Ed503), name=(Ed 503 feedback), wt=0

Here the second example does not have an res=() declaration, and Lertap will use its default assignment for affective subtests, which is res=(1,2,3,4,5).</td></tr>
<tr><td>*alt</td><td colspan="2">This control card is used when not all items use all of the response codes found in a subtest's res=() declaration. For example, if res=(1,2,3,4,5,6,7), and the following card is used

*alt 55555 77777

then Lertap will know that the first five items of the subtest use only the first 5 response codes, while the last five items use all 7. Optional.</td></tr>
<tr><td>*pol</td><td colspan="2">This control card is used to tell Lertap to reverse the scoring for some of the subtest's items. By default, an affective item is weighted in a forward manner, with the first response getting a weight of 1 point, the second a weight of 2 points, the third a weight of 3 points, and so on. To reverse this pattern, that is, to give the highest weight to the first response, and the lowest weight to the last response, use a *pol card with + and – signs.

*pol +-+-+ -+-+-

This card reverses the weights for every other item. The *pol card is optional—it's required only when one or more items are to have their weights reversed.</td></tr>
<tr><td>*mws</td><td colspan="2">The "multiple-weights specification" card is used to change the response weights for a designated item. For example, if a subtest is using response weights of res=(1,2,3,4,5), and the following *mws card is used</td></tr>
</table>

	*mws c3, 5, 4, 3, 2, 1 then the weights for the item whose responses are found in column 3 of the Data worksheet will be 5 for the first response, 4 for the second response, 3 for the third response, 2 for the fourth response, and 1 for the last (fifth) response. This would be the same as reversing the item's weights on a *pol card. If only a few items in a subtest are to have their weights reversed, then using *mws cards, one for each item, is sometimes easier than using a *pol card. The weights found on the *mws card do not have to be integers: *mws c17, 0.00, 0.25, 0.50, 0.75, 1.00 this card applies to the item whose responses are found in column 17 of the Data worksheet. For this item, the first response is to have a weight of 0.00, the second a weight of 0.25, the third a weight of 0.50, and so forth. *mws Call, -3, -2, -1, 0, 1, 2, 3 this card's "Call" means all columns, that is, all items which belong to the respective subtest. There are seven (7) weights shown in this *mws specification—the Call word means that all items belonging to the subtest will use these weights. There are many countries which use a decimal separator different to the full stop (or period). Users in these countries are required to express decimal values as shown here, with the full stop, but Lertap will convert them correctly.

Example sets

Here are some examples of sets of control cards for affective subtests.

Set 1:

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

There are 10 affective items in this subtest. As the *sub card does not carry an res=() declaration, Lertap will assign res=(1,2,3,4,5), the default for affective subtests. Since there is no *alt card, all 10 items will be assumed to use all five response codes. And, since there is no *pol card, the items will all be forward weighted, that is, a response of "1", the first response in the res=() list, will get a weight of 1 point; the second response, "2", will get 2 points, …, the last response, "5", will get 5 points.

What would be the minimum possible score on this 10-item subtest? Ten. The minimum possible "score" for each item is one; on the other hand, the maximum possible score on an *item* is 5, so the maximum possible *subtest* score is 50. If a

person does not answer an item, a score of 3 points will apply since the missing data option has not been turned off (there is no MDO control word seen on the *sub card). If MDO had been used, then the minimum possible *item* "score" would be zero—in turn, this would make the minimum possible *subtest* score zero as well.

Set 2:

*col (c28-c37)
*sub Aff, Title=(Comfort), Name=(User satisfaction), Wt=0
```
*pol +---- ++--+
```

This example adds a *pol card. There's still no explicit res=() declaration on the *sub card, so res=(1,2,3,4,5) will be applied by default. Six of the 10 items are to be reverse scored—the *pol card directs Lertap to reverse the scoring for the second, third, fourth, fifth, eighth, and ninth items. For these items, a response of "1", the first response code found in the res=() declaration, will get a score of 5 points; the second response, "2", will get a score, or weight, of 4 points, and so on, down to the last possible response, "5", which will have a weight, or score, of just 1 (one) point. (There's a bit more about using *pol cards later.)

In this example a Wt= assignment appears on the *sub card, which will mean that scores from this subtest will not be added to those from other subtests to make a Total score. What other subtests, you might well ask? We can't tell—we'd have to assume that the three cards shown above were not the only cards in the CCs worksheet. If there's only one subtest, a Wt= assignment doesn't make sense, and will be ignored by Lertap.

Set 3:

*col (c28-c37)
*sub Aff, Title=(Comfort), Name=(User satisfaction), Wt=0
```
*mws c29, 5, 4, 3, 2, 1
*mws c30, 5, 4, 3, 2, 1
*mws c31, 5, 4, 3, 2, 1
*mws c32, 5, 4, 3, 2, 1
*mws c35, 5, 4, 3, 2, 1
*mws c36, 5, 4, 3, 2, 1
```

This example amounts to the same thing as the previous example. Now (for some reason), *mws cards have been used to reverse the scoring for six items.

Set 4:

*col (c28-c37)
*sub Aff, Title=(Comfort), Name=(User satisfaction), Wt=0
```
*pol +---- ++--+
*mws c32, 1, 2, 3, 4, 5
```

In this example the *pol card reverses the item scoring for six items, including the item whose responses are found in column 32 of the Data worksheet. However, the *mws card gets Lertap to change the weights for the item whose

responses are in column 32, and effectively re-weights this item in the standard forward manner. When used, *mws cards override whatever has come before.

Set 5:

*col (c28-c37)
*sub Aff, Res=(1,2,3,4,5,6,7), T=(Comfort), N=(User satisfaction)

```
*alt 55555 77777
*pol +---- ++--+
```

It's a bit unusual to see a *alt card used with affective subtests, but we've got one here, to be sure. It says that the first five items uses the first 5 responses seen in the res=() declaration, while the last five use all 7 response codes. What's the minimum possible score on this 10-item subtest? Ten. There are 10 items, and, since there's no MDO word on the *sub card, the minimum possible item score is 1 (one) for all 10 items.

What's the maximum possible subtest score? Sixty (60). There are five items whose maximum possible score is 5 points each, and there are another 5 items with a maximum possible score of 7. Five times 5, plus five times 7, gives 60.

Chapter 4 mentioned that the control words used on the *sub card, such as Name and Title, could be dramatically abbreviated, if wanted. The person who typed up these control cards took advantage of this possibility.

Set 6:

*col (c28-c37)
*sub Aff, Res=(1,2,3,4,5,6,7), T=(Comfort), N=(User satisfaction)

```
*mws Call, -3, -2, -1, 0, 1, 2, 3
```

The "Call" statement on this example's *mws card says that all items in the subtest are to have their scoring changed so that a response of "1" gets –3 points, "2" gets –2 points, "3" gets –1 point, "4" gets zero points, "5" gets 1 point, "6" gets 2 points, and "7" gets 3 points. (There are a few affective instruments which score responses in this manner.)

Set 7:

```
Lertap syntax, 21 July 2543.
Part1
*col (c4 - c23)
*sub aff, name=(CAQ 5.22, Part 1), title=(Part1)
*pol +-+++ +++++ ++--- ---+-
Part2
*col (c24 - c38)
*sub aff, name=(CAQ 5.22, Part 2), title=(Part2)
Part3
*col (c39 - c48)
*sub aff, name=(CAQ 5.22, Part 3), title=(Part3)
Part4
*col (c49 - c61)
*sub aff, name=(CAQ 5.22, Part 4), title=(Part4)
Part6
*col (c80 - c90)
*sub aff, name=(CAQ 5.22, Part 6), title=(Part6)
Part7
*col (c91 - c100)
*sub aff, name=(CAQ 5.22, Part 7), title=(Part7)
*pol +-+-+ ---++
```

The cards above are from a research project which Dr Nanta Palitawanont of Burapha University, Thailand, conducted in July of 2543 (year 2000 in the Gregorian calendar). She was using version 5.22 of the Computer Attitude Questionnaire from the University of North Texas to collect information from students and staff. Above you can see how Dr Palitawanont defined various subtest, or Part, scores, with some of the Parts having items which had to be reverse scored.

In this example there are some lines which do not begin with an asterisk (do not have a * at the beginning). Such lines are ignored by Lertap—they provide the chance to add comments to the CCs worksheet. Take some care, however: a line which is totally blank is interpreted by Lertap as the last line to be processed—if there are more lines in the CCs sheet after the blank line, they will not be read even if they have the magic asterisk as their first character.

Chapter 4 has another example of a job which involved processing multiple affective subtests.

More about the *pol card

This small section is provided for those readers who might want more clues about the use of the *pol card.

Consider these three items, taken from the Lertap Quiz data set (Nelson, 1974):

(28) I have used item analysis programs superior to LERTAP.

1) strongly disagree
2) disagree
3) have not used such programs before
4) agree
5) strongly agree

(31) I will recommend to others that they use LERTAP.

1) strongly disagree
2) disagree
3) undecided
4) agree
5) strongly agree

(32) I don't think I could design my own LERTAP analysis.

1) strongly disagree
2) disagree
3) undecided
4) agree
5) strongly agree

These three items are affective ones. They all use the same response codes, which is the set {12345}. They do not have a right answer, but they might be said to have a "best" answer, that is, an answer which reflects the most positive response to the statement posed. For Item 28 the "best" answer is the first, for Item 31 it's the last, while for Item 32 it's the first.

If we wanted to add the responses to these items together so as to indicate whether or not people had a positive outlook to their Lertap experience, we'd want to first reverse the scoring of Items 28 and 32. On both of these items, the best response, the most positive response, is "strongly disagree".

If we did this, then someone with a total score of 15 over these three items will have answered "strongly disagree" to Item 28, "strongly agree" to Item 31, and "strongly disagree" to Item 32.

In Lertap, the way to accomplish this sort of scoring, wherein some items are reverse scored, is via the *pol card. *pol -+- is what would do the job in this three-item example.

What did people do before they had Lertap and its *pol card? Some of them used to reverse items as they were entered into the data set. If someone answered "1" on Item 28, a "5" would be entered in the data set. Others, SPSS users, would create a new variable using a statement similar to this:

Item28R = 6 – Item28

As you have now seen, the *pol makes these shenanigans unnecessary.

Peeking at Sub worksheets

If you're not sure how Lertap will interpret your control cards, use the Run option on the toolbar to "Interpret CCs lines", and then look at the subtest's Sub worksheet.

Sub worksheets are normally hidden from view. To unhide them, use Excel's Format / Sheet option. Sub sheets are not spectacularly formatted, but you will probably be able to understand most of their contents.

The *tst card

There is another card which may be used with any subtest, including affective ones. The *tst card is used to break out certain data records from the Data worksheet, after which Lertap's Run options are used to get results for these records only.

For example,

*tst c12=(1,2), c13=(F)

will have Lertap make a new Excel workbook containing only those cases in the Data worksheet which have a 1 or a 2 in column 12, and an F in column 13. Once this workbook is created, all options on Lertap's toolbar are available for use, including, of course, the Run options.

There is more information on the use of the *tst card in Chapter 4.

Chapter 7

Interpreting Lertap Results for Cognitive Tests

Contents

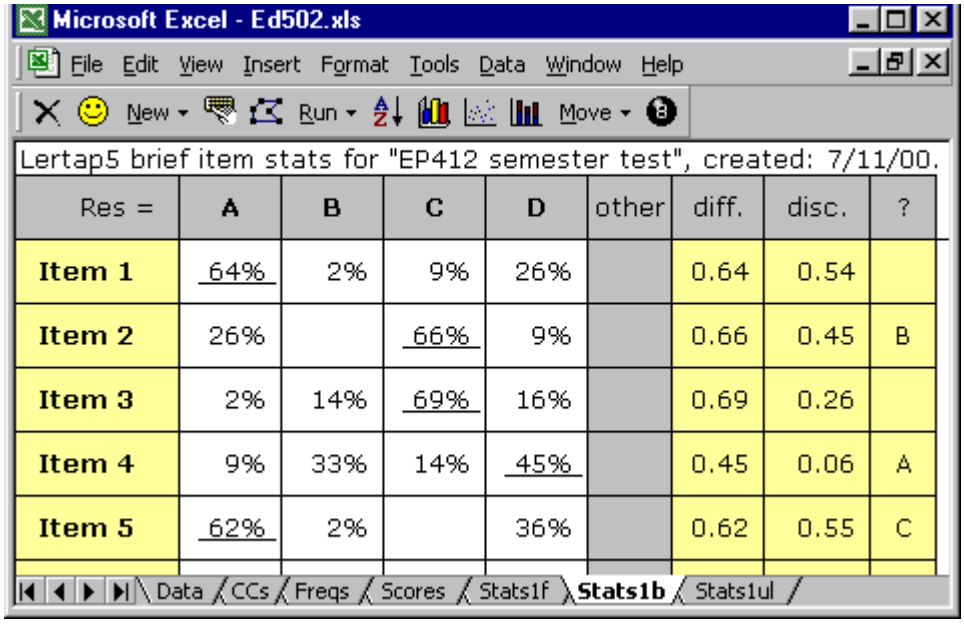
Well, there you are, and you've done it. You've digested all the technical, nitty-gritty material found in previous chapters, prepared and processed your Data and CCs sheets, used "Interpret CCs lines" and "Elmillon item analysis", and found that Lertap has produced its results sheets, worksheets with 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[18]. 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

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

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.

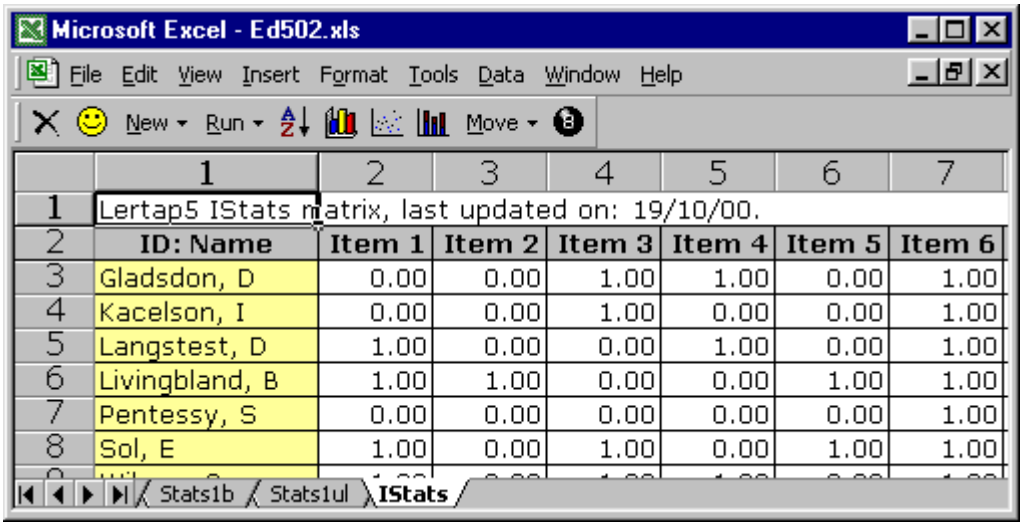
If all seems in order, he'll then review the topics corresponding to the items which students did poorly on, and go over them again with the class.

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.

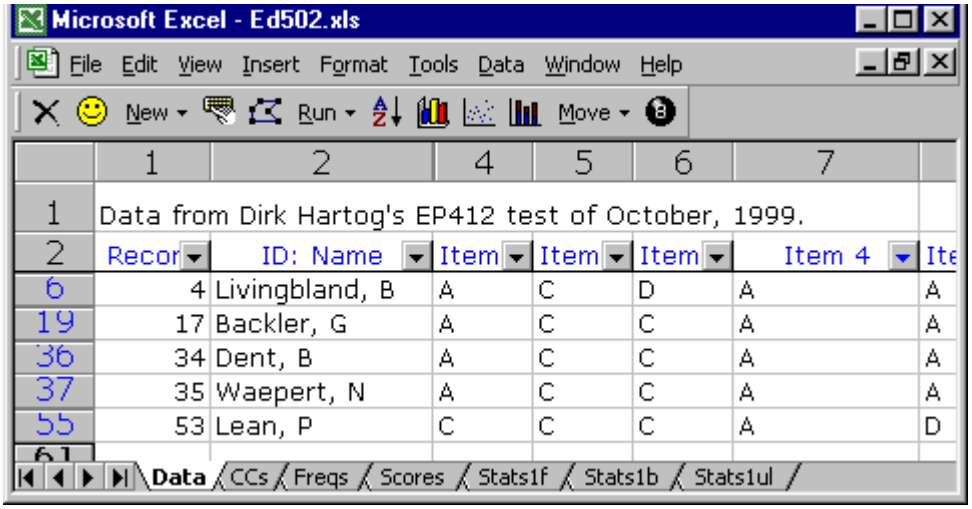
To get results for each student at an item level, the "Output item scores matrix" capability on Lertap's Run menu is useful. It produces a report such as the following:



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.

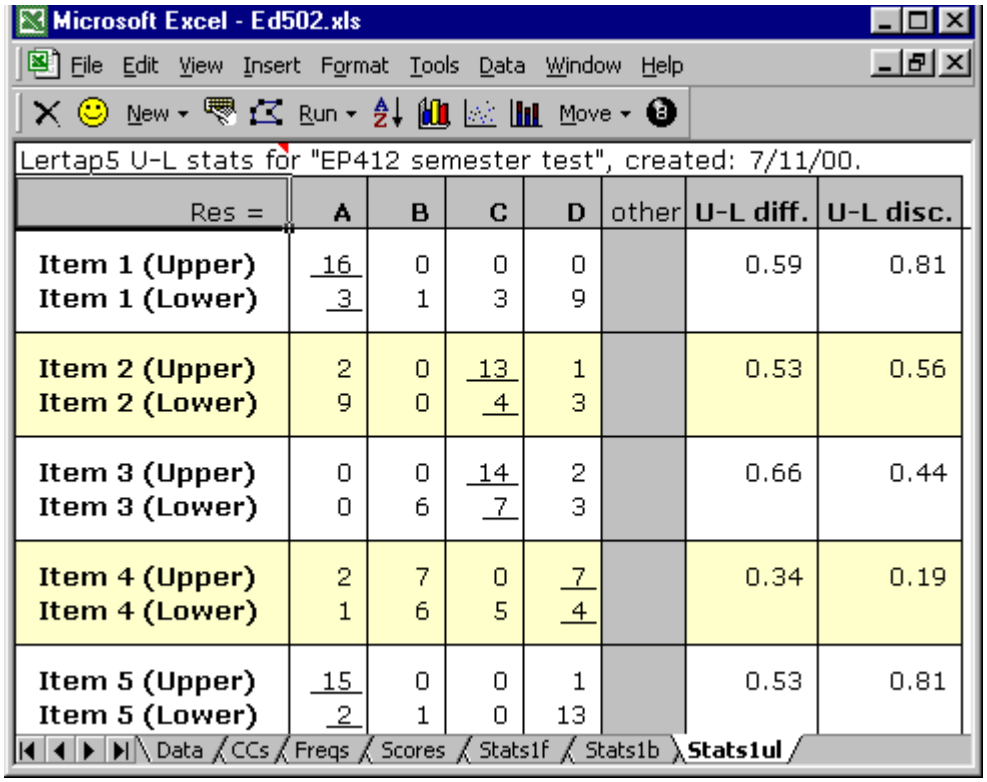
But there are many times when instructors wish to do more with Lertap's results.

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:



Lertap5 U-L stats for "EP412 semester test", created: 7/11/00.

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			

Data | CCs | Freqs | Scores | Stats1f | Stats1b | **Stats1ul**

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[19]. 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

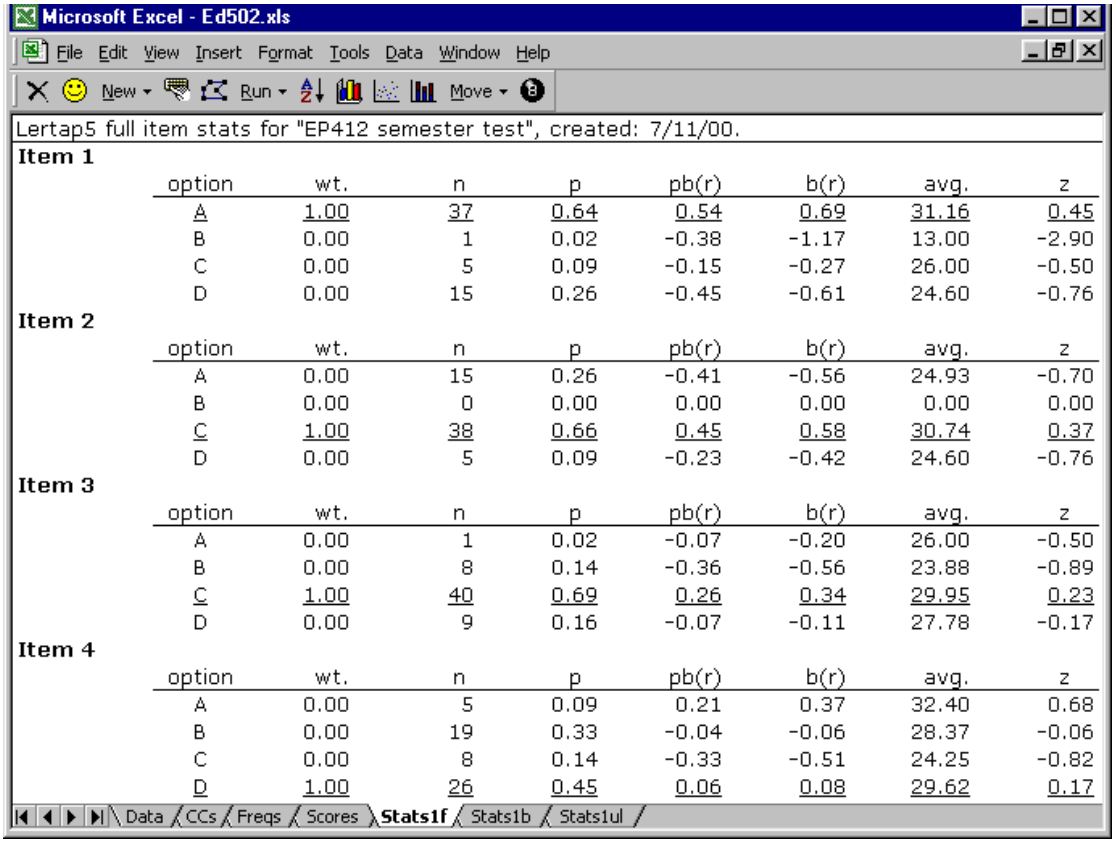
[19] Ebel & Frisbie, and most other texts, call the U-L item discrimination index "D".

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.

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)[20].

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

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.

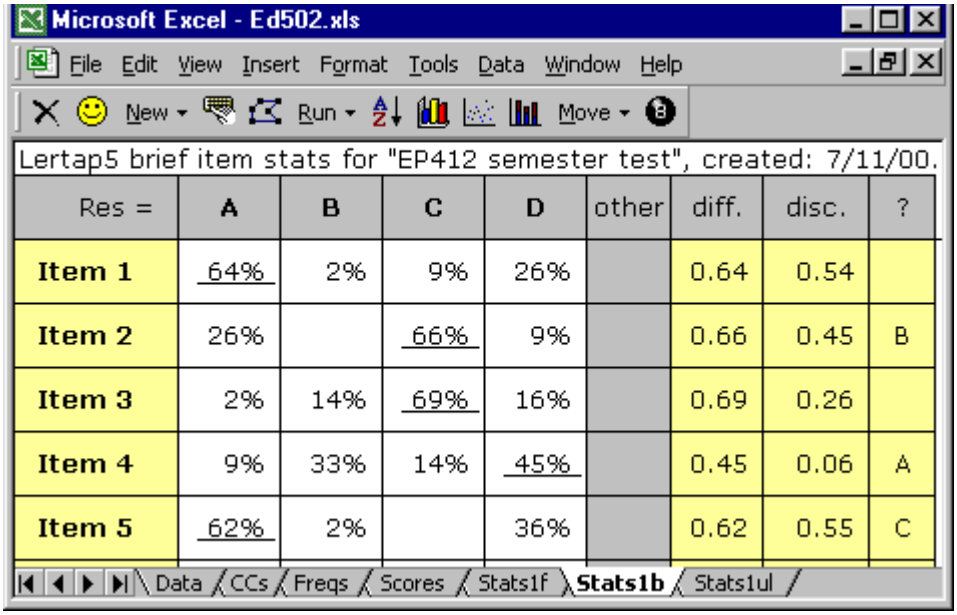
If an item is discriminating, helping to identify who's strong on the criterion measure and who's not, the pb(r) and b(r) figures for the correct answer will be high, while corresponding figures for the distractors will be negative. The avg. value for the right answer will be higher than the avg. values for the distractors, 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[21]. 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

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

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.

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.

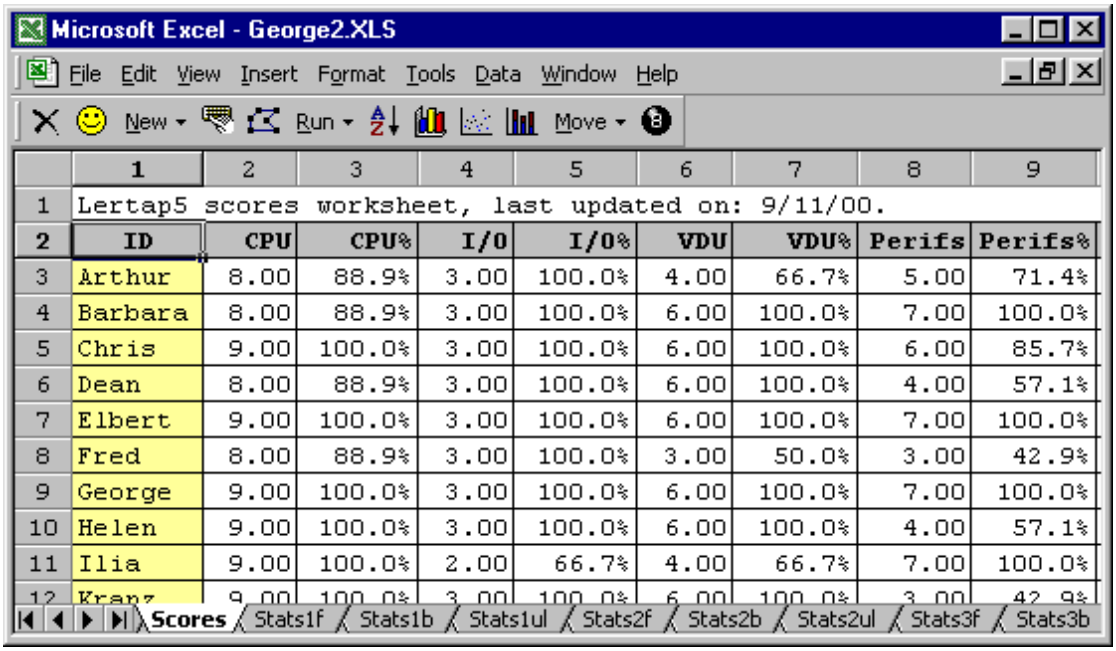
To begin, CRT users will want to have Lertap report scores in terms of percentage correct, as opposed to number of items right. This is easy to do: as mentioned in Chapter 5, using the control word "PER" on a *sub card will see to it that Lertap computes and displays percentage scores. For example, the 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[22]:

```
*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:



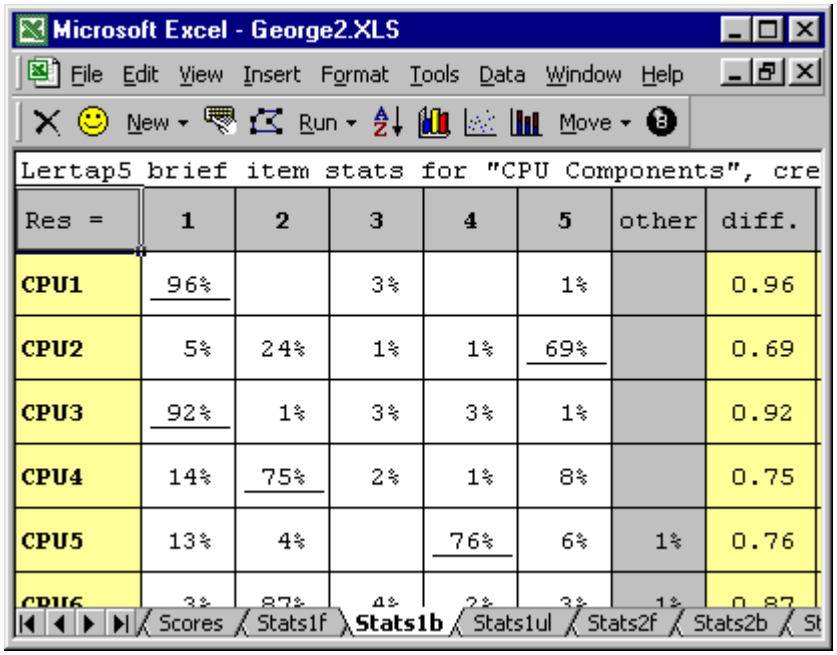
	1	2	3	4	5	6	7	8	9
1	Lertap5 scores worksheet, last updated on: 9/11/00.								
2	ID	CPU	CPU%	I/O	I/O%	VDU	VDU%	Perifs	Perifs%
3	Arthur	8.00	88.9%	3.00	100.0%	4.00	66.7%	5.00	71.4%
4	Barbara	8.00	88.9%	3.00	100.0%	6.00	100.0%	7.00	100.0%
5	Chris	9.00	100.0%	3.00	100.0%	6.00	100.0%	6.00	85.7%
6	Dean	8.00	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.

[22] 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:

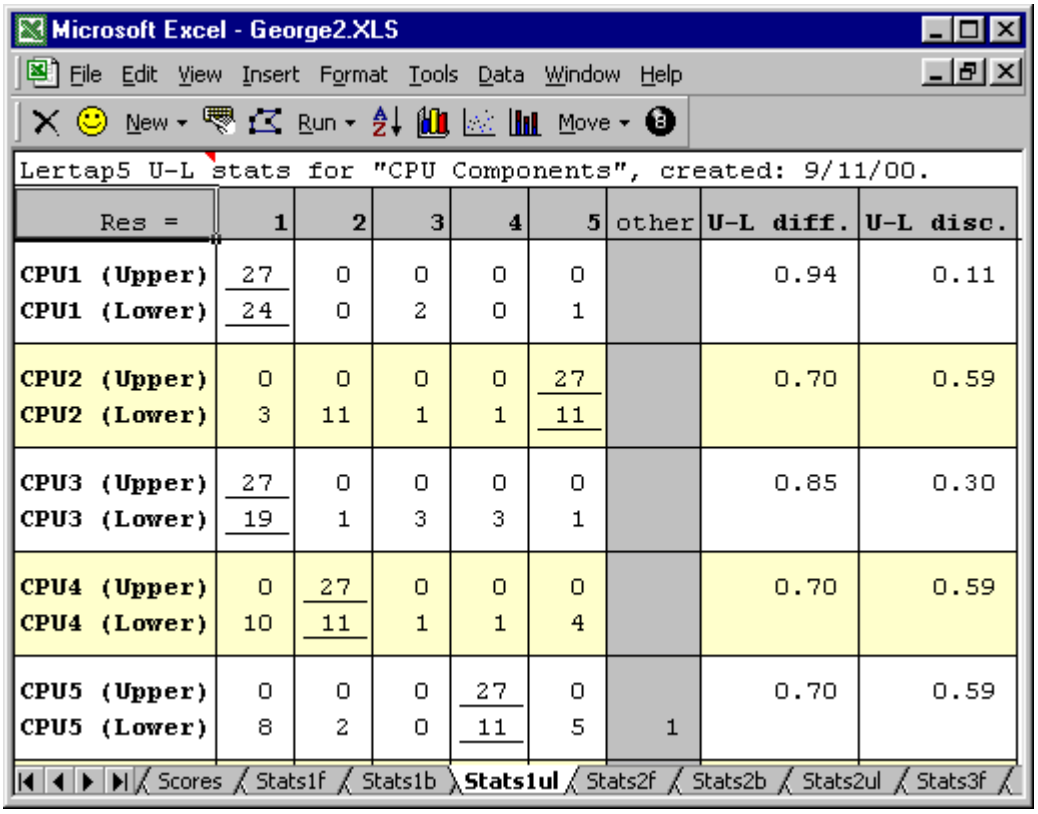


Lertap5 brief item stats for "CPU Components", cre

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

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?[23])

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.

[23] 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.

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:

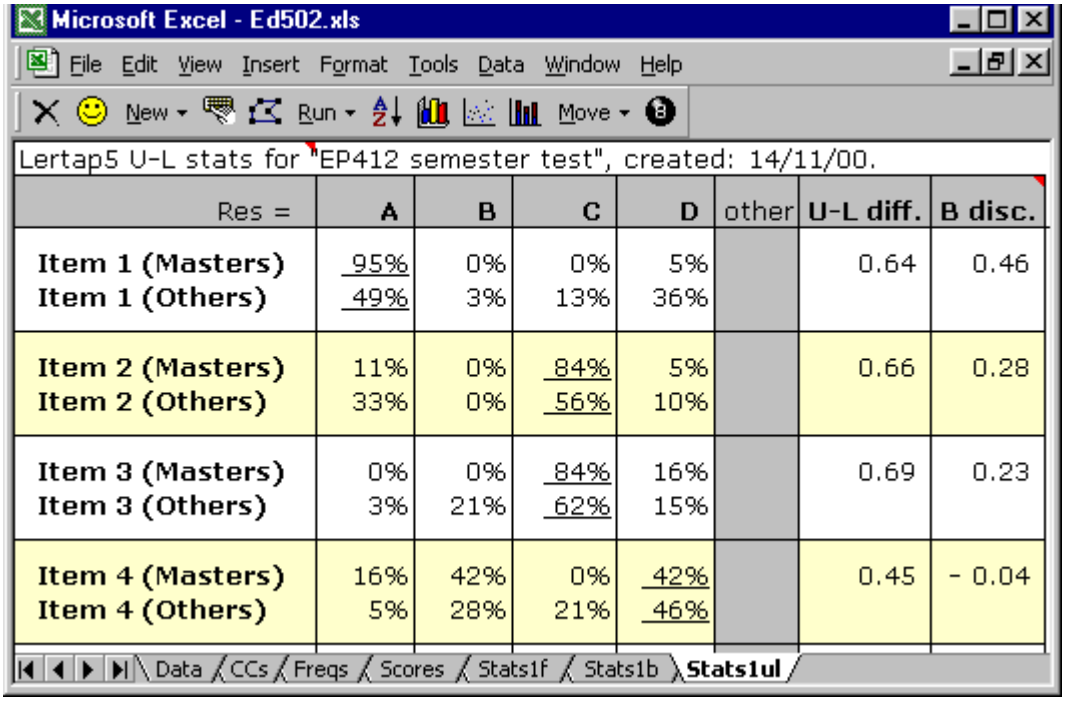
```
From EP 412 class of Semester 2, 1999.
*COL (C4-C48)
*SUB Name=(412 semester test), Title=(EP412), Mastery
*KEY ACCDA CCDBB DBCDB ADDAC BDDCC BDCBA BCCBD CBCCC ACDCA
```

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.

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):



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[24].

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.

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).

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

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[25], 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 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,

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

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[26] 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.

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....".

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

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 <u>not</u> free from error; in this regard, Lertap provides the figures needed to form confidence intervals, score ranges which reflect the imprecision of our tests.

Chapter 8

Interpreting Lertap Results for Affective Tests

Contents

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 used Lertap to process results from surveys as often as they have to process cognitive test results.

There are, of course, other systems useful for processing surveys. Of them, the SPSS statistical package[27] 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*)."

---

[27] www.spss.com

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 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 names[28].

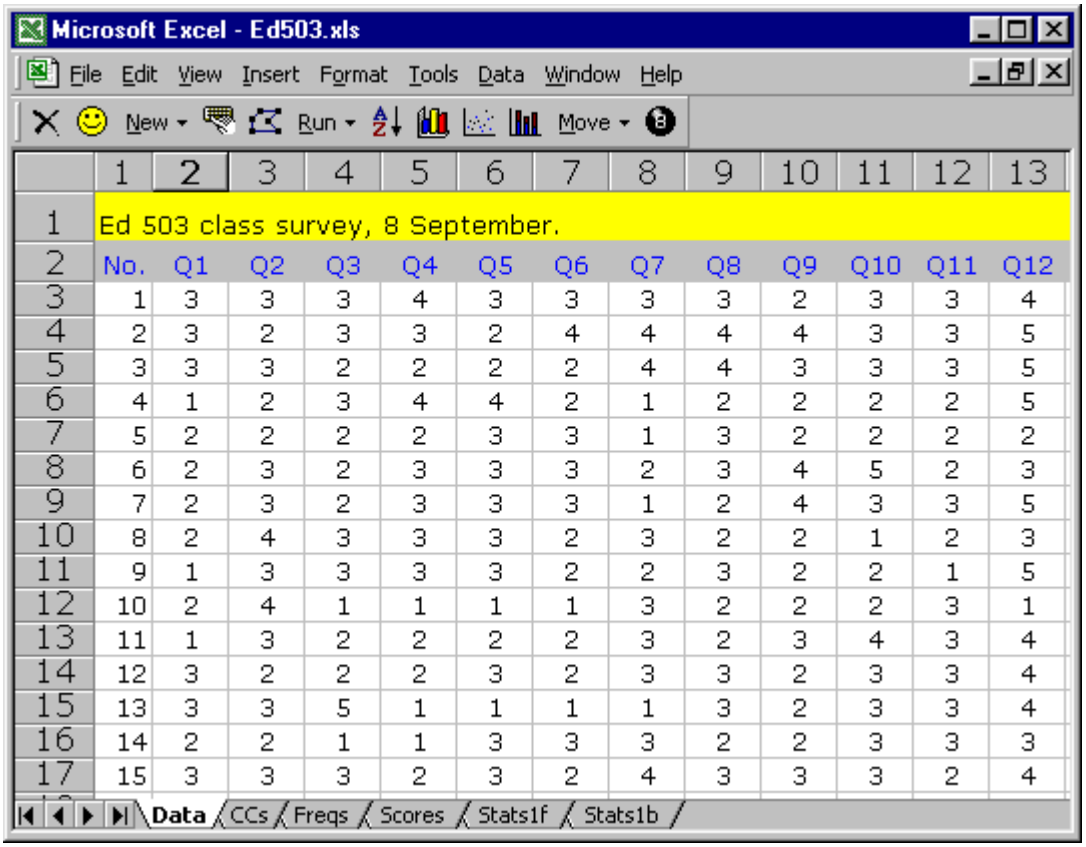
1.*The amount of work I did for this unit was*						
very great	1	2	3	4	5	quite small
2.*The quality of my work for this unit was*						
excellent	1	2	3	4	5	poor
3.*I learned from this unit*						
very much	1	2	3	4	5	very little
4.*The skills learned during the unit will be*						
very useful	1	2	3	4	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*						

[28] The survey was used at Curtin University of Technology. What are called "courses" in North America, and "papers" in New Zealand, are referred to as "units" at Curtin.

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
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 below:



Microsoft Excel - Ed503.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

	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

Data CCs Freqs Scores Stats1f Stats1b

The initial CCs worksheet had just these two entries:

```
*Col (C2-C13)
*Sub Affective
```

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

The Stats1b report gives a brief summary of item responses, and tosses in a few item statistics as well:

Microsoft Excel - Ed503.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 brief item stats for "Test1", created: 14/11/00.

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

Data CCs Freqs Scores Stats1f **Stats1b**

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).

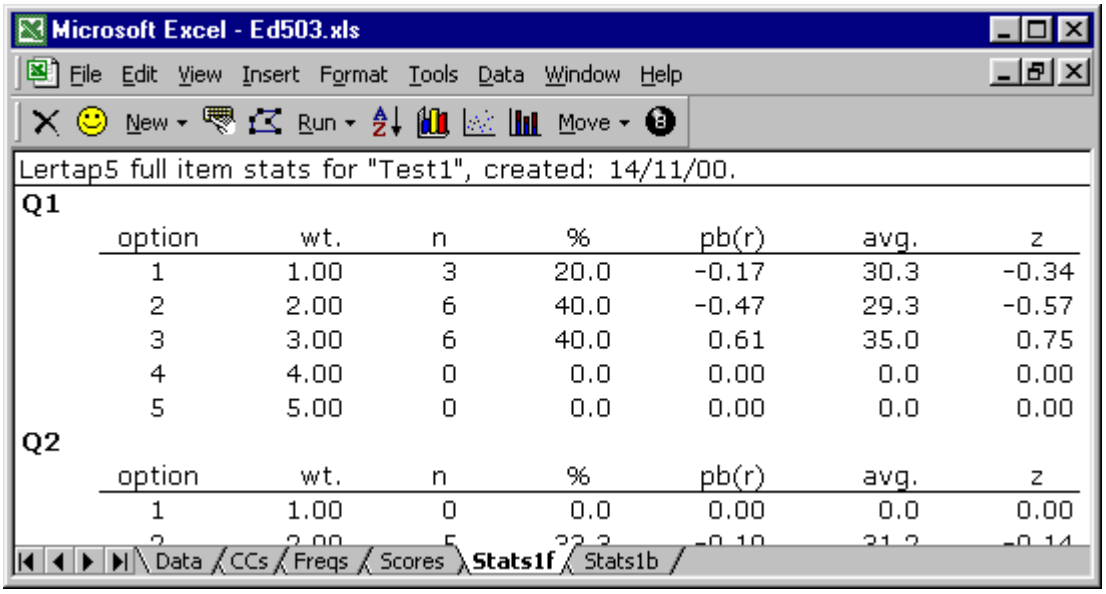
The columns to the right of "other" indicate the type of scoring applied to each item (the pol. column[29]), the mean of the responses, their standard deviation, and the degree to which responses on each item correlate with the sum of the responses on the other items[30]. These columns are not always used in interpreting results. In this case, the 12 items were <u>not</u> 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 <u>not</u> 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.

[29] "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.
[30] 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:



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 example[31].

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.

[31] See Chapter 10 for more discussion of the Stats1f reports for affective subtests.

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 Lertap 2, a system which appeared in 1973:

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.xls file. If you have Lertap running on your computer, you have a copy of the data.

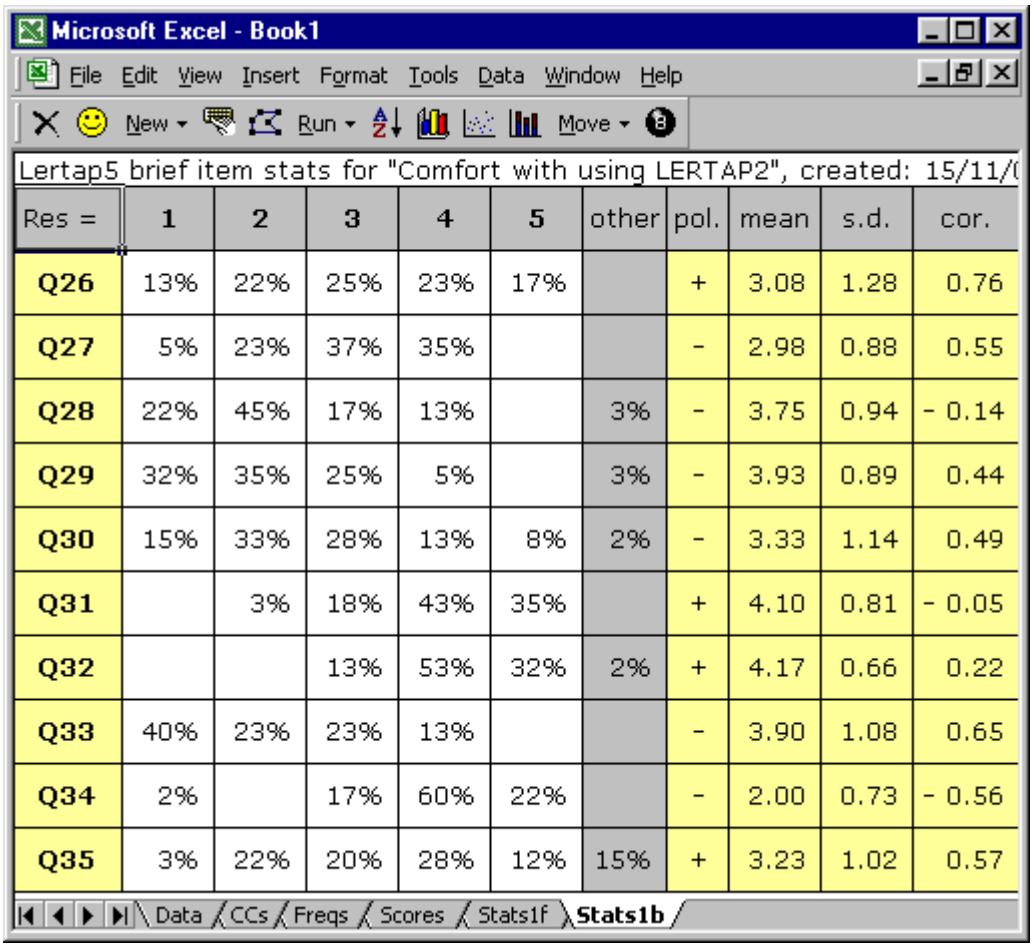
Answers to these 10 items were processed by using five digits: 1 for SD, through to 5 for SA. To get survey results, Nelson used these CCs lines:

```
*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. Nelson 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 points[32].

Here's the brief stats report for this survey:



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

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

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.

[32] The ease with which Lertap allows items to be reverse-scored is a feature not currently found in some other systems, SPSS 10 among them.

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.

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

without	*alpha*	*change*
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 / nonmastery 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 0.65 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:

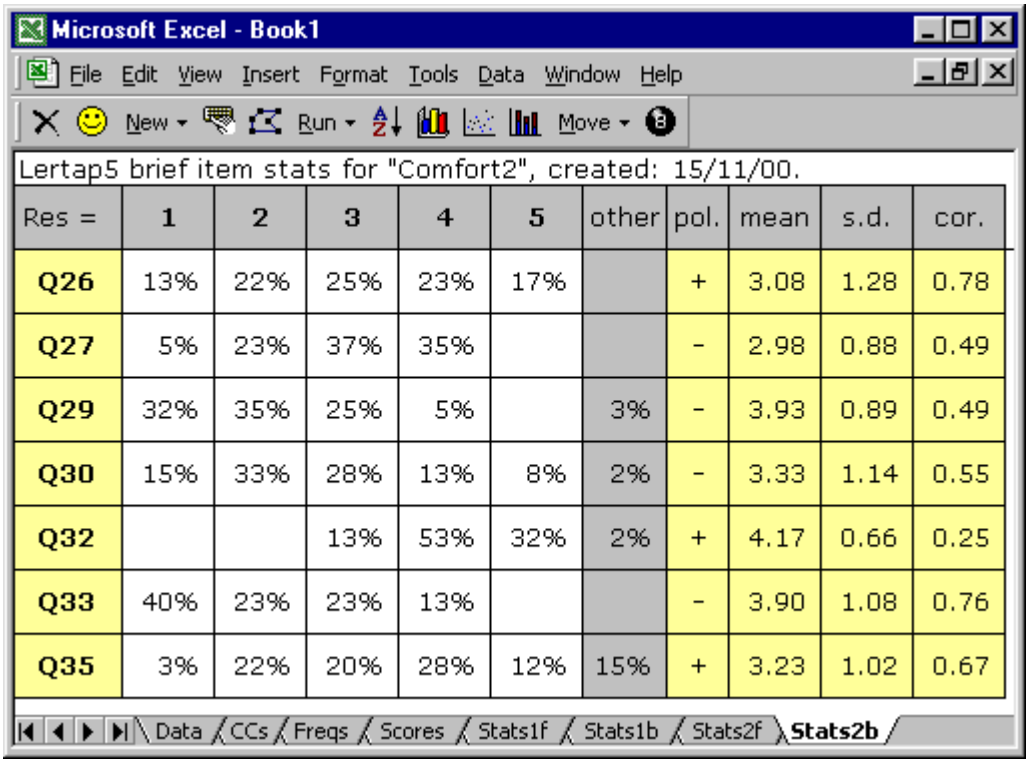
```
*col (c28-c37)
*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort)
*pol +---- ++--+
*col (c28-c37)
*sub Aff, Name=(Comfort2), Title=(Comfort2)
*pol +---- ++--+
*mws c30, *
*mws c33, *
*mws c36, *
```

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)[33].

[33] There would have been other ways to remove the items from the subtest. For example, *col (c28-c29,c31-c32,c34-c35,c37) would have accomplished the same thing, with *pol then becoming +---+-+.

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



Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

New Run Move

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

Data CCs Freqs Scores Stats1f Stats1b Stats2f **Stats2b**

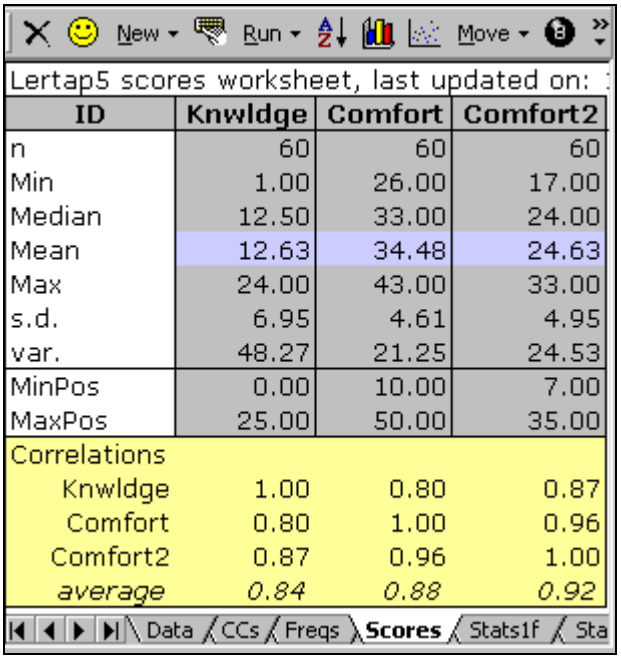
This is about as good as it gets. The seven items in the new subtest have very good correlations, all comfortably positive. The Stats2f report indicated that coefficient alpha was 0.83, a nice increase over the value of 0.63 found when all ten items were included.

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:

```
*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:

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

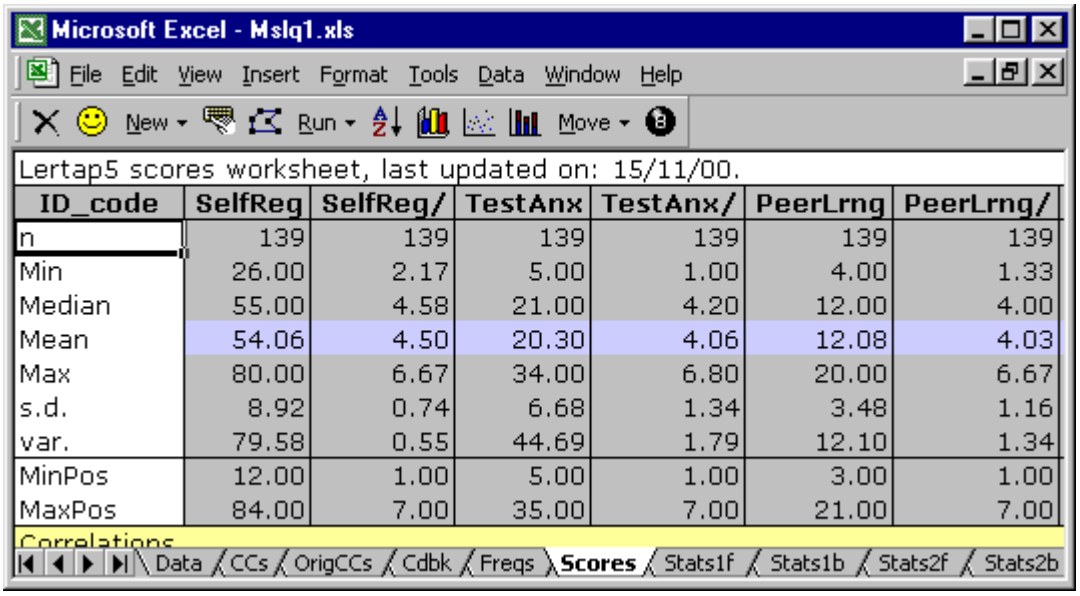
The Lertap control cards for four of the ten subtests, the MSLQ scales, are shown below. Note that ampersands (&s) have been used to separate the subtests—this is not necessary, but it makes it a bit easier to see where each subtest's specifications begin.

```
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 21[34]. 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.

[34] 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.



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

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. These values would be considered reasonable, particularly when we take into account the small number of items in some of the scales.

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 provide. 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 a bit 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 of Technology surveyed a sample of their first-year students, searching for an indication of the extent of their satisfaction with a new outcomes-focused program.

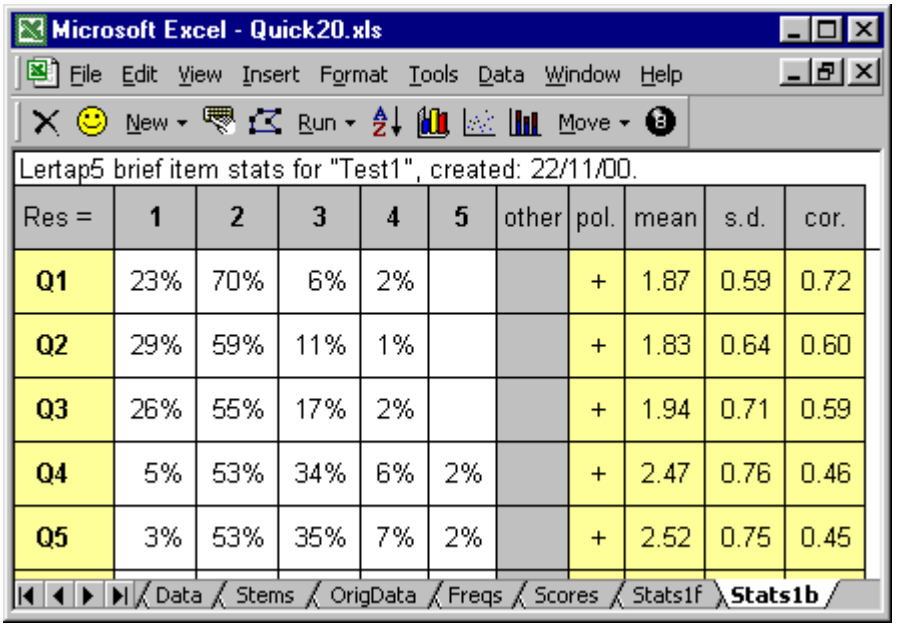
Twenty questions were answered by students in three groups: early childhood education (ECE), primary education (Pri), and secondary education (Sec). The questions had to do with how effectively the new program had helped them "learn about how to become a competent teacher"; "understand the role of a 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, 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:



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 subgroups

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 worksheet[35].

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 options. 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 option used to "Interpret CCs lines" and apply the "Elmillon 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):

[35] There is more about the *tst card in Chapters 4, 5, and 6.

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

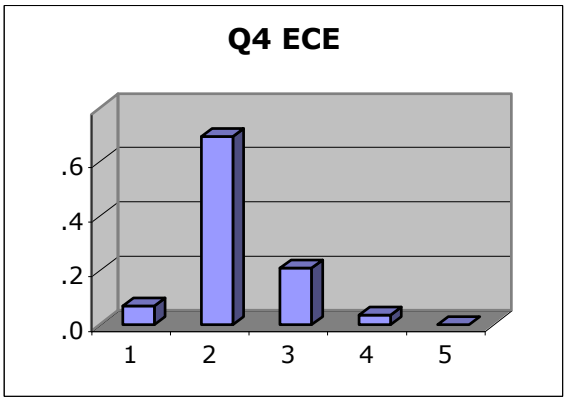
New Run Move

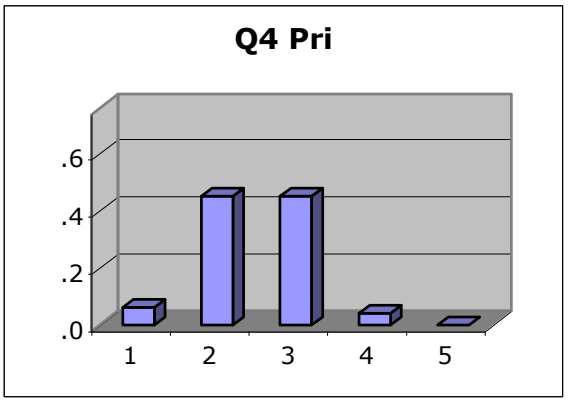
Lertap5 brief item stats for "Test1", created: 22/11/00.

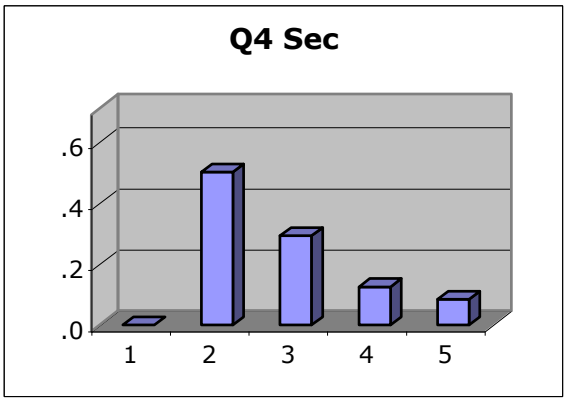
Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1	25%	67%	8%				+	1.83	0.55	0.64
Q2	46%	50%	4%				+	1.58	0.57	0.49
Q3	29%	58%	13%				+	1.83	0.62	0.48
Q4		50%	29%	13%	8%		+	2.79	0.96	0.58
Q5		50%	29%	13%	8%		+	2.79	0.96	0.52

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 option). 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.

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 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.

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.

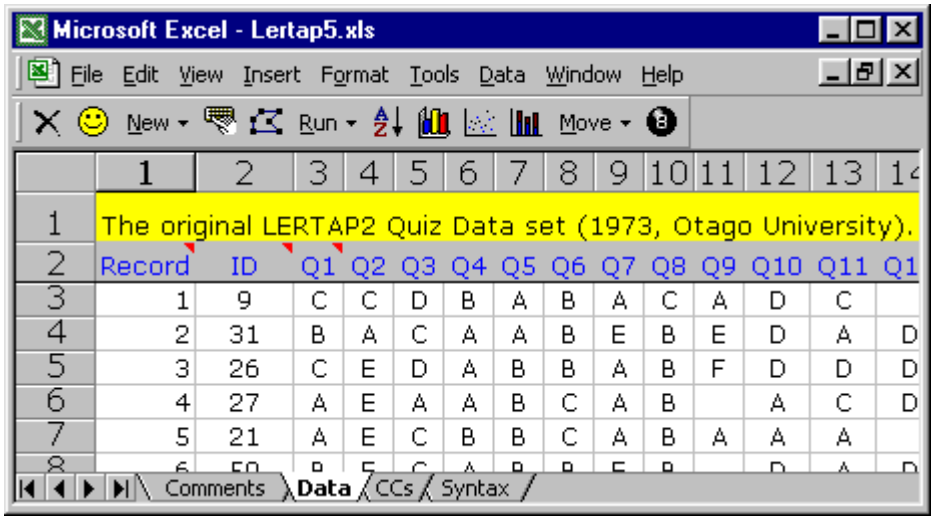
We'll walk you through two examples.

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.xls file.

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.

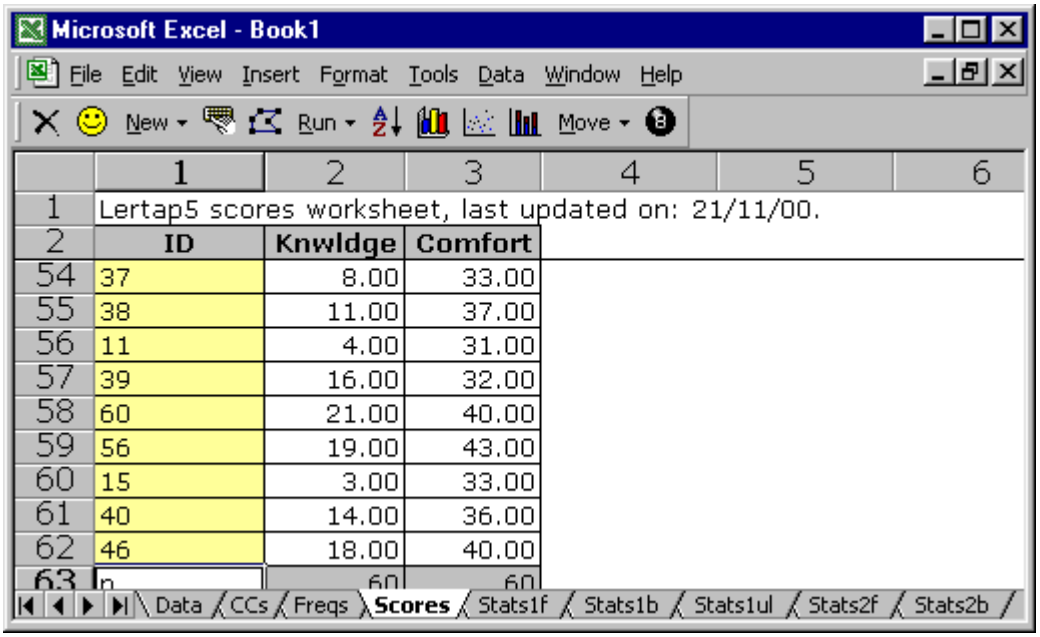
We want to do two things: (1) look at how responses to each of the ten affective questions correlated with a person's score on the 25-item cognitive test, and then (2), how responses to the same ten affective items correlated with Q37, denoted as "YrsTest" in the data set. Q37 asked respondents to indicate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

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



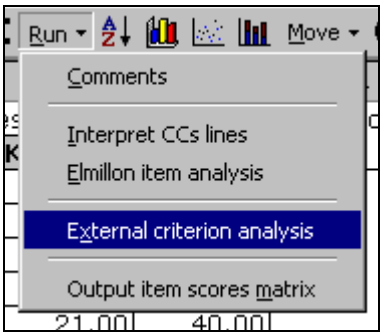
You'd use Lertap's New option to "Make a new Lertap workbook which is a copy of the present one". Then you'd access the Run options to "Interpret the CCs lines", and to get an "Elmillon item analysis".

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



The workbook above is displaying 9 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, click on Run, and then click on "External criterion analysis".

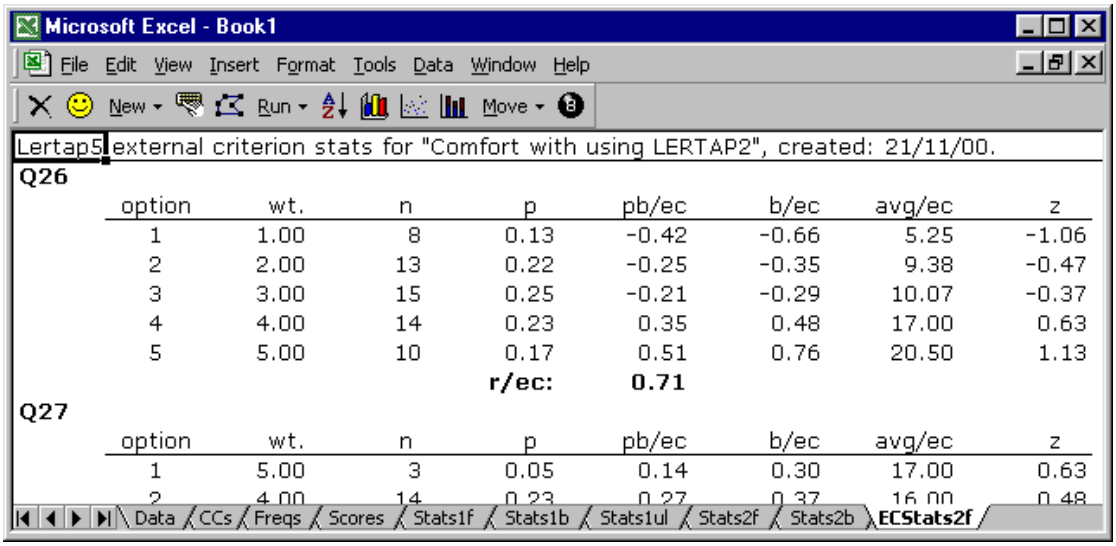


Lertap will ask for the "column number of the score which will serve as the external criterion". It's referring to the columns in the Scores worksheet. In this case, the column with the scores to use as the external criterion is the second column—we want to use the Knwldge score as the external criterion.

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.

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:



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)

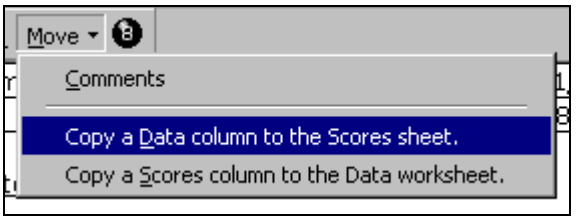
.00: Q28 Q31 Q34
.10:
.20:
.30: Q27
.40: Q32
.50:
.60: Q29 Q30
.70: Q26 Q33 Q35
.80:
.90:

Making a move for another external criterion

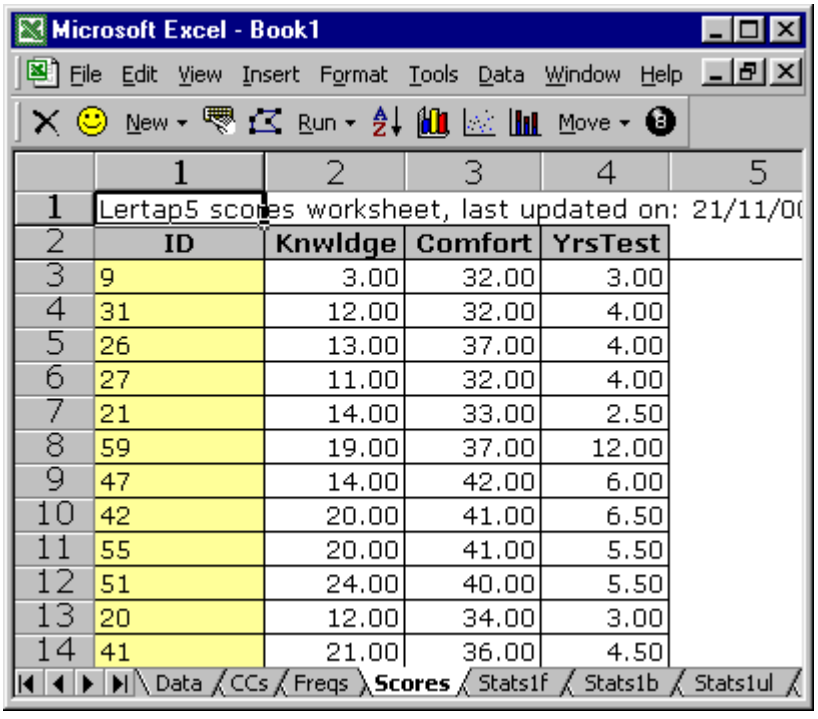
We posed two questions above. We've found the correlations of the affective items with Knwldge, the cognitive subtest, finding five items with high correlations: Q26, Q29, Q30, Q33, and Q35. But we also wanted the same correlations with another criterion, Q37, a question which had to do with the number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

If we use the Run options 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:

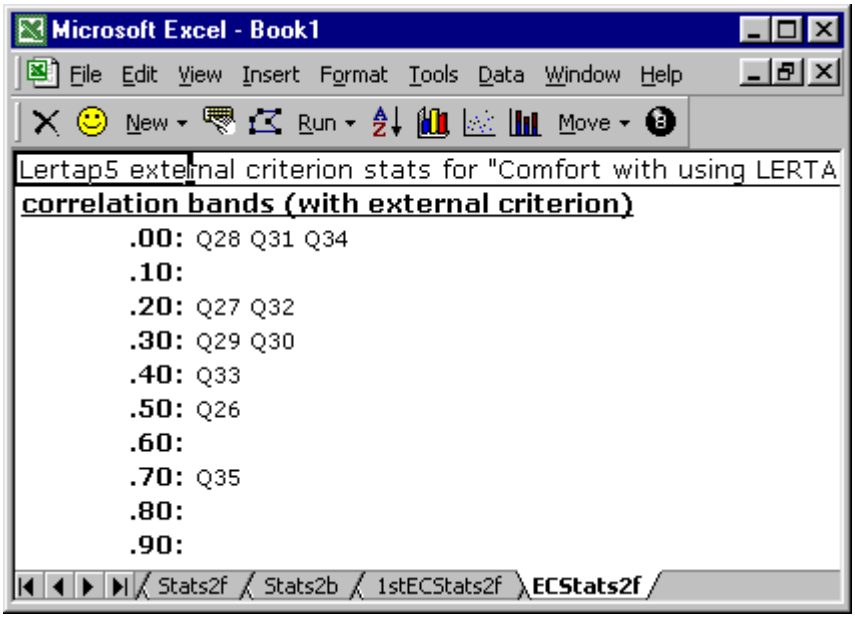


The use of this option is straightforward. We tell Lertap that column 39 of the Data sheet has the scores to be copied and pasted. Lertap checks to make sure that the column seems to have valid data, and then makes the copy, giving us an amended Scores worksheet:



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).

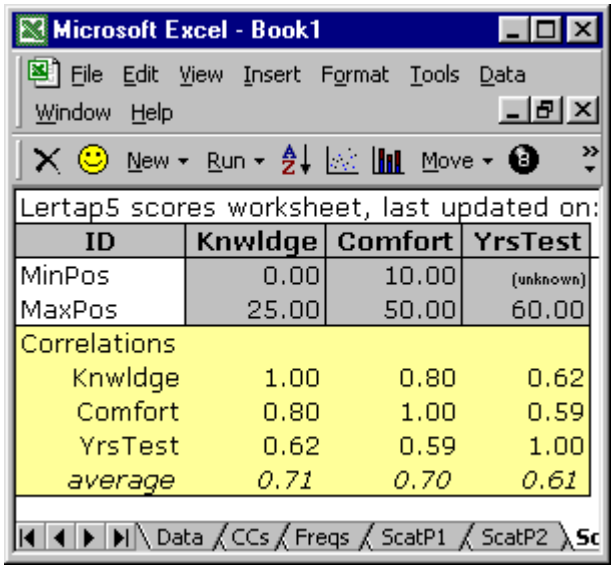
The next hurdle: Lertap announces that this subtest already has an external criterion report. And it's right. It does. The ECStats2f sheet from our first external criterion analysis is still there. We need to rename this worksheet, or delete it, so that Lertap can create a new one. Once we've done one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



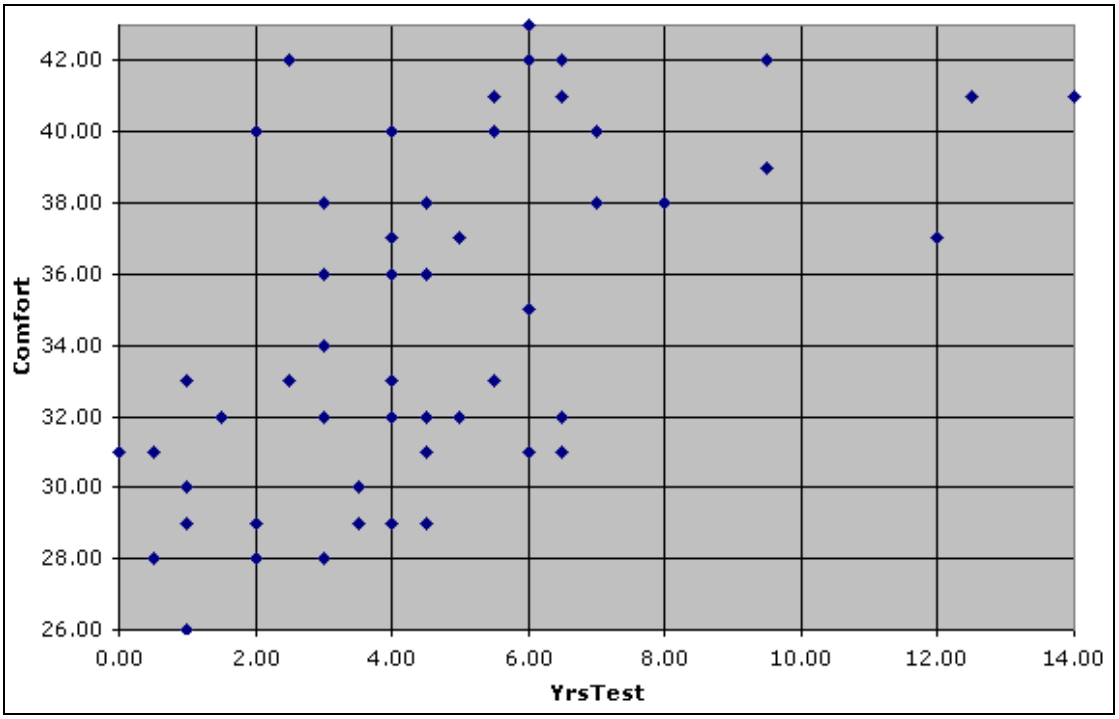
In the screen capture above, only two of the ten Comfort items, Q26 and Q35, have correlations with YrsTest in excess of **r/ec: 0.50**. Once again we see the trio of Q28, Q31, and Q34 having low correlations.

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:



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 experi-

ence 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.

Chapter 9

Lertap, Excel, and SPSS

Contents

This chapter's objective is to impart some helpful pointers on using Lertap within Excel, getting Lertap and SPSS to communicate, fixing font problems which occasionally arise, and importing data from other systems.

How Lertap worksheets are linked

By now readers are well aware of the two main steps to using Lertap's Run option: "Interpret CCs lines", and "Elmillon item analysis". Before these two actions can be used, a Lertap Excel workbook must contain a worksheet called "Data", and a worksheet called "CCs".

It might be useful to examine what happens when Lertap's Run options are taken.

When "Interpret CCs lines" is used, new worksheets are added to the workbook by Lertap. Only one of these, a sheet called "Freqs", will be visible—the others, called "Sub" worksheets, will be hidden from view, waiting for Elmillon to run.

The Freqs worksheet may be deleted at any time. None of the Lertap procedures requires information from the Freqs worksheet.

The hidden Sub worksheets, on the other hand, are vital to the functioning of Elmillon, as well as some other procedures. If the Sub worksheets are unhidden and deleted, the "Elmillon item analysis" will be unsuccessful. Should this happen, "Interpret CCs lines" must be used again.

When Elmillon runs, it creates more worksheets. For example, it always creates a worksheet called "Scores", a sheet which has two sections—test scores, and

summary statistics. The Scores worksheet is required by such Lertap procedures as the histogrammer, the scatterplotter, and the external criterion analyser.

It is generally safe to delete columns from the Scores worksheet without major consequences, but *the rows in this sheet should not be tampered with*. If by chance something were to happen to one or more of the rows of the Scores sheet, the functioning of all those procedures which look to Scores for information, such as the histogrammer, the scatterplotter, and the external criterion analyser, will be jeopardised. *The rows in the Scores sheet should not be tampered with.*

Experienced Excel users may be tempted to get into the Scores sheet, and have Excel sort it in some manner. This should not be done. There is a one-to-one correspondence between the rows in the Scores sheet and the rows in the Data sheet; sorting the Scores sheet will destroy this crucial linkage.

Lertap has an icon on its toolbar which will sort scores. It does this by first making a copy of the Scores sheet, which it calls "Sorted". Once the Sorted sheet is ready, all Lertap sorts are done on it. This is the pattern which should be followed if users want to use the results in the Scores worksheet: have Lertap make its Sorted copy, or use Excel to make your own copy using your preferred name, and then work with the copy.

The rows at the bottom of the Scores sheet have "headers" in the first column, such as "n", "Min", "Median", and so on, and these too should be left as Lertap creates them. One or two Lertap procedures search down the first column of the Scores sheet, looking for the statistic they need, such as "n".

Besides making and controlling the Scores worksheet, Elmillon also creates worksheets with item and subtest results. Each subtest, cognitive or affective, will usually have two statistics worksheets, such as "Stats1f" and "Stats1b". Cognitive subtests will generally have a third worksheet with upper-lower results, called "Stats1ul".

Only one of these statistics worksheets is used by another procedure. It's the "b" sheet, such as Stats1b. The procedure which creates the item response charts requires the "b" sheet(s) in order to do its job.

The histogrammer will make two worksheets if the Analysis ToolPak is installed. One of these, the "L" sheet, as in Histo1L, feeds the second histogram sheet, the "E" sheet. Once the E sheet is made, either or both of the histogram worksheets may be deleted without consequence.

Changing the Data and CCs sheets

Changes may be made at any time to the Data and CCs sheets. However, any changes made in these worksheets will not automatically ripple through to the other sheets. If a change is made in one of the sheets, or in both, "Interpret CCs lines" and "Elmillon item analysis" must be used again.

When "Interpret CCs lines" is run a second time, Lertap will warn that it is about to delete the results worksheets, also called the secondary worksheets—by which

it means Freqs, Scores, Stats1f, and so on. In fact, every worksheet in the workbook will be deleted, except Data and CCs, unless they are first renamed.

If wanted, worksheets may be saved from this automatic deletion process by renaming them. A suggested procedure for doing this is to put something in front of the original name. For example, if you want to save the Stats1ul sheet, rename it as OrigStats1ul, or Orig1ul—don't use a name which begins with the letters "stat".

Preparing a Lertap workbook for hibernation

It's possible for a Lertap workbook to come to have a dozen or even more secondary worksheets. Users who have completed their analyses may want to save storage space on their computer or server by deleting the results sheets, knowing they can always recreate them with the "Interpret CCs lines" and "Elmillon item analysis" options.

The left-most icon on the Lertap toolbar, the X to the left of the yellow smiley face, is there for just this purpose.

From Lertap to SPSS

Any Lertap worksheet may be made to work with SPSS, the Statistical Package for the Social Sciences[36]. The 8-ball towards the right-hand side of Lertap's toolbar eases the process of preparing a Lertap worksheet so that it may be used in SPSS.

We'll work through an example.

We have a data set named MSLQ1.xls. Its Data worksheet contains information collected from 139 undergraduate students studying at Curtin University's Faculty of Education. The data in the worksheet are from the students' responses to a survey, an instrument which asked them for two types of information: demographic information, such as gender, age, and course of study, and information about their study habits and strategies. The latter information was collected by using the University of Michigan's Motivated Strategies for Learning Questionnaire, the MSLQ (Pintrich, *et al*, 1991; see Chapter 3 section, "A large-scale survey").

The version of the MSLQ used in this study had 10 affective scales, or subtests. Four of the scales used items which had to be reversed before they were scored.

Lertap was used to create the 10 scale scores for each student, and to get the response- and item-level reports which are its forte, such as Stats1f and Stats1b.

Once the scales were shown to have adequate alpha values, some of their scores were added to the Data worksheet. For example, one of the scales had Title=(SelfReg), with Name=(Self-regulation). SelfReg consisted of 12 items scattered throughout the data set in non-contiguous columns, with two of its items requiring reverse scoring. Lertap found SelfReg's alpha reliability to be

[36] The SPSS website is www.spss.com

0.77, a value which confirmed that the scale's scores could be used as a "consistent" affective measure.

It was decided to use SPSS in order to see if there might be a relationship between course of study and SelfReg. We needed to export a Lertap worksheet to SPSS, but we had some preparatory work to do first. Course of study was coded in column 4 of the Data sheet, and had a column header of "Degree". The SelfReg scores were in column 4 of Lertap's Scores worksheet. It's only possible to export one worksheet, and here we had the relevant variables in two different sheets. What to do?

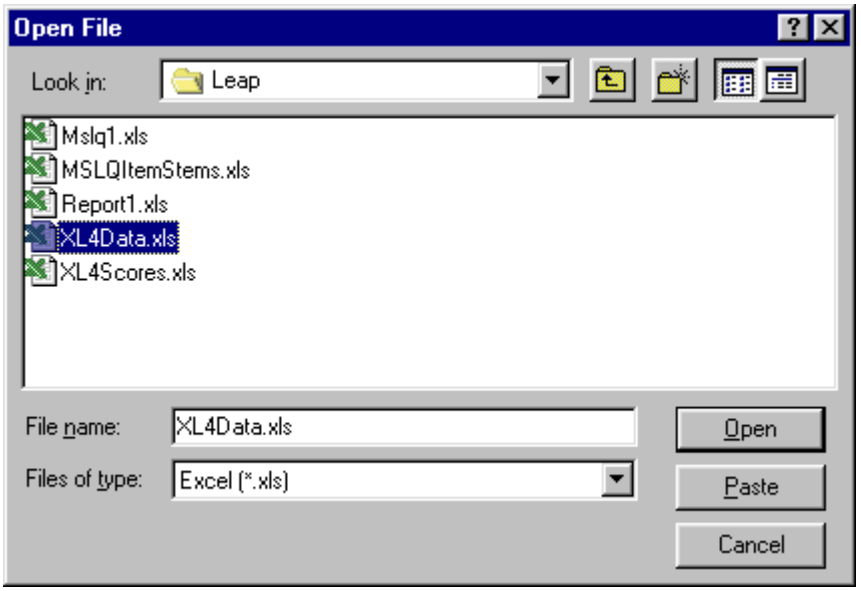
Easy. We used Lertap's Move option to copy SelfReg from the Scores worksheet to the Data sheet. For some reason, Lertap put SelfReg into column 71 of the Data worksheet, even though columns 68-70 were empty. We used Excel to move SelfReg to column 68.

Then, with the Data sheet as our active worksheet, our highlighted worksheet, we clicked on Lertap's 8-ball. Lertap said it would save the Data worksheet as an Excel 4 sheet, having a title of XL4Data, and it did.

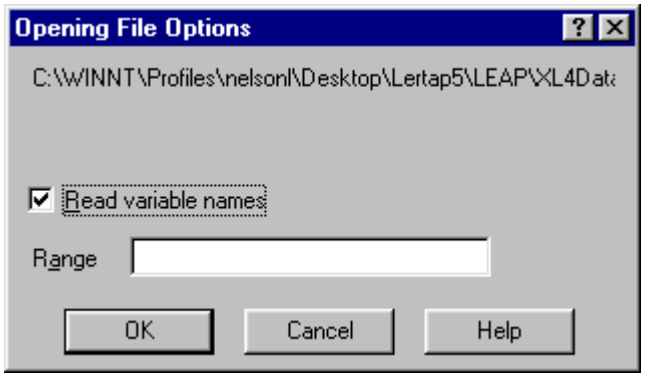
As mentioned in the next chapter, Lertap fixed the XL4Data sheet so that it had only one row of header information at the top, not the two which it had in the original workbook. It did this to maximise compatibility with SPSS, a system which wants to have a single row of headers, nothing more.

We looked at the XL4Data sheet to satisfy ourselves that it seemed okay. Then we closed it. We closed it as we were going to go to SPSS to open it, and knew we'd be likely to get a "sharing violation" is we didn't close Excel's copy.

We started SPSS 10, and used its File / Open / Data / option, going to the right folder on our computer, and telling it to look for Excel files, as shown below:



We clicked on Open, and got the following box:

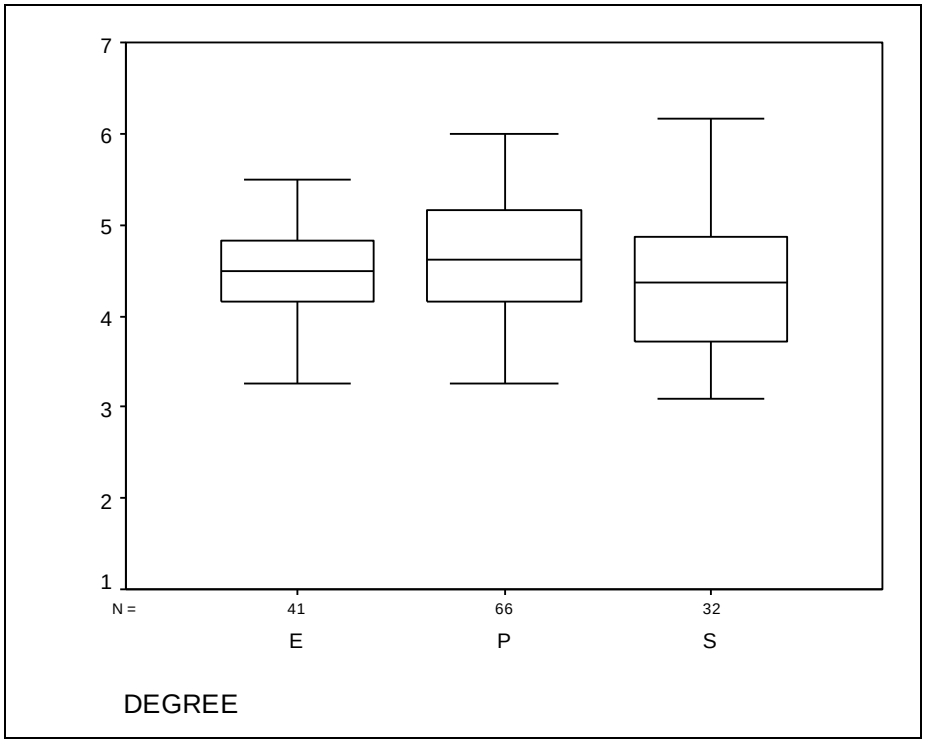


Note the tick in the Read variable names box? This is important. It gets SPSS to interpret the first row of the in-coming Excel worksheet as a header row.

We clicked on OK; there being no need to fill in the Range box.

SPSS said some unpleasant things about our XL4Data.xls file, giving us a variety of "invalid data" statements. Mostly it was referring to our DOB column in the Lertap Data worksheet, which we used for date of birth information. Many of the dates were missing, or typed in as a text value instead of an Excel date. This was not of particular concern for what we wanted to do, and so we proceeded.

We like Boxplots, and asked SPSS to make one, plotting SelfReg scores for each of the three Degree values. Here's what we got:



We then went on to do some means comparisons. By the time we had finished our analyses, we had copied all ten of the scale scores from Lertap's Scores sheet to its Data sheet, used the 8-ball, and picked up the lot in SPSS.

An important Lertap / SPSS issue

An important matter to keep in mind when setting up column headers in Lertap's Data worksheet, and when using the Title=() specification on *sub cards, is to realise that these headers and titles will become "variable names" if you later use SPSS. There are some restrictions on how such names are formed.

SPSS variable names can contain up to eight characters, the first of which must be a letter. The names should not contain a space. They may contain the underscore character, and the "period" (full stop), but, if they do, these characters should not be at the end of the name.

SPSS variable names must not be the same as SPSS reserved keywords. These keywords include the set {ALL AND BY EQ GE GT LE LT NE NOT OR TO WITH}.

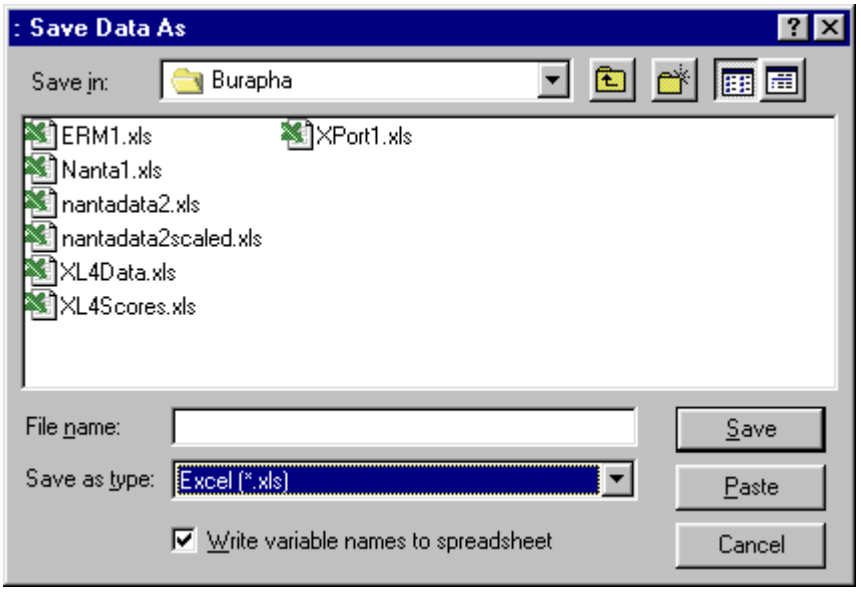
These naming rules make it unwise to use item headers such as 1, 2, 3, and Item 1, Item 2, Item 3, or Q 1, Q 2, Q 3. However, Item_1 would be okay, as would Q_1.

If you have no intention of using SPSS, Lertap's column headers and Title=() specifications can be just about anything you want. Making an effort to keep them at eight characters or less is a good idea (in fact, if the Title is longer than eight characters, Lertap uses just the first eight).

From SPSS to Excel to Lertap

If item response data have been entered into SPSS, it is possible to have SPSS' data editor save the contents of its "Data View" as an Excel file, or workbook.

To do this, in the SPSS data editor, use the File / Save As... option and complete the following dialog box:



Make sure that the "Save as type:" field is as above, and also make sure there's a tick in the "Write variable names to spreadsheet" box.

SPSS will produce a log of what it has done. After it has done this, close SPSS, go to Excel, and open the workbook you saved via the "File name:" box above (it's empty above, but you will, of course, supply a name).

There are then several steps to follow before the new workbook will be ready for use with Lertap. With the new workbook open in Excel,

1. Use Excel's Tools / Options... tabs, look at the General tab, and place a tick in the "R1C1 reference style" box. This will make sure that the new workbook uses digits instead of letters for its columns.
2. Rename the workbook's worksheet to Data.
3. Insert a new row at the top of the Data worksheet, and type a title or description of some sort in it. This row is required as Lertap wants its Data worksheet to have two header rows—the first can have a description of any sort, while the second must have headers for the columns.
4. Use Excel's Insert / Worksheet option to add a new worksheet to the workbook. Rename it as CCs.
5. Prepare appropriate control cards, or job definition statements, in the CCs sheet.
6. Use Excel's File / Open... option to open the Lertap5.xls file. This will install Lertap's toolbar, and give access to the Run option, the one which leads to "Interpret CCs lines", and "Elmillon item analysis".

That's it, you should be ready to roll. If you someday have the chance to follow these steps, however, you may notice an annoying problem with the new workbook: all of its sheets will be set to work in a fixed-pitch font, usually Courier.

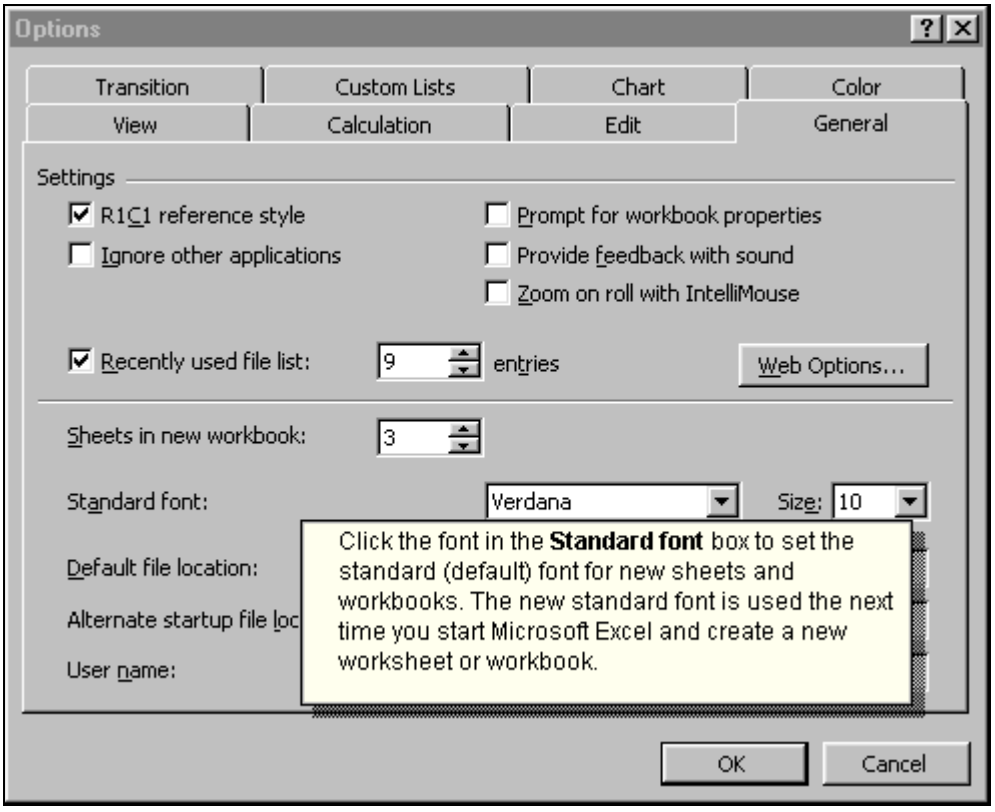
Lertap's reports don't really sparkle with Courier, they look better with the Verdana font, or with Arial. It's supposed to be feasible to set the default Excel worksheet font on one of the Options tabs, but doing so may not work as one might expect. When you run into font problems of this nature, the following little section may be useful.

More about Excel fonts and Lertap

The font used in any Excel worksheet may be changed quickly. Use the CTRL and A keys at the same time, and then select, say, Verdana from the Font box on Excel's Formatting toolbar (use View / Toolbars if the Formatting bar is not displayed).

Lertap is programmed to use the Verdana font, point size 10, in all but one of its worksheets. The exception is the CCs worksheet, where Courier New, a fixed-pitch font, is preferred (but not required).

Excel has a setting called the "Standard Font". It may be set using the General tab under Excel's Tools / Options, as seen below[37]:



Setting the Standard Font does not seem to work as one might often hope. If a new workbook is opened when the Standard Font is set to, for example, Arial, changing the Standard Font from within the same workbook to, say, Verdana, will not "stick". We can close the workbook, exit from Excel, start up again, reopen

[37] We were using Excel 2000 when we took this screen shot.

the workbook, find its Standard Font set to Verdana, but any new worksheets in that same workbook will inherit the original Standard Font, Arial.

Reading the fine print provided by Excel's Help system will reveal that the Standard Font only applies to new workbooks. Workbooks which existed before the Standard Font was changed are not affected—in other words, the default font for existing workbooks is fixed at whatever the Standard Font was when the workbooks were first created.

How to work around this? Let's assume you have set up a workbook outside of Lertap, that is, used Excel without Lertap's toolbar showing. Say you've set up a Data worksheet following Lertap's requirements (first two rows used for headers, as described in Chapter 3). Say this worksheet's font is, perhaps, Times New Roman. If you start up Lertap at this point, and access its Run options, the reports which Lertap creates, such as Stats1f and Stats1b, will (more likely than not) use Times New Roman. They may not look as good as they would with Lertap's preferred font, which is Verdana.

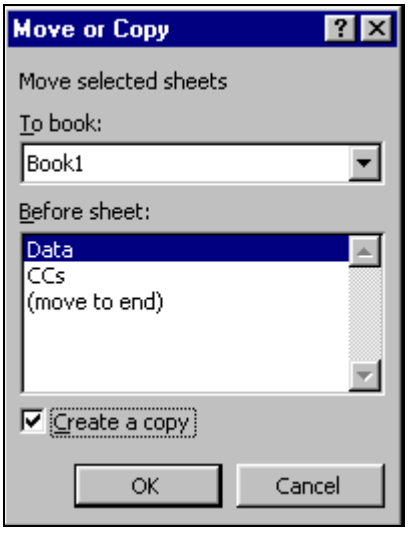
How to get the reports in Verdana? You could try reformatting each report using the CTRL and A keys at the same time, as mentioned above. However, things may still be less than optimal—the contents of some cells may not show in their entirety; you may have to turn on "Row & column headers"[38], and then resize the columns.

How to circumvent font problems such as these? Begin by using Lertap's New option to "Make a new blank Lertap5 workbook". Take note of the name of the new workbook which Excel creates—it'll usually be something like Book1.xls.

Next, go to the workbook which contains the Data worksheet set up outside of Lertap. Make a copy of this worksheet, and move the copy to the new Lertap workbook (Book1.xls, or whatever it's called).

How to make a copy of a worksheet and have it moved to another workbook? Excel's Edit options include one to "Move or Copy Sheet"[39]. Taking this option will result in a dialog box like this one:

[38] Use the View Tab on Excel's Tools / Options for this.
[39] Right-clicking on the worksheet's tab is a useful shortcut.



Above, we placed a tick in the "Create a copy" box, and chose Book1 from the drop-down list of open workbooks which appeared when we clicked on the little arrow on the right-hand side of the "To book". Once the OK button is clicked, Excel will copy the worksheet to Book1. At this point you'd want to delete the Data sheet from Book1, and rename the worksheet you've just copied so that its name is Data.

Next, place appropriate control "cards" in Book1's CCs sheet. You should then find that the report worksheets which Lertap creates when its Run options are used are well formatted, and employ the Verdana font.

Importing other data formats

Excel can import data from a variety of other systems. We asked Excel's Help system to list them, and here's what it said:

> Microsoft Office provides drivers that you can use to retrieve data from the following data sources:
>
> Microsoft SQL Server OLAP Services (OLAP provider)
> Microsoft Access 2000
> dBASE
> Microsoft FoxPro
> Oracle
> Paradox
> SQL Server
> Text file databases
>
> Notes:
>
> You can retrieve data from a Microsoft Exchange or Lotus 1-2-3 data source by using DAO in Microsoft Visual Basic.
>
> You can use ODBC drivers or data source drivers from other manufacturers to get information from other types of databases that are not listed here, including other types of OLAP databases. For information about other drivers that might be available from Microsoft, see the xlreadme.txt file. For

> information about installing an ODBC driver or data source driver that is not listed here or in the xlreadme.txt file, check the documentation for the database, or contact your database vendor.

Importing data from a text file is directly supported from Excel's Data / Get External Data / Import Text File... option. You'd use this option to import "csv" files (comma-separated values files), or text files where the tab character has been used to separate fields.

Excel's Data / Get External Data / Import Text File... option provides good support for importing data from text files which have fixed-width fields.

Chapter 10

Computational Methods Used in Lertap 5

Contents

Overview

Lertap 5 produces a variety of scores and statistics, and a few graphs. This chapter presents a summary of these, and details most of the procedures and methods behind them.

This version of Lertap is written in Microsoft's Visual Basic for Applications language, or "VBA". VBA is common to all the applications found in the suite of programs known as Microsoft Office, a collection which includes Word, Excel, Access, and PowerPoint. Of these, it's Excel, Microsoft's spreadsheet program, which acts as the host application for Lertap 5.

Understanding how Lertap works is aided by knowledge of Excel's basic structure, which consists, at the top level, of a "workbook". A workbook is a collection of

"worksheets". Each worksheet is a matrix, or grid, of rows and columns. The intersection of a row and column produces a "cell".

Readers familiar with other spreadsheet programs, or with older versions of Excel, will want to take particular note of the "workbook" structure of Excel. This structure was first found in the version of Excel known as "Excel 95". It permits multiple spreadsheets, or "worksheets" as Excel calls them, to exist as a bound collection within a single file. Under Windows, this file usually has an extension of "xls". These xls files are referred to as workbooks; each workbook may contain from one to hundreds of individual worksheets.

Excel has traditionally used numbers to label worksheet rows, and letters to label columns. It refers to this row/column naming method as the "A1" referencing system. Under the A1 method, cell A1 refers to the intersection of column A with row 1 (one).

Lertap much prefers another referencing system, one which uses numbers as labels for both rows and columns, one which follows the mathematical convention of referring to a cell in a matrix first by its row number, then by its column number. The top left cell in a matrix, or worksheet, is cell (1,1), denoting the intersection of the first row with the first column. Cell (2,1) refers to the intersection of row 2 with column 1 (one). This method of labelling cells is called the "R1C1" referencing system in Excel. Each time it starts, Lertap sends a message to Excel, directing it to use R1C1 referencing (without such a message Excel is inclined to start up using A1 referencing).

Lertap 5's basic method of operation is based on the use of worksheets. In this version of Lertap, a data set is an Excel workbook. In previous versions of Lertap, a data set consisted of sets of cards, or card images on magnetic tape, or, starting in late 70s, files on a floppy or fixed disk.

In its most elemental form, a Lertap workbook has two worksheets, one with data records, the other with job definition statements. The job definition statements are often referred to as "Lertap control cards".

A Lertap workbook may have any name. Its two fundamental worksheets, however, may not—they must be called Data and CCs. As users direct Lertap to take action, additional worksheets are added, by Lertap, to their workbook. These Lertap-generated worksheets can be seen as secondary sheets—they result from Lertap running its various analyses, using the two primary sheets of Data and CCs as its source. Names of these secondary sheets include "Freqs", "Scores", "Stats1f", "Statsb" (and others).

It is generally the case that users need only save the two primary sheets, Data and CCs, in their workbook. This is so as the other sheets, the so-called secondary ones, are derived from the primary ones. Any or all of the secondary sheets may be deleted, and then re-generated from the primary ones. Of course, there is no harm, none whatsoever, in having workbooks with many worksheets--there is no need to delete the secondary sheets at all. However, users will sometimes want to share their workbooks with colleagues, and may want to do so by sending them over the Internet. In cases such as this it is often useful to pare

back on the number of worksheets. (Note that it is easy to have Excel extract a copy of any worksheet in a workbook, sending the copy to a new xls file. See Excel's on-line help for instructions.)

The Lertap5.xls workbook

All of Lertap 5's VBA code modules are contained as Excel-based macros within the file Lertap5.xls. This file also contains worksheets, including four visible ones, Comments, Data, CCs, and Syntax, and a special hidden one called System.

When Lertap5.xls is opened, all of its code modules are made available (exposed) to all other open Excel workbooks. Access to these modules is via the Lertap5 toolbar, a toolbar having a smiley yellow face towards its left side, and an 8-ball towards its right side. As of October 2000, Lertap's standard toolbar looked like this[40]:

Users may have Lertap workbooks open without opening the Lertap5.xls file. This they will do when they want to review results previously obtained, or when they are preparing new Data worksheets. The Lertap5.xls file need be opened only when access to the Lertap toolbar is desired.

It is possible to have Lertap5.xls open every time Excel is started, something which guarantees that the Lertap toolbar is always present. Information on how to do this may be found in Excel's on-line help manual.

Interpret CCs lines (the Sub worksheets)

Say a user has saved a Lertap5 data set in a workbook named QuizData.xls. Within this file, or workbook, the user has saved all data records in a worksheet named "Data", and has recorded some initial job definition statements, or control cards, in another worksheet named CCs.

The Data worksheet has its data records starting in the third row; as required by Lertap the first two rows in the Data sheet are used for titles and headers.

At this point say Lertap's toolbar is accessed, and the Run menu is used to "Interpret CCs lines". The respective VBA code module, or macro, is called into action, parsing the lines found in the CCs worksheet, and, if no syntax errors are detected, writing core operational information to new worksheets, called the "Sub" sheets. One Sub worksheet will be added to the workbook for every *****col statement found in the CCs worksheet. The first Sub worksheet will be named Sub1, the second Sub2, and so forth.

The Sub worksheets are usually hidden by Lertap. They may be unhidden easily—here again Excel's on-line help manual will be of assistance. It may be instructive to users to examine the contents of a Sub file as they contain the fundamental subtest and item data used in Lertap analyses. In fact, after the

[40] There is another toolbar. See the "advanced level toolbar" discussion later in this chapter.

option to "Interpret CCs lines" has once been taken, the contents of a workbook's CCs worksheet are never again referred to, that is, not until the "Interpret CCs lines" option is taken again.

Whenever "Interpret CCs lines" is selected, Lertap examines the worksheets in the currently-active workbook, and deletes all secondary worksheets if any are found. It gives a warning before it does this. It then proceeds to work its way through the lines in the CCs worksheet.

Users who wish to keep their secondary worksheets will want to rename them before returning to "Interpret CCs lines". The preferred, strongly-recommended, procedure for renaming worksheets is to add something in front of their original names. For example, to preserve the "Stats1f" worksheet, a user might rename it as "OrigStats1f".

In summary, "Interpret CCs lines" is the option, and the only option, which gets Lertap to read the contents of the CCs worksheet. In the process of reading CCs lines, Lertap adds new worksheets to the workbook. It adds one Sub worksheet for each *col statement, and it also adds a worksheet called "Freqs".

The Freqs worksheet

The *col "cards" in the CCs worksheet inform Lertap of the location of columns in the Data worksheet which contain item response characters. For example, the card

*col (c3, c5, c10)

informs Lertap that columns 3, 5, and 10 of the Data worksheet contain item responses.

Item responses are single characters. They may upper case letters from the Roman alphabet (A to Z), lower case letters from the Roman alphabet (a to z), or the ten digits, that is, the Arabic number characters (from zero to 9).

The Freqs (for frequencies) worksheet simply lists the number of times each of these characters is found in the columns. If the columns in the Data worksheet are headed by a title in Row 2, such as "Q1", for example, then these headers will appear in the Freqs listing.

It is possible for a column in the Data worksheet to be referenced by more than one *col card. However, a column's frequency tally will appear only once in the Freqs worksheet.

If a column is found to be empty, to contain something which is not a letter or a digit, or to have contents longer than one character, it is said to have an "other" response. These are denoted as "?" in the Freqs listing. Response characters which would fit Freqs' "other" label would be, for example, punctuation characters, such as the comma, semicolon, colon, and full stop (or period), special characters such as *()&^%$#@, and the space (or blank).

Note that what gets tallied as a response character in the Freqs listing may later be classed as "other" in some subsequent Lertap analysis. In item analyses, for example, Lertap requires that items use no more than ten response characters. Say that a particular subtest has items which use four response characters, perhaps ABCD. Then a lower case equivalent to these letters will, in item analyses, fall into the "other" response category. The Freqs listing is not so particular, allowing 62 characters through its filter (26 upper case letters, 26 lower case letters, and 10 digits)—as a consequence, Freqs will tally the frequency of the lower case equivalents, even though they may well be classed as "other" responses when item analysis results are reported.

A Freqs display serves two immediate purposes: it gives a frequency response tally, and it indicates if there are any strange responses in the Data worksheet. A strange response is an unexpected one. For example, in a column where M and F have been used to code respondent gender, an "m", a "f", a "1", or a "2" would all be unexpected.

When the Freqs listing indicates the presence of unexpected responses, Excel's data filter capability may be used to quickly locate the Data records containing the responses. This is a powerful utility, easy to use (after some practice)--refer to Excel's on-line help manual for instructions.

It is possible to have Lertap make a Freqs worksheet without also making Sub worksheets. To do this, the first line in the CCs sheet must be a *col line, and the next CCs line should be empty.

The Freqs worksheet may be deleted at any time. None of the analyses which may run subsequent to "Interpret CCs lines" require information from the Freqs worksheet.

Elmillon item analysis

Elmillon is Lertap's main program, responsible for producing scores and item and test statistics. Its scores may be referred to as "scale scores" in the case of affective instruments, or, in the cognitive case, as either "test scores", or "subtest scores".

Elmillon begins its work by looking through the currently-active workbook for worksheets whose names begin with the letters "Sub" (or "SUB", or "sub"—case is not important). These sheets, usually hidden from view, are created by the "Interpret CCs lines" option (see above), and will have names such as Sub1, Sub2, Sub3, and so forth. There will be one Sub sheet for each *col card in the CCs worksheet.

When a Sub sheet is encountered, its contents are read, and basic calculation accumulators are dynamically dimensioned in memory. Elmillon then makes a complete pass through the Data worksheet, filling up its accumulators with response frequencies, and either forming scores, in the case of an internal criterion, or, in the case of an external criterion, reading scores from another worksheet. If Elmillon is forming test scores for the first time, it writes them to the scores worksheet during this pass.

In this phase, Elmillon also opens a temporary, hidden, scores worksheet called "ScratchScores", and writes copies of the criterion scores to it, along with a pointer to their respective records in the Data worksheet. The pointer goes into the first column of the new scratch sheet, with the corresponding score getting recorded in the second column.

If the subtest being processed is a cognitive one, an Upper-Lower analysis is usually called for. This requires sorting the criterion scores, after which the scores corresponding to the upper group are written into the third column of the scratch scores sheet, with the lower group's scores going into the fourth column.

Elmillon then uses the criterion scores to update its statistics accumulators. Statistics are aggregated at three distinct levels: item responses, items, and subtest.

With statistics in hand (or memory, as it were), Elmillon then adds two or three new worksheets to the workbook. One of these will have a name similar to "Stats1f", while another will be named "Stats1b", or something similar. If Upper-Lower statistics are required, a worksheet with a name similar to "Stats1ul" is created.

The little "f", "b", and "ul" letters at the end of the names of these new worksheets signify "full", "brief", and "upper-lower". The full worksheets have very detailed item and subtest performance data, while the brief worksheets have concise item performance summaries. The full worksheets have report formats which will be familiar to users of Lertap 2 and Lertap 3. The brief reports are new to this version, as are the upper-lower ones.

The digit which precedes the "f", "b", and "ul" is simply a sequential counter indicating the ordinal position of the subtest in the CCs worksheet. The first subtest corresponds to the first *col card in the CCs sheet.

The worksheet with subtest scores will be called "Scores". There is always but one Scores worksheet per workbook (there may, however, be off-shoots of Scores in the workbook, such as the "Sorted" scores worksheet, but these are not produced by Elmillon itself).

Elmillon will refuse to work if there are no Sub worksheets in the workbook. It regards this to be the case whenever its scan of worksheet names fails to uncover any which begin with the letters "Sub", or "sub", or "SUB"; case is not important.

We turn now to a detailed discussion of the contents of worksheets such as Stats1f, Stats1b, and Stats1ul, and mention how their statistics are derived. The nature of these statistics will vary, depending on whether the respective subtest is cognitive or affective in type.

Stats1f for cognitive subtests

Cognitive subtests are comprised of items which have a correct answer. Consider the typical Stats1f item response statistics shown below:

Item 7

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	20	0.33	-0.26	-0.34	10.05	-0.37
B	0.00	1	0.02	-0.18	-0.56	3.00	-1.39
C	0.00	7	0.12	-0.32	-0.53	6.43	-0.89
D	0.00	0	0.00	0.00	0.00	0.00	0.00
E	1.00	31	0.52	0.40	0.50	15.71	0.44
other	0.00	1	0.02	0.18	0.54	22.00	1.35

Here (above) the label of "Item 7" has come from row 2 of the Data worksheet. This is the row reserved for column headings.

Item 7 used five options, having response codes: A, B, C, D, and E. The correct answer was E, a fact which is denoted by the underlining in the table. Elmillon discovered one Data record which had a response which was not one of these five characters, and has, consequently, shown an "other" line for this item. At this point reference to the Freqs display for this item might reveal what the other response was; for cognitive items it's often a non-response—these are sometimes spaces, or blanks, and sometimes special characters reserved to code missing data, depending on how the user has processed the data.

The column headed "wt." indicates the response weight, or number of points associated with each of the response codes. Above we see a typical cognitive item: only one of the responses has a weight other than zero. If a respondent selected E as his or her answer for Item 7, then that respondent would get one point. This is so as 1.00 is the "score" associated with a response of E.

The "n" column indicates the number of respondents selecting each response, while "p" is this number expressed as a proportion.

How many respondents were there? Sum down the "n" column. There were 60 respondents. Thus "p" for response A is 20 divided by 60, or 0.33.

"p" for the correct answer is often called the item's ***difficulty***. For Item 7 the difficulty is 0.52. This is the proportion of respondents who got the item right.

If more than one response to an item has a non-zero weight, then Lertap defines item difficulty as the number of people who got some points for their response to the item, divided by the total number of respondents. As an example, consider this table:

option	wt.	n	p
A	0.50	20	0.33
B	0.00	1	0.02
C	0.00	7	0.12
D	0.00	0	0.00
E	1.00	31	0.52
other	0.00	1	0.02

Above, options A and E both have non-zero weights, and the number of people who got some points for their answer to this item is 51 (being 20 plus 31). The total number of respondents is 60, so the item's difficulty would be 51/60, or 0.85[41].

The column headed "pb(r)" indicates the point-biserial correlation of each response with the criterion score. The usual criterion score is the subtest score, and, when this is the case, the criterion is said to be an internal one.

There are many equations which may be used to calculate the value of the point-biserial correlation coefficient. Lertap uses an equation which may be seen in editions of Glass & Stanley (1970, p.164, eq. 9.6; 1974, p.163, eq. 9.6); Kaplan & Sacuzzo (1993), and Magnusson (1967) also have useful equations for the point-biserial coefficient.

The point-biserial correlation is interpreted as a conventional Pearson product-moment coefficient. In fact, if we gave all those who selected any given response a "1", and all those who did not a "0", and then correlated these "scores" with the respondent's subtest score using a conventional product-moment equation, we'd get the same result as that obtained by applying the point-biserial computation equation.

The value of pb(r) for the correct answer to a cognitive item is generally referred to as the item's ***discrimination*** index. Item 7's point-biserial discrimination index is 0.40.

Now, when two scores are correlated, if one score forms part of the other, the value of the correlation coefficient will be artificially inflated. This will be the case when an item analysis uses an internal criterion—such a criterion score is simply the subtest score, a score which is obtained by summing the points earned on each item. Each item's "score", or points, forms part of the subtest score, and this will serve to inflate the value of the correlation coefficient.

Lertap corrects for this part-whole inflation by calculating what the point-biserial value would be if the item were removed from the subtest. For cognitive items, this correction to the point-biserial figures applies to the response with the greatest "wt." value.

The effect of the correction for inflation on pb(r) values may be seen by using Lertap to run an "External criterion analysis" using as the criterion the column in the Scores worksheet where the subtest's scores are found. In the case of Item 7, for example, the value of pb(r) for the right answer was 0.46 before the correction was applied (compare with 0.40 above). The corrected figure will always be lower than the uncorrected value unless the number of items is very great. Here, Item 7 was from a subtest with 25 items. When the number of items is smaller than this the effect of the correction will be even more substantial.

[41] The *mws card is used to multiply-key cognitive items, as mentioned in Chapter 5. A card such as *mws c7, .5, 0, 0, 0, 1 would produce the response weights shown.

Users of older versions of Lertap might want to note that Lertap 2 did not apply this correction to cognitive items. Lertap 3 did.

The column headed "b(r)" indicates the biserial correlation of each response with the criterion score. Biserial figures are corrected for part-whole inflation in the same manner as followed for the pb(r) values.

Methods for determining the biserial correlation coefficient may be seen in Allen and Yen (1979, p.39); in editions of Glass & Stanley (1970, p.171, eq. 9.11; 1974, p.171, eq. 9.11); in Gulliksen (1950, p.426); and in Magnusson (1967, p.205, eq. 14-7). Lertap uses the equation seen in Glass & Stanley. Values of the normal density function required in the equation are based on an algorithm known as "NDTRI", found in the IBM Scientific Subroutine Package[42].

It is possible for biserial correlation coefficients to have a magnitude greater than one. A particularly complete discussion of the biserial coefficient, with comparisons to the point-biserial coefficient, is found in Lord & Novick (1968, pp.337-344).

The "avg." column indicates the average criterion score earned by the respondents who selected each response. The average criterion score for the 20 respondents who selected response A on Item 7 was 10.05. As a z-score, this average is –0.37, a figure which is computed by subtracting the mean criterion score from the "avg." figure, and dividing the result by the standard deviation of the criterion scores. In this case, the mean criterion score was 12.63, and the standard deviation of the criterion scores was 6.95. Subtracting 12.63 from 10.05, and dividing the result by 6.95, gives a z-score of –0.37.

One could make the point that the part-whole inflation issue applies to the "avg." value as much as it does to the pb(r) values. That is, in the case of internal criteria, the "avg." value is inflated if a response has a "wt." greater than zero. While this is so, Lertap does not apply a correction to the "avg." and "z" values.

Lertap allows cognitive items to have more than one correct answer. More precisely, Lertap allows any response to have any weight (wt.). The correction for inflation will, however, be applied to only one response, the one having the greatest positive weight. If two or more responses share the greatest positive weight, the correction is applied to the response which comes last when Elmillon outputs its results to the Stats1f worksheet.

Above, it was mentioned that Elmillon calculates statistics at three levels: response, item, and subtest. For all items, cognitive and affective, the statistics computed at the item level include the item mean, and the item's product-moment correlation with the criterion score, corrected for part-whole inflation when the criterion is internal, which is the usual case.

In the case of cognitive items, the normal situation is for each item to have one correct answer, with a weight of one point. When this is true, the p value shown

[42] See http://pdp-10.trailing-edge.com/www/lib10/0145/

for the correct response will equal the item's mean, and the pb(r) value will equal the item's product-moment correlation.

The subtest statistics found in sheets such as Stats1f have the following format:

Summary statistics

number of scores (n):	60	
lowest score found:	1.00	(4.0%)
highest score found:	24.00	(96.0%)
median:	12.50	(50.0%)
mean (or average):	12.63	(50.5%)
standard deviation:	6.95	(27.8%)
standard deviation (as a sample):	7.01	(28.0%)
variance (sample):	49.08	

number of subtest items:	25	
minimum possible score:	0.00	
maximum possible score:	25.00	
reliability (coefficient alpha):	0.91	
index of reliability:	0.96	
standard error of measurement:	2.03	(8.1%)

The number of scores is usually equal to the number of records in the Data worksheet. However, if a Data record is missing data for all of the items belonging to a subtest, that record is omitted from calculations.

The first standard deviation value reported by Elmillon is the "population" standard deviation, computed by using "n" in the denominator of the relevant equation. The "sample" standard deviation and variance are computed using "n-1" in the denominator. Lertap commonly uses population variance and standard deviation values in its calculations. For example, it uses the population standard deviation for z-scores.

For a discussion of variance and standard deviation calculations, see Glass & Stanley (1970, p.82 & 1974, p.82), Hays (1973, p.238), Hopkins & Glass (1978, p.78), or Magnusson (1967, p.8).

The figures shown in parentheses are calculated using the maximum possible score. Above, for example, the average subtest score, that is, the mean subtest score, was found to be 12.63, which is 50.5% of the maximum possible score, 25.

The reliability figure reported by Elmillon is Cronbach's coefficient alpha. Lertap 2 used Hoyt's analysis of variance procedure to derive an estimate of reliability. As is now well known, Hoyt's procedure produces the same value as alpha. Lertap 3 used alpha. For a discussion of coefficient alpha, and its interpretation, Pedhazur & Schmelikn (1991, pp.92-100) have a particularly thorough presentation. It has long been known that coefficient alpha is not an index of the factorial complexity

of a test; high values of alpha cannot be interpreted as meaning that a test is measuring just one factor, or that the test's items are necessarily highly homogeneous.

The index of reliability is the square root of the alpha value, while the standard error of measurement equals the standard deviation times the square root of the quantity (1-alpha).

For further discussion of the calculation and interpretation of these statistics, see Ebel & Frisbie (1986), Hopkins, Stanley, & Hopkins (1990), Linn & Gronlund (1995), Magnusson (1967), Mehrens & Lehmann (1991), Oosterhof (1990), and Pedhazur & Schmelikn (1991).

After Stats1f's subtest statistics, Elmillon summarises item difficulties using a series of 10 bands.

item difficulty bands

.00: Item 22
.10:
.20:
.30:
.40: Item 1 Item 2 Item 9 Item 11 Item 14 Item 18 Item 19 Item 20 Item 21 Item 25
.50: Item 3 Item 4 Item 6 Item 7 Item 10 Item 12 Item 15 Item 17 Item 24
.60: Item 8 Item 13 Item 16 Item 23
.70: Item 5
.80:
.90:

Above we see that Item 22's difficulty fell into the first band, labelled ".00". This means that its difficulty was below .10. To see what the actual value was one may either scroll up in the Stats1f worksheet to see the item's response statistics, or look up its value in the Stats1b worksheet.

After the item difficulty bands, Elmillon creates a similar display for item discrimination values:

item discrimination bands

.00:
.10:
.20: Item 4 Item 22
.30: Item 5 Item 14 Item 24
.40: Item 7 Item 9 Item 16 Item 23
.50: Item 3 Item 10 Item 12 Item 15 Item 17
.60: Item 1 Item 2 Item 6 Item 8 Item 11 Item 18 Item 21 Item 25
.70: Item 13 Item 19 Item 20
.80:
.90:

It may be seen, above, that Item 7's discrimination value falls in the .40 band, meaning that it was equal to or greater than .40, but less than .50.

Here Elmillon is reporting item-level data. Item discrimination is now the item's product-moment correlation with the criterion.

For the usual case of a cognitive item with one correct answer, and a weight of one point, an item's mean will be equal to p, its difficulty, and its product-moment correlation will equal its pb(r) value for the correct answer.

The item difficulty and discrimination bands did not feature in Lertap 2; they were introduced in Lertap 3.

New to Lertap 5 is a display of adjusted alpha reliability values. This appears at the very end of the Stats1f worksheet, and has a format such as that seen below (note that to save space the display below has been truncated after the tenth item; there were 25 items in the complete display).

alpha figures (alpha = .9149)

without	*alpha*	*change*
Item 1	0.909	-0.006
Item 2	0.909	-0.006
Item 3	0.911	-0.003
Item 4	0.917	0.002
Item 5	0.915	0.000
Item 6	0.910	-0.005
Item 7	0.914	-0.001
Item 8	0.910	-0.005
Item 9	0.914	-0.001
Item 10	0.911	-0.003

The display above begins by repeating the subtest's overall alpha value, as reported earlier under summary statistics. It then indicates what the value of alpha would be if each item were omitted from the subtest. For example, if Item 1 were removed from the subtest (for some reason), the alpha figure would decrease to 0.909, a change of –0.006 from the original value of alpha.

Correction for chance scoring

It is possible to correct cognitive test scores for the possible effects of guessing by using the "CFC" control word on a *sub card.

The correction is based on what the literature often refers to as the "standard correction". Item weights are changed so that distractors have a weight equal to minus one (-1.00) divided by the total number of distractors used by the item.

For discussions on the pros and cons of using the correction for chance formula, or, more generally, "formula scoring", see Ebel & Frisbie (1986); Hopkins (1998); Linn & Gronlund (1995); Mehrens & Lehmann (1991); Oosterhof (1990); and Wiersma & Jurs (1990).

Assessing the impact of applying Lertap's CFC scoring may be accomplished by duplicating a cognitive subtest's control cards, and using the CFC control word on one of the *sub cards. The following cards indicate how this might be done:

```
*col (c3-c27)
*sub Res=(A,B,C,D,E,F), Name=(Knowledge), Title=(Knwlg)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
*col (c3-c27)
*sub Res=(A,B,C,D,E,F), CFC, Name=(Knowledge CFC), Title=(KnwlgCFC)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
```

The two subtests referred to by these cards are identical, except that the second *sub card carries the CFC control word. The ramifications of using CFC may be seen by looking at the correlation between `Knwlg` and `KnwlgCFC`, by having Lertap make a scatterplot of these two scores, and by comparing sorted listings of the scores.

Stats1b for cognitive subtests

It has been mentioned that Elmillon, Lertap's item analysis program, adds two worksheets for each Sub worksheet found in a workbook. The first of these is called Stats1f; snapshots from a typical Stats1f sheet are shown above.

New in Lertap 5 is a condensed summary of item-level statistics. This is found in the sheet called Stats1b; a sample from a typical Stats1b sheet for a cognitive subtest is shown below.

Lertap5 brief item stats for "Knowledge of LERTAP2", created: 23/08/00.

Res =	A	B	C	D	E	F	other	diff.	disc.	?
Item 1	43%	42%	15%					0.43	0.66	
Item 2	7%	20%	12%	13%	48%			0.48	0.66	
Item 3	3%	2%	53%	42%				0.53	0.54	
Item 4	55%	45%						0.55	0.23	
Item 5	22%	70%	8%					0.70	0.33	
Item 6	27%	50%	23%					0.50	0.62	
Item 7	33%	2%	12%		52%		2%	0.52	0.40	D
Item 8	2%	63%	35%					0.63	0.61	D
Item 9	15%	43%	8%	7%	8%	13%	5%	0.43	0.40	
Item 10	17%	10%	12%	53%			8%	0.53	0.54	

The Stats1b display tries to fit as much item performance data as possible on a single line. It begins with a header line showing possible item responses; above these show as A B C D E and F.

Following the header row, a mixture of response and item-level statistics are given for each item. Above it may be seen that, for Item 7, 33% of the respondents selected option A, corresponding to the same response's p value of 0.33 found in the Stats1f worksheet. The underlining indicates that option A had a response weight greater than zero. (Option A was the correct answer to this item.)

Item 7's "diff." index shows above as 0.52. This value will equal the p value found in the full statistics sheet, Stats1f, if the item has only one correct answer, with a weight of one point, which is the usual case for cognitive items.

Item 7's "disc." index is an item-level statistic, equal to the product-moment correlation of the item with the criterion, corrected for part-whole inflation (discussed above). The disc. value will equal the Stats1f pb(r) value if the item has only one correct answer.

The last column in the Stats1b display is headed with ?, a question mark. If an item has an entry in this column, Lertap is pointing out that the item has one or more distractors which *may* be functioning in a questionable manner. Above we

see, for example, that Lertap has flagged one of Item 7's distractors, D. No-one selected this distractor. This is usually an undesired outcome for a distractor—the role of a distractor is to serve as a foil, a decoy which will appear attractive to weaker respondents. To Lertap, weaker respondents are those with criterion scores below the criterion mean.

To repeat, Lertap regards a distractor as effective if it is selected by respondents, and if these respondents have a below-average score on the criterion. Distractors which are not selected by anyone, and distractors which are selected by respondents whose average criterion score equals or exceeds the criterion mean, are candidates for the ? column.

This summary of distractor performance had no equivalent in Lertap 2. Lertap 3 was the first version to have a distractor adequacy index, reporting the percentage of distractors which were selected by weaker respondents.

The Stats1b report is obviously a brief one. The most complete information about item functioning is given in the Stats1f worksheet.

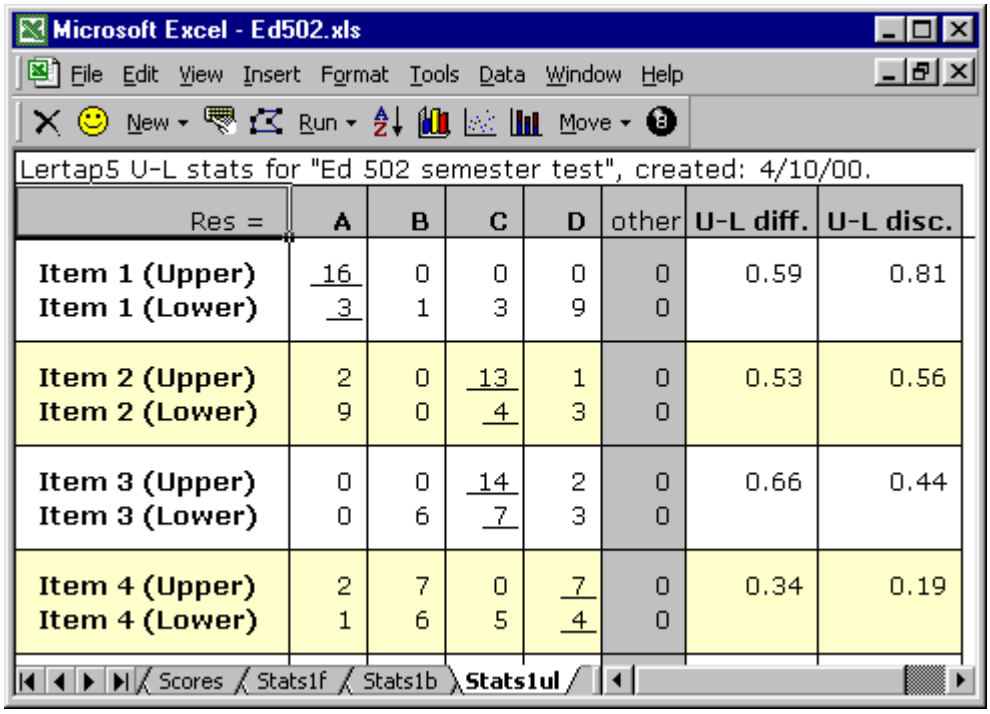
Stats1ul for cognitive subtests

The Stats1b worksheet is new to this version of Lertap, as is another, one which is added only for cognitive subtests: Stats1ul.

The item discrimination statistics seen in both the Stats1f and Stats1b sheets are based on correlation coefficients. There is another way of indexing item discrimination, one which was originally advocated before the birth of desktop computers, and a method which remains popular with many users. It's called the "upper-lower" (U-L) method.

An advantage of the U-L approach is its conceptual simplicity. Test results are used to define two groups—the "upper" group, with high test scores, and the "lower" group, with low scores. If an item, a cognitive item, is discriminating, it should be the case that the upper group is more successful in identifying the correct answer, and less prone to fall for the distractors than is the lower group.

Lertap's U-L stats sheet, Stats1ul, has two sections. The first section summarises item results in the format seen here:



Microsoft Excel - Ed502.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "Ed 502 semester test", created: 4/10/00.

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

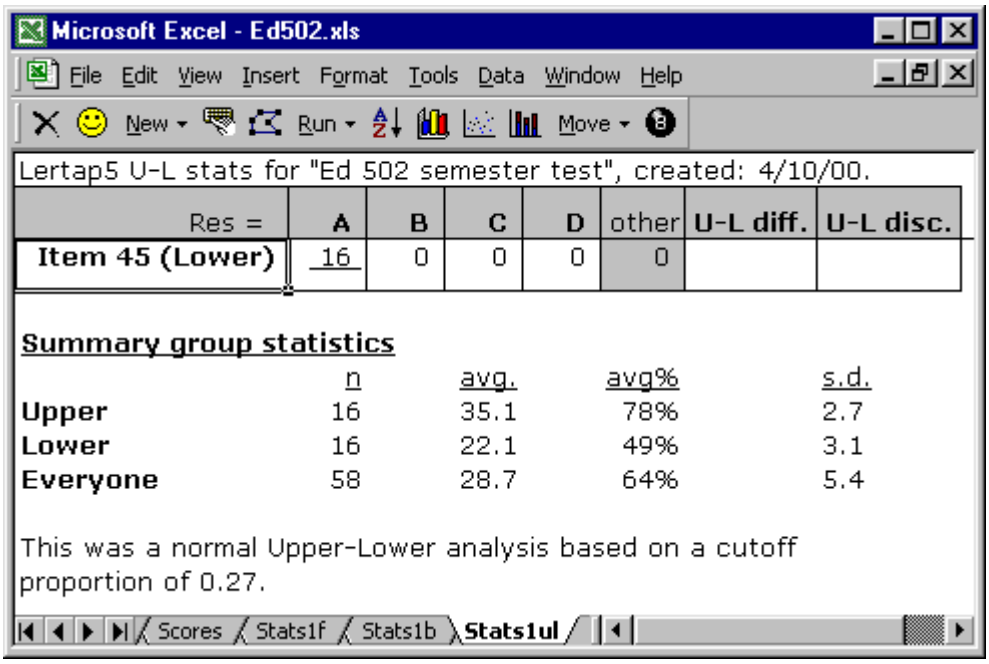
Scores Stats1f Stats1b **Stats1ul**

The U-L item stats show how many people in each group selected each item option. The keyed-correct answer is underlined (all responses with weights greater than zero are underlined—usually there's only one such).

The U-L diff. index is the number of people in both groups who selected the right answer, divided by the sum of the number of people in both groups. In this example, 19 respondents got the Item 1 right, and there was a total of 32 respondents, 16 in the upper group, and 16 in the lower.

The U-L disc. is based on computing two difficulty figures, one for each group. In this example, the difficulty of Item 1 in the upper group was 16/16, or 1.00. In the lower group, the difficulty was 3/16, or 0.19. The U-L disc. is the difference between these two, 1.00 – 0.19, or 0.81.

The lower section of the Stats1ul sheet looks like this:



Microsoft Excel - Ed502.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "Ed 502 semester test", created: 4/10/00.

Res =	A	B	C	D	other	U-L diff.	U-L disc.
Item 45 (Lower)	16	0	0	0	0		

Summary group statistics

	n	avg.	avg%	s.d.
Upper	16	35.1	78%	2.7
Lower	16	22.1	49%	3.1
Everyone	58	28.7	64%	5.4

This was a normal Upper-Lower analysis based on a cutoff proportion of 0.27.

Scores Stats1f Stats1b **Stats1ul**

The statistics shown in this summary section are based on results from the top 27% of the overall group, and the bottom 27%. In this example, the whole group, "Everyone", consisted of 58 people—27% of 58, rounded to the nearest integer, is 16, the "sample size" of each group.

The usual Lertap procedure is to use an internal criterion, the subtest score, to define the groups. To do this, Lertap scores each person's responses, forms an array in memory with copies of the scores, then sorts the scores and picks off the top and bottom groups, writing their results to a temporary, hidden, worksheet called "ScratchScores". It is possible to use an external criterion to define the groups, as mentioned later in this chapter.

The avg. shown in the table is the mean of the scores in each group. The avg.% figure is avg. divided by the maximum possible score. The s.d., standard deviation, is computed using the "population" equation (see the discussion on standard deviation calculations mentioned earlier).

Why are the groups based on 27%? It's a standard, often-recommended figure. According to Hopkins, Stanley, & Hopkins (1990, footnote on p.269), the reason for selecting 27% can be traced to Kelly (1939), who presented a case for this being the optimal value for defining the groups. Garrett (1952, footnote on p.215) wrote "*....There are good reasons for choosing 27%. When the distribution of ability is normal, the sharpest discrimination between extreme groups is obtained when item analysis is based upon the highest and lowest 27 per cent in each case....*".

Many authors suggest that departures from 27% will not have adverse effects. Linn & Gronlund (1995), and Garrett (1952), write that 25% would be fine. Ebel

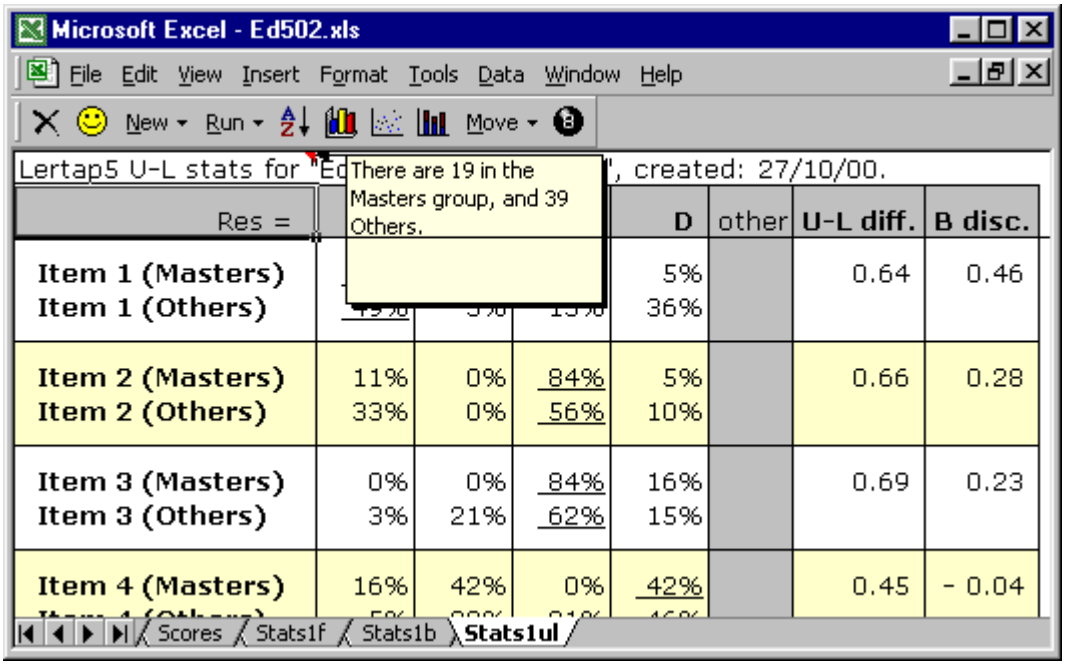
& Frisbie (1986) suggest that 25% or even 33% would be workable, but then clearly state (p.229) that "*....The optimum value is 27 percent....*".

It is possible to have Lertap use something other than 27%, if wanted. The "System" worksheet holds crucial operational settings such as this. It is even possible to have the U-L analysis turned off altogether; this is also managed by changing one of the settings in the System sheet. Refer to a following section on how to access this worksheet.

Mastery and criterion-reference testing

Chapter 5 provides comments on the use of the Mastery and Mastery= control words. These words are used to alter Lertap's U-L analysis so that the groups are defined by reference to a criterion level.

When the Mastery word is used, the upper group will be those who have equalled or bettered the criterion level, and the lower group will be all others. Lertap's default mastery level is 70% of the criterion score. This level is easily changed by using, for example, Mastery=80, which changes the level to 80%.



The screen capture above displays Lertap's standard "mastery" analysis format[43]. The U-L diff. index is calculated as for a normal U-L analysis. The U-L disc. is now called the B disc. index after Brennan (1972). The B disc. index is formed in a manner identical to that used to determine U-L disc: from two difficulty values. In this example, the difficulty of Item 2 in the Mastery group was 0.84, while in the other group it was 0.56. Subtracting the lower group's difficulty from the upper group's, 0.84 – 0.56, gives a B disc. value of 0.28.

[43] The box with group sizes is a comment attached to cell R1C1, flagged by a small red triangle. It displays whenever the mouse pointer moves into the cell.

The report of U-L diff. and B disc. indices is followed by small tables of additional statistics, as shown below:

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.

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

The "Summary group statistics" are similar to those presented in the normal, non-mastery analysis. The "Variance components" section is based on the work of Brennan & Kane (1977, 1984), who suggested that an items-by-persons analysis of variance could be used to estimate errors associated with domain-referenced measurement. Lertap's "Error" variance component is Brennan & Kane's "interaction" component (a more accurate term might be "interaction and error" since both interaction and error variance are involved, and inseparable).

The "Index of dependability" is M(C) in Brennan & Kane (1977). It may be interpreted as a reliability coefficient—it has the same zero-to-one range of a classical reliability index, with values closer to one signifying less error in the measurement process. In Brennan & Kane's approach, error variance has two main components: the standard error of measurement from classical test theory (shown in Lertap's full statistics report), and a component due to item sampling. It is this latter component which distinguishes Brennan & Kane's method from classical test theory, and makes their 68% confidence interval larger than that found in the classical case (for the example above, the classical standard error of measurement, as a proportion, was .062).

The two last lines in the section above are based on the work and recommendations of Subkoviak (1976, 1984). The "Prop. consistent placings" is the statistic commonly referred to in the literature as $\hat{p}_0$, an estimate of the proportion of test takers who have been correctly classified as either master or nonmaster. Lertap follows the Peng and Subkoviak approximation method to derive the estimate, as seen in Subkoviak (1984, pp.275-276). Algorithm 462 from Collected Algorithms

of the CACM[44] is employed to obtain bivariate normal probability values, with coefficient alpha passed as the correlation between the two normal variates.

Since it will always be expected that some will be correctly placed by chance alone, the second statistic, "Prop. beyond chance", estimates the proportion correctly classified as a result of the testing process itself. This statistic is known as kappa in the literature.

Stats1f for affective subtests

When it goes about its calculations, Lertap does not make major distinctions between cognitive and affective subtests. In fact it uses the same methods for deriving statistics at all three levels: response, item, and subtest. In the Stats1f worksheet, differences arise in the application of the correction for part-whole inflation, and in the display of what are called "mean/max" bands instead of item difficulty bands.

Consider the following Stats1f display for an item from a 10-item affective subtest:

Item 30

option	wt.	n	%	pb(r)	avg.	z
1	5.00	9	15.0	0.26	39.8	1.15
2	4.00	20	33.3	0.24	36.1	0.34
3	3.00	17	28.3	-0.12	33.6	-0.19
4	2.00	8	13.3	-0.37	30.1	-0.95
5	1.00	5	8.3	-0.31	29.8	-1.02
other	3.00	1	1.7	-0.15	29.0	-1.19

The results above use a display format essentially identical to that found in Lertap's display of response-level data for cognitive items (see, for example, the results presented earlier for "Item 7"). The "p" column from the cognitive case has been replaced by "%", but otherwise the columns are the same.

The statistical procedures used to derive the values given in the columns are also the same, with one exception: at this level Lertap does not correct any of the pb(r) values for part-whole inflation.

Notice the "other" line above, and, in particular, its weight of 3.00. How did the weight of 3.00 get there?

For affective items, Lertap automatically derives and applies a missing data weight unless told not to. The weight will be equal to what Lertap considers to be the "mid-scale" value. This item portrayed above has weights which range from 5 to 4 to 3 to 2 to 1. The mid-scale weight is 3, and this is what Lertap will apply to every respondent whose answer to this item was "other", that is, not one of the response codes seen under the option column.

[44] Communications of the ACM.

Had the response weights ranged from 1 to 2 to 3 to 4 to 5 to 6, Lertap's weight for "other" responses would be 3.5.

How to defeat this? Use the MDO control word on the subtest's *sub card in the CCs worksheet. Note: this missing data weight applies only to affective subtests, not to cognitive ones. Lertap 2 and Lertap 3 users will want to note that the MDO control word works opposite to what they're used to. In previous versions the missing data weight was derived and applied only when MDO was specifically mentioned; now it's always present, and MDO must be used to turn it off.

The **subtest statistics**, or Summary statistics, for affective subtests are the same as those found in the Stats1f sheets made for cognitive subtests (see above).

A difference in the contents of Stats1f for the affective case is that the item difficulty bands seen in cognitive reports do not appear. The concept of "item difficulty" is not used in the affective case. Instead, Lertap derives what it calls mean/max figures, and summarises these in a series of bands.

mean/max bands

.00:
.10:
.20:
.30:
.40: Item 34
.50:
.60: Item 26 Item 27 Item 30 Item 35
.70: Item 28 Item 29 Item 33
.80: Item 31 Item 32
.90:

An item's mean/max figure is calculated by dividing the item mean by the maximum item weight. For example, Item 30's mean was 3.33; its maximum weight was 5; 3.33 over 5 gives 0.67. Consequently, Item 30 appears on the mean/max band of .60, meaning that its mean/max figure was equal to or greater than .60, but less than .70. Where is an affective item's mean reported? Not in the Stats1f sheet—it will be found in Stats1b.

The mean/max bands are particularly useful in the case of Likert items, where a scale of "strongly disagree" to "strongly agree" is deployed. An item whose mean/max figure falls into the .80 or .90 band is one where respondents reported strong agreement, assuming the strongly agree response to be the one with the maximum weight, of course.

The **correlation bands** seen for affective items indicate where each item's product-moment correlation with the criterion lies. Negative correlation values will map to the first band. Where is an affective item's correlation found? In the Stats1b worksheet (see below).

The **alpha figures** for affective subtests follow the pattern, and method, used in the cognitive case, and are interpreted in the same way.

Stats1b for affective subtests

The differences between Stats1f sheets for cognitive and affective subtests may be slight, but this is not so much the case in the Stats1b worksheet.

Lertap5 brief item stats for "Comfort with using LERTAP2", created: 23/08/00.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Item 26	13%	22%	25%	23%	17%		+	3.08	1.28	0.76
Item 27	5%	23%	37%	35%			-	2.98	0.88	0.55
Item 28	22%	45%	17%	13%		3%	-	3.75	0.94	- 0.14
Item 29	32%	35%	25%	5%		3%	-	3.93	0.89	0.44
Item 30	15%	33%	28%	13%	8%	2%	-	3.33	1.14	0.49

The differences in the Stats1b sheet are seen in the right-most columns. The column labelled "pol." indicates the scoring polarity of each item, that is, whether or not it was reversed. Items which have been reversed are denoted by a minus sign. Item 30 is one such item; when response-level statistics for this item were reported earlier, its weights could be seen running from high to low down the response options.

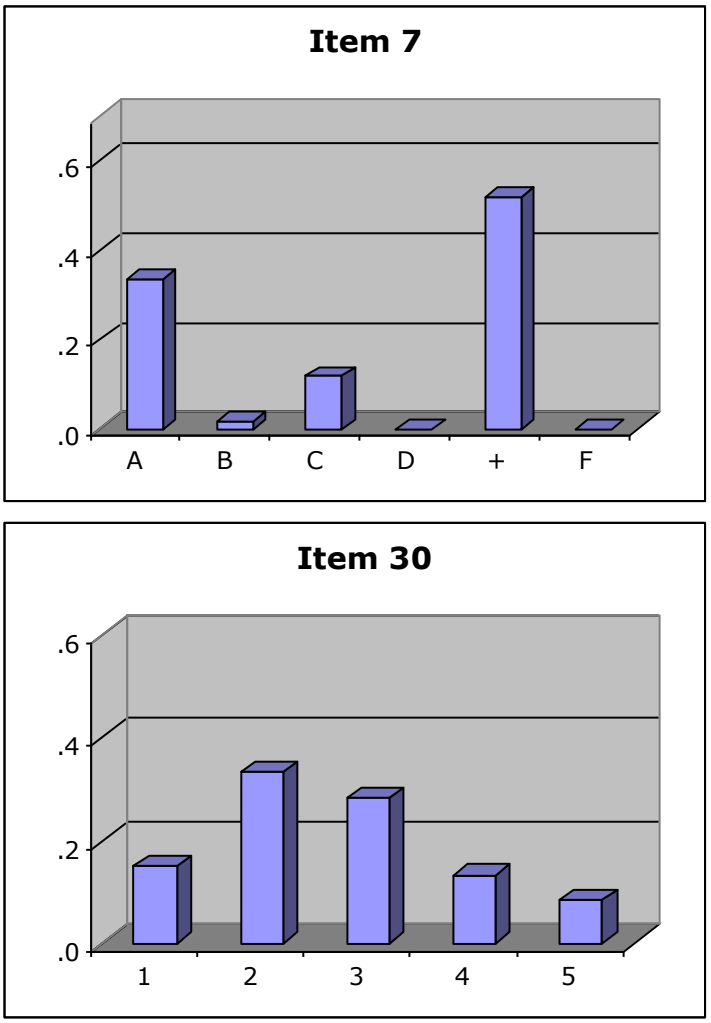
Item mean is what would be expected, that is, the average item score over all respondents, where an item's "score" is its weight—a respondent who chose option 2 on Item 30, for example, got a "score" of 4 on this item.

Item standard deviation is a population value, computed using "n" in the denominator of the appropriate equation (as discussed above).

Item correlation is the Pearson product-moment correlation between the item score and the criterion score, corrected for part-whole inflation.

Item response charts

It is possible to have Lertap make charts from a brief stats sheet, such as Stats1b. The charts option is activated via one of the icons on the Lertap toolbar; it produces displays such as the following:



When charts are made, they generally have corresponding summary item performance data attached to them (not shown here).

Lertap calls on Excel to make these charts, passing it information from the rows of response statistics found in the Stats1b worksheet. It instructs Excel to open a new worksheet, Stats1bCht, just for the charts.

Users of Excel 97 should note that the number of charts which may be made is limited to about 50, sometimes less, unless they have installed a special fix to overcome a system limitation specific to this version of Excel. The fix may be found on the Lertap website (www.lertap.com).

Scores

The Lertap program which creates the Stats1f and Stats1b worksheets is Elmillon. Before Elmillon creates these worksheets, it opens and starts to fill another worksheet called Scores. A sample from a Scores sheet is shown here:

ID	Knwldge	Comfort
9	3.00	32.00
31	12.00	32.00
26	13.00	37.00
27	11.00	32.00
21	14.00	33.00
59	19.00	37.00
47	14.00	42.00
42	20.00	41.00

The Scores worksheet always has some sort of ID information in its first column. The numbers or letters or names seen in this column may come directly from one of the initial two columns in the Data worksheet, providing the column heading in the Data worksheet begins with letters "ID" (or "id"—case is not important).

There will usually be one score for each subtest. The scores will be labelled by the characters found in the Title=() control word on *sub cards.

A respondent's score on each subtest is derived by summing over the respondent's score on each of the items which comprise the subtest.

It is possible to copy columns from the Data worksheet to the Scores worksheet, providing the columns contain only numeric data. It is also possible to copy a column from the Scores worksheet to the Data worksheet. The options for copying columns is found under the Move menu on the Lertap toolbar. It might be emphasised that these options copy and paste, they do not actually pick up a column and move it.

At the bottom of the Scores worksheet appear a variety of statistics, such as those shown below:

n	60	60
Min	1.00	26.00
Median	12.50	33.00
Mean	12.63	34.48
Max	24.00	43.00
s.d.	6.95	4.61
var.	48.27	21.25
MinPos	0.00	10.00
MaxPos	25.00	50.00
Correlations		
Knwldge	1.00	0.80
Comfort	0.80	1.00
average	*0.80*	*0.80*

All of these statistics, except two, are computed by Excel, not Lertap. To see how they're derived, select a cell and examine the contents of Excel's Formula Bar.

The two statistics not computed by Excel are MinPos and MaxPos, which come from Lertap. The first gives the minimum possible score on each subtest, while the second gives the maximum possible. What is the difference between "Min" and "MinPos"? The first is computed by Excel, and is simply the lowest score found in the subtest's score column. The second is calculated by Lertap using the item weights found in the subtest's Sub worksheet, and is the rock-bottom, absolute minimum score anyone could get on the subtest.

The *average* correlation shown in the very last line is the simple average of each score's correlations with all the other scores. (In this example there is only one other score.)

Histograms

An option on the Lertap toolbar will create a "Lertap2-style" histogram from any of the columns in the Scores sheet. This is not a chart, but a special table made with text characters. A snippet from one of these histograms is shown below:

z	score	f	%	cf	c%	
-1.67	1.00	1	1.7%	1	1.7%	□
-1.53	2.00	0	0.0%	1	1.7%	
-1.39	3.00	6	10.0%	7	11.7%	□□□□□□
-1.24	4.00	5	8.3%	12	20.0%	□□□□□
-1.10	5.00	3	5.0%	15	25.0%	□□□
-0.95	6.00	0	0.0%	15	25.0%	
-0.81	7.00	4	6.7%	19	31.7%	□□□□
-0.67	8.00	2	3.3%	21	35.0%	□□
-0.52	9.00	0	0.0%	21	35.0%	
-0.38	10.00	2	3.3%	23	38.3%	□□
-0.24	11.00	3	5.0%	26	43.3%	□□□
-0.09	12.00	4	6.7%	30	50.0%	□□□□

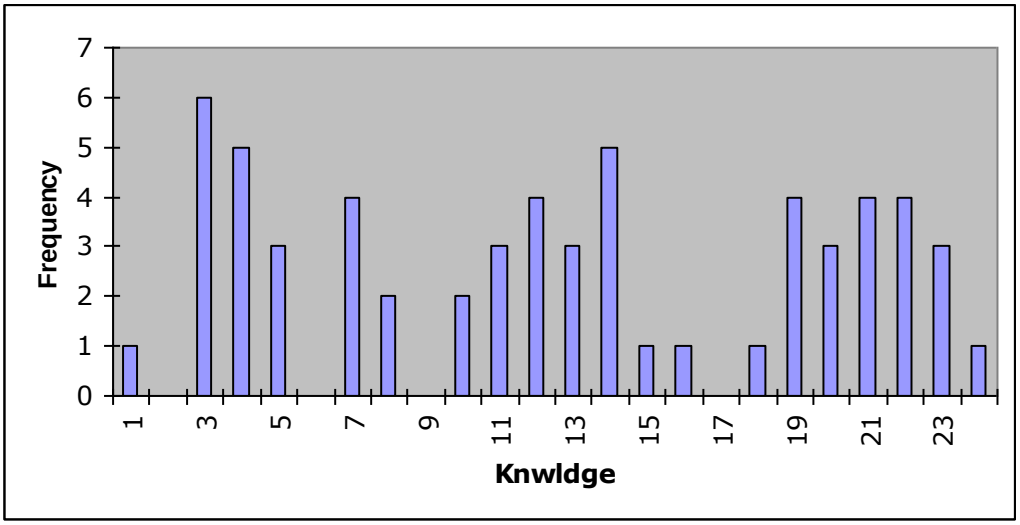
The column headed "z" shows the z-score corresponding to the numeric value found in the adjacent "score" column. The statistics required to calculate z-scores, namely the mean and standard deviation of all scores, are taken from the bottom of the Scores worksheet.

The "cf" column shows the cumulative frequency associated with each score, that is, the number of scores at and below the given score, divided by "n", the total number of scores. The value of n is taken from the bottom of the Scores worksheet.

The symbols used to "plot" the frequency bars change, becoming thinner as the bars become longer. If the frequency, "f", of any score exceeds 200, the bars are rescaled so that each symbol represents more than one case. This maximum length of 200 may be changed—see the section below on the System worksheet.

Lertap2-style histograms are provided in worksheets called, for example, Histo1L.

If users have installed an Excel option, the Analysis ToolPak, Lertap gets Excel to make the Histo1E worksheet, into which it places an Excel chart, such as this one:



Charts like this one are based on an Excel-created "Bin / Frequency" table which appears in the first two columns of the Histo1E sheet, immediately to the left of the chart. The table looks like this:

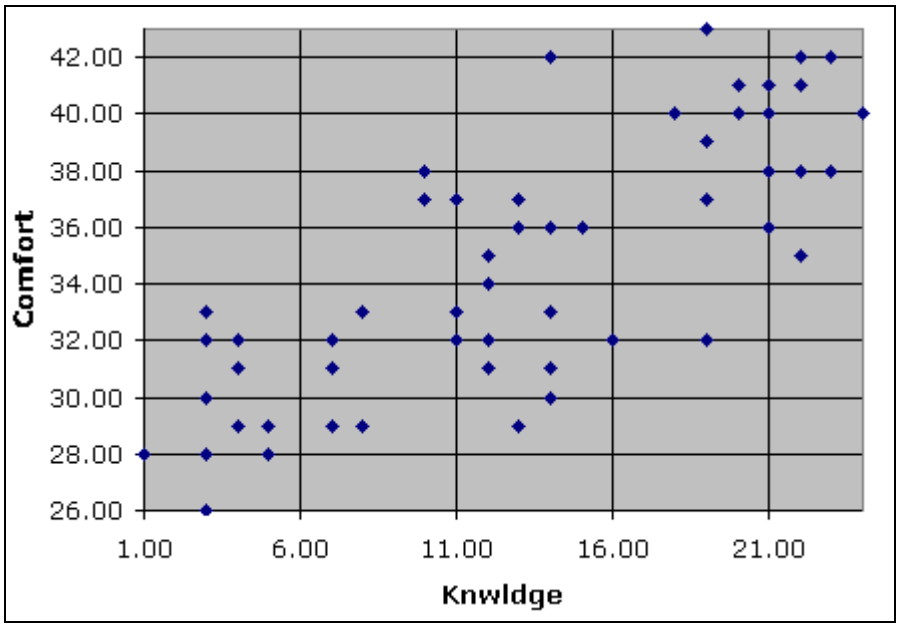
Bin	*Frequency*
1	1
2	0
3	6
4	5
5	3

This table is live—change one of its values and the chart will also change. Because of this, the table cannot be deleted. However, the chart may be dragged over the table, and the table may be shifted to the right of the chart, if desired.

The Analysis ToolPak is supplied with Excel, but has to be added in before it will work. In fact, it's called an "Add-in"—see Excel's Help for instructions.

Scatterplots

Excel has many built-in charting routines, and "XY (Scatter)" is one of the easiest to use. Lertap's scatterplot icon asks users to nominate the two scores to be plotted, and then gets Excel to make an XY (Scatter) chart from the corresponding columns of the Scores worksheet.



Lertap uses the Min and Max figures from the bottom of the Scores sheet to get Excel to start and end its X and Y axes at corresponding scores. Users may alter the way the scatterplot looks—there are *many* options which may be applied via Excel's Chart menu.

The Analysis ToolPak is not needed for charts such as this one.

External criterion statistics

It is possible to have the items of any subtest correlated with any of the scores in the Scores worksheet. The option for doing this is found under the Run menu on the Lertap toolbar.

Item 30

option	wt.	n	p	pb/ec	avg/ec	z
1	5.00	9	0.15	0.43	19.78	1.03
2	4.00	20	0.33	0.18	14.45	0.26
3	3.00	17	0.28	-0.04	12.24	-0.06
4	2.00	8	0.13	-0.38	5.88	-0.97
5	1.00	5	0.08	-0.35	4.60	-1.16
other	3.00	1	0.02	0.01	13.00	0.05
			r/ec:	**0.63**		

The table above shows the figures which resulted by having Lertap correlate each of the items on an affective subtest, "Comfort", with the "Knwldge" score found in the Scores worksheet.

The pb/ec column gives the point-biserial correlation for each response with the criterion (Knwldge). The avg/ec column indicates the average criterion score for all those respondents who selected each response option. For example, the average criterion score for the 9 respondents who selected option 1 on Item 30 was 19.78. The "z" value of 1.03 is 19.78 expressed as a z-score. (The average criterion score was 12.63, standard deviation 6.95.)

The **r/ec** figure is the product-moment correlation between the item scores and the criterion scores.

External criterion statistics for cognitive items are identical to those given for affective items. In the case of cognitive items having one correct answer, the r/ec figure will equal the pb/ec value found for the item's correct answer.

External criterion reports are given in worksheets with names such as ECStats1f, ExStats2f, and so forth.

After the item statistics are given, summary statistics for the external criterion are provided. These should be checked with the same statistics which appear at the bottom of the respective Scores column—the two sets of statistics should be identical (if not, an error has occurred—deleting or changing records in the Data and Scores worksheets without going all the way back to the "Interpret CCs lines" option, for example, will cause errors).

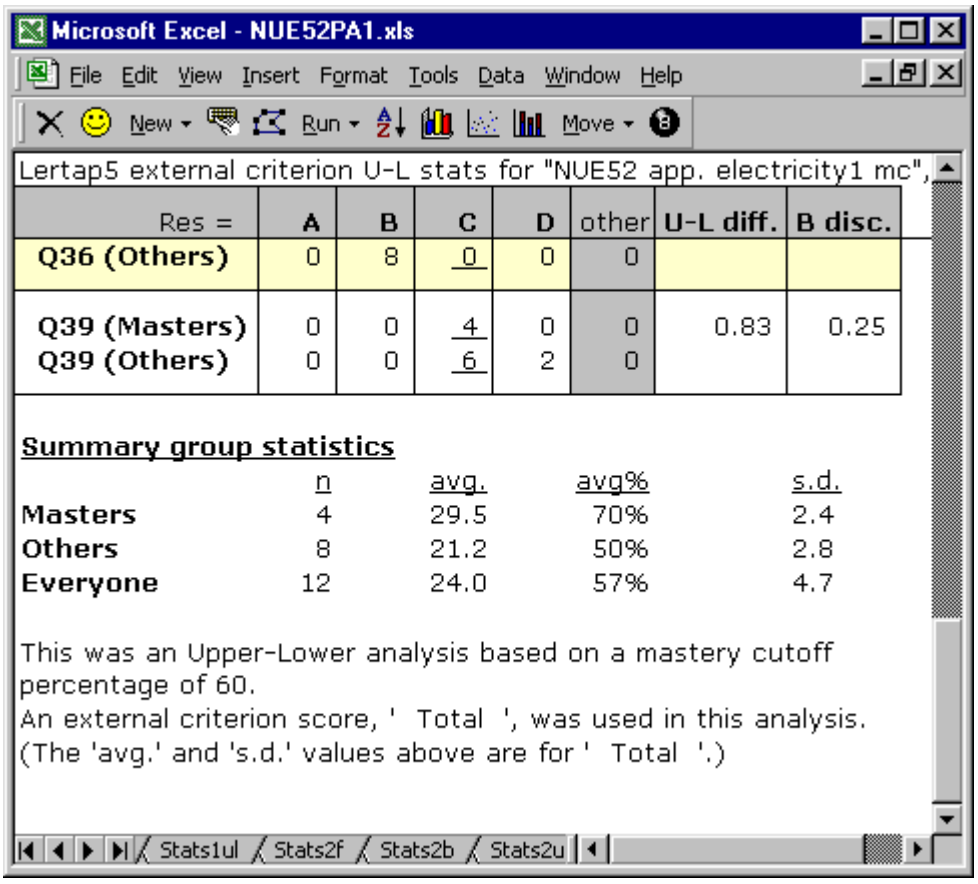
correlation bands (with external criterion)

.00:
.10: Item 4
.20: Item 5 Item 22
.30: Item 7 Item 14 Item 23 Item 24
.40: Item 3 Item 8 Item 9 Item 10 Item 11 Item 12 Item 15 Item 17 Item 18
.50: Item 1 Item 16 Item 20
.60: Item 6 Item 13 Item 19 Item 21 Item 25
.70: Item 2
.80:
.90:

These correlation bands appear at the very end of the ECStats1f worksheet. They summarise the r/ec values for the items. If an item's r/ec value is negative, the item will be listed in the **.00** band.

External statistics for U-L analyses

The sample screen snapshot below exemplifies the report format seen when an external criterion score is used with a U-L analysis:



Microsoft Excel - NUE52PA1.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 external criterion U-L stats for "NUE52 app. electricity1 mc",

Res =	A	B	C	D	other	U-L diff.	B disc.
Q36 (Others)	0	8	0	0	0		
Q39 (Masters)	0	0	4	0	0	0.83	0.25
Q39 (Others)	0	0	6	2	0		

Summary group statistics

	n	avg.	avg%	s.d.
Masters	4	29.5	70%	2.4
Others	8	21.2	50%	2.8
Everyone	12	24.0	57%	4.7

This was an Upper-Lower analysis based on a mastery cutoff percentage of 60.
An external criterion score, ' Total ', was used in this analysis.
(The 'avg.' and 's.d.' values above are for ' Total '.)

Stats1ul Stats2f Stats2b Stats2u

In cases such as this, the upper and lower groups are defined with reference to the external criterion. Here (above) the external criterion was a score called 'Total', and the Masters group was defined as those having a score of 60% or better on this criterion.

Item scores matrix

An option on Lertap's Run menu will produce matrices of item scores and intercorrelations for any subtest. An example of item scores is shown here:

Lertap5 IStats matrix, last updated on: 24/08/00.

ID	Item 26	Item 27	Item 28	Item 29	Item 30
9	2.00	3.00	5.00	4.00	4.00
31	3.00	3.00	2.00	3.00	3.00
26	4.00	4.00	4.00	5.00	3.00
27	2.00	2.00	3.00	5.00	4.00
21	2.00	2.00	3.00	5.00	4.00
59	4.00	3.00	4.00	4.00	4.00

This example is from an affective subtest. It displays the "score", or "points", earned by each respondent on each item of the subtest. When the subtest is a cognitive one, this matrix will usually consist of zeros and ones, as shown below:

Lertap5 IStats matrix, last updated on: 24/08/00.

ID	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
9	0.00	0.00	0.00	0.00	0.00	1.00	0.00
31	0.00	0.00	1.00	1.00	0.00	1.00	1.00
26	0.00	1.00	0.00	1.00	1.00	1.00	0.00
27	1.00	1.00	0.00	1.00	1.00	0.00	0.00
21	1.00	1.00	1.00	0.00	1.00	0.00	0.00
59	0.00	1.00	1.00	1.00	1.00	1.00	1.00

Summary item statistics and intercorrelations are given at the bottom of the IStats worksheet, and have the format shown here:

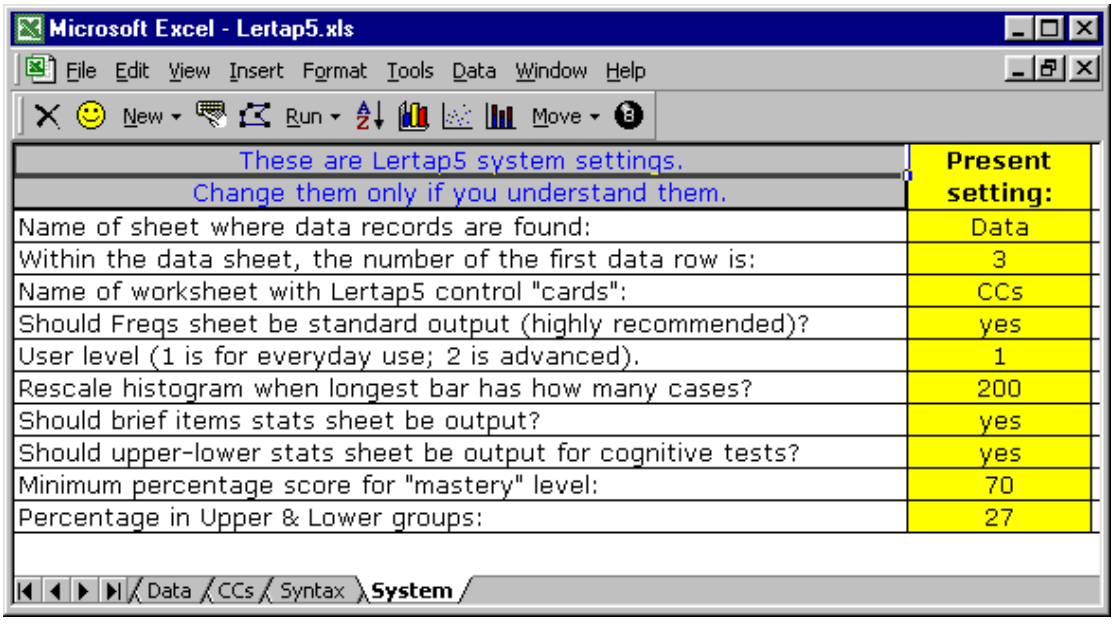
n	60	60	60	60	60
Min	1.00	2.00	2.00	2.00	1.00
Median	3.00	3.00	4.00	4.00	3.00
Mean	3.08	2.98	3.75	3.93	3.33
Max	5.00	5.00	5.00	5.00	5.00
s.d.	1.28	0.88	0.94	0.89	1.14
var.	1.64	0.78	0.89	0.80	1.29
Correlations					
Item 26	1.00	0.57	-0.02	0.41	0.45
Item 27	0.57	1.00	0.17	0.23	0.34
Item 28	-0.02	0.17	1.00	-0.14	-0.12
Item 29	0.41	0.23	-0.14	1.00	0.37
Item 30	0.45	0.34	-0.12	0.37	1.00
Item 31	-0.01	0.14	-0.05	0.01	-0.04
Item 32	0.24	0.06	-0.07	0.27	0.04
Item 33	0.74	0.45	-0.19	0.43	0.55
Item 34	-0.41	-0.34	-0.02	-0.31	-0.36
Item 35	0.66	0.34	-0.20	0.36	0.51
average	*0.29*	*0.22*	*-0.07*	*0.18*	*0.19*

These statistics are all produced by Excel. Respective equations may be seen, in Excel, by clicking on any of the statistics cells, and then examining the contents of Excel's Formula Bar.

The *average* item correlation figure, shown in the last row, is the mean of an item's correlations with the other items in the subtest.

The System worksheet

One of the five worksheets in the Lertap5.xls file is a hidden one called System.



The System sheet contains settings which are used as parameters for Lertap's operations. All of these settings may be changed; however, changing one of the first three settings ("Data", "3", "CCs") is *not* recommended at all, and may result in unpredictable consequences.

There are three "yes" settings, each referring to the creation of one Lertap's reports. If these are changed to "no", the corresponding report (worksheet) will not be created.

The user level setting determines the number of icons which display on the Lertap toolbar—see the following section for more details.

The rescale histogram setting makes reference to the length of the bars produced in the Lertap2-style histogram known as "Histo1L". In large data sets these bars can become too long, going off the screen, and off the printer. Whenever a single bar becomes longer than 200 symbols, Lertap increases the number of cases represented by the symbol so that the bar shortens. For example, if the frequency corresponding to a particular test score is, say, 800, Lertap will divide this value by 200, and let each bar symbol represent four (4) cases.

How to access the System worksheet? It's a hidden sheet in the Lertap5.xls workbook. The workbook is protected with a password of "shack25"—use Excel's Tools / Protection option to Unprotect Workbook, type in the password without the quotation marks, and then use the Format / Sheets option to Unhide System.

Advanced level toolbar

One of the settings in the System worksheet has to do with what is called the "user level". If this setting is changed from its default value of 1 (one) to 2, a different Lertap toolbar results[45]:

This toolbar differs from the standard one in the two icons seen immediately to the right of the Run option. The first of these icons toggles hide / unhide for a workbook's Sub sheets, sheets which are normally hidden. This icon makes it easy to unhide all of them with just one click.

The second icon, one we call the "Liberty Bell", gets Elmillon to act on a single Sub sheet, as opposed to all of them. Normally, when the "Elmillon item analysis" option is taken, all of a workbook's Sub sheets are read, and scores and reports are produced for each. Use of the Liberty Bell makes it possible to act on just one selected Sub sheet.

The idea here is to give users the power to make alterations to their subtests at the most basic level. To investigate the effects of different item response weights, for example, a user could use Excel to copy a given Sub worksheet, go to the copy, change response weights, and then apply the Liberty Bell to get

[45] The workbook has to be saved and then re-opened before the change will become effective.

results. For some users this will be more convenient than having to go all the way back to the CCs sheet, add new control cards, and then go through the "Interpret CCs lines" and "Elmillon item analysis" options again.

Readers might be surprised to hear that they can get Lertap to do its things without having to create lines in a CCs worksheet. The program which does the real analysis work for Lertap is Elmillon. Elmillon does not read CCs lines—it looks for Sub sheets, from which it derives all of its information. If users can create their own Sub sheets, using the right format, they need not be concerned about lines of job definition statements in the CCs file.

Exporting worksheets

An option on Lertap's toolbar will take any of the worksheets in the currently-active workbook, and turn them into a single Excel 4 worksheet. This option is provided as Excel 4 worksheets are quite easy to import in other programs, such as SPSS, something which cannot always be said of worksheets from Excel 95, 97, 98, and 2000. For example, version 9 of the SPSS package will hiccup a bit if asked to import from Excel 97—users generally have to work through an ODBC (open data base connectivity) interface, which is not overly straightforward. (It seems that version 10 of SPSS may have fixed this problem, but read on.)

The Lertap worksheets which users may want to convert to Excel 4 format will usually be Data, Scores, or IStats. These worksheets have their two top-most rows reserved for titles and headers. A program such as SPSS will want to see just one such row, so Lertap's Excel 4 converter will strip the first row from the Data, Scores, and IStats worksheets' leaving just the row of column headers, which SPSS is happy to see (it uses them as variable names—when opening an xls file from within SPSS, be sure and tick the SPSS box which says Read variable names). Lertap's Excel 4 converter will also strip the statistics rows from the bottom of the Scores worksheet.

Time trials

In October 2000 we selected five data sets and ran them through Lertap 5 on three quite different computers. The results reported below should be seen as relative ones—the system used in October was not the final version.

The data sets

"LRTP Quiz" is the standard data set which comes built into the Lertap5.xls file. It has 60 respondents and two subtests. The first subtest is a cognitive one, with 25 items, while the second is an affective one with 10 items.

"Ed 502" is typical of the type of results which many teachers collect. It's from an end-of-semester test given to 48 graduate students. It had one subtest, a cognitive one with 63 items.

"Nanta" is the name of a data set from Dr Nanta Palitawanont of Burapha University, Thailand. It involved 300 students responding to an affective instrument with 10 scales, with an average of seven items per scale.

"UCV 1" is from la Universidad Central de Venezuela, and involved 1,494 students responding to three cognitive tests, each with 20 items[46].

"UCV 2" is from the same university. This data set had results from 11,190 students who sat an entrance exam having two 25-item cognitive subtests.

The computers

"LeoPAD1" was a generic laptop computer running a Pentium MMX CPU, 166 MHz clock, and 32 M RAM. It was running Excel 2000 under Windows 98. This computer would, these days, be considered an "old" Pentium.

"G3" was a Macintosh G3 computer, running at 350 MHz, and having 132 M RAM. G3 was using Excel 98. Readers should note that we're not experienced Mac users. The G3 did very little Lertap work for us until we increased Excel's memory slice to 50 K (which seemed a small amount, but got the job done; a larger memory chunk may have resulted in better performance).

"Marek" was a Pentium III machine with 128 M RAM, running at 450 MHz. It was running Excel 2000 and NT 4.0.

The results

We present results for three levels of processing, "Int. CCs", which means "Interpret CCs lines", "Elmlln", which refers to "Elmillon item analysis", and "No U-L", which means a job with U-L analyses disabled. We included the latter level as it is likely some users will turn the U-L analysis off; we knew it added to the processing burden, and results confirm this.

In the table, "secs." means seconds, and "mins." means minutes. The reason "n.a." is given for the Nanta data set has to do with the fact that this data set involved only affective subtests; U-L analyses are only relevant to cognitive tests.

[46] Thanks to Carlos Gonzalez of UCV for making these data sets available for our benchmarking tests.

	LeoPAD1	G3	Marek
LRTP Quiz			
Int. CCs	30 secs.	12 secs.	3 secs.
Elmlln	63 secs.	23 secs.	8 secs.
No U-L	34 secs.	14 secs.	5 secs.
Ed 502			
Int. CCs	28 secs.	13 secs.	5 secs.
Elmlln	81 secs.	22 secs.	9 secs.
No U-L	50 secs.	13 secs.	5 secs.
Nanta			
Int. CCs	3 mins.	1 min.	18 secs.
Elmlln	7 mins.	2 mins.	37 secs.
No U-L	n.a.	n.a.	n.a.
UCV 1			
Int. CCs	7 mins.	2 mins.	28 secs.
Elmlln	9 mins.	4 mins.	53 secs.
No U-L	6 mins.	2 mins.	41 secs.
UCV 2			
Int. CCs	33 mins.	10 mins.	3 mins.
Elmlln	45 mins.	24 mins.	5 mins.
No U-L	31 mins.	19 mins.	4 mins.

It is likely readers will have a variety of reactions to these performance figures. Some may be surprised that the analyses can take minutes to complete in some cases. Others, perhaps more experienced data analysts, may marvel that data sets with more than ten thousand cases can be processed so quickly on a desktop. It used to be that we'd have to carry five boxes of punch cards to the "computing centre", or a magnetic tape, and generally plan to spend at least a couple of hours waiting for results.

Chapter 11

A History of Lertap

The first version of Lertap was developed for the Venezuelan Ministry of Education in the years 1971 through 1972. It was called "DIEitem", with DIE referring to el Departamento de Investigaciones Educacionales.

At the time, the Ministry was embarking on a national assessment program, with emphasis on mathematics and language achievement. The Kuhlmann-Anderson aptitude, or "IQ", tests were also used on a national scale by the Ministry, and a general-purpose item analysis program was required, one which could handle conventional achievement tests, and the Kuhlmann-Anderson forms.

The development of the Ministry's assessment centre was under the direction of Rogelio Blanco, with Richard Wolfe, of OISE (Ontario), overseeing the technical services part of the operation. Richard created a general front-end to set up data sets for subsequent analyses, using the PL/I programming language. The first version of Lertap, programmed in FORTRAN II, picked up data sets pre-processed by the PL/I program, and output classical item statistics. The first Lertap could not only handle the idiosyncrasies of the Kuhlman-Anderson tests, but could also entertain multiple tests within the same data set. Thus one could submit a data set with results from the mathematics test, the Spanish-language test, and the Kuhlmann-Anderson forms, all strung together in a lengthy input string.

Work on the first Lertap was supported by the Ford Foundation, and by the Organization of American States.

In 1973 work on the second version began at the University of Colorado, home of the Laboratory of Educational Research. The PL/I front end was replaced by another, written in FORTRAN, and featured the use of a set of free-form control cards to describe a job. These control cards were the forerunners of those seen in the latest version of the software, described in Chapters 4, 5, and 6 of this book[47].

At the time, free-form control cards were not at all common, and, in this regard, Lertap 2 could be considered as being slightly ahead of its time. In 1973 the SPSS system had not yet appeared—many universities used the "BMD" series from the Health Sciences Computing Facility of University of California at Los Angeles (UCLA).

Lertap 2 also introduced support for processing affective tests. Bob Conry of the University of British Columbia provided strong support for the "aff" subtest capability, while Ken Hopkins and Gene Glass, at LER in Boulder, supported and encouraged the development of the overall package.

---

[47] The cards used in version 2 were almost identical to those seen in the new version.

The work started at LER was transferred to the University of Otago, in Dunedin, New Zealand, late in 1973. By the end of 1974 Lertap 2 was stable, and in use in a variety of centres in Canada and the United States.

Lertap 2's development was supported by several people at Otago, especially Dan McKerracher, Department of Education, and Brian Cox, Computing Centre. The user guide which emerged from Otago, the *Guide to Lertap Use and Interpretation*, was widely circulated, and use of the system grew steadily in the 70s.

The development of a microcomputer version, to be called Lertap 3, began in earnest at Otago in 1980, using a CP/M card on an Apple II computer, and then on an Osborne 1 system. Pascal and BASIC 80 were used to write initial modules, with everything eventually translated to BASIC 80 as it had a clear performance edge. The first working version was comprised of a series of interlinked modules which would load and unload themselves in just 56K of core memory.

By 1983 a reliable version of Lertap 3 was ready, accompanied by a comprehensive user guide. Barbara Calvert keenly supported the development of this version, and the resources of Otago's Department of Education stood behind the effort. A few hundred copies of the user guide were printed, and made ready for distribution.

At this time, however, the IBM Corporation decided to produce a microcomputer of its own. By late 1983 Lertap 3 had been altered so as to operate within IBM microcomputers, and National Computer Systems of Minneapolis had purchased non-exclusive rights to market it. NCS repackaged Lertap 3 as two stand-alone programs, MicroTest1, and MicroSurvey1. The work required to get the IBM version ready was partially supported by Hans Wagemaker of New Zealand's Department of Education, by Evelyn Brzezinsk of the Northwest Regional Educational Laboratory, Portland, and by Larry Erikson of National Computer Systems.

By the late 80s the Otago version of Lertap 3 was in use in many sites, with the NCS versions finding a home in many (many) more. It was unfortunate that a user guide for this version was never thoroughly developed. The guide printed at Otago covered the pre-IBM version, but the operation of this version differed much, and IBM users found it to be of limited use.

In 1987 a series of brief Lertap 3 user help sheets were circulating from Curtin University of Technology in Perth, Western Australia. These were later assembled as a small book, *Lertap 3 General Notes*, printed by Curtin University's Printing Services.

Lertap 3 remains a potent data analysis system. The scope of analyses it supports includes those related to cognitive and affective tests, general surveys, and classroom gradebooks. Its data preparation facilities include a module for complete date entry verification. And, it can handle results from the Kuhlmann-Anderson tests. (Not that they're used that much anymore, but Form B of the K-A is complex, making more exacting demands of a test-analysis program.)

In 1992 Piet Abik translated Lertap 3 to the Indonesian language, and it was later purchased by Indonesia's Ministry of Education and Culture for country-wide use in secondary schools, with the support of Bambang Irianto.

Lertap 2, Lertap 3, and the NCS versions came to be used throughout the world.

When Microsoft released the Windows 3 operating system, in 1992 (in Australia), it was clear that Lertap had to move to Windows. Users started to write to ask when the Windows version would be ready.

Work on Lertap 4 began in 1993, but was never finished. It came to have a facility for processing survey results, but not much more. It was clear that the production of a Windows version would require a dedicated block of time, and at least two programmers, if it was ever to get off the ground. It never did.

In 1999 Curtin University's Division of Humanities approved an application from the senior author to devote an extended sabbatical period to the development of an Excel-based Lertap system. The feasibility of using Excel as a serious developmental platform was obviously quite on—several small Excel-based engineering systems had emerged, and Prentice Hall had released the Excel-based "PHStat" system for business managers (Levine, Berenson, & Stephan, 1999).

So it was that Lertap 5 was born. Excel, and Visual Basic for Applications, proved more than equal to the task. And, while switching to an Excel base was obviously of enormous benefit, a decision to build the new system on the control "card" syntax seen in the second version was another telling factor behind the new system's relatively rapid emergence.

Appendix A

The Original (1973) Quiz on LERTAP 2

LERTAP 2 Quiz, Version 1.

(1) Which of the following is not included in standard LERTAP output?

a) individual scores.
b) subtest histograms.
c) correlation among subtest and total test scores.

(2) What control word is used on which control card to activate the correction-for-chance scoring option?

a) MDO on *ALT.
b) CFC on*TST.
c) WT on*SUB.
d) MDO on*FMT.
e) CFC on *SUB.

(3) Which of these control cards is not used by affective subtests[48]?

a) *FMT
b) *POL
c) *KEY
d) *MWS

(4) The minimum number of control cards used for any subtest is two.

a) true
b) false

(5) The RES control word refers to

a) subtest score rescaling.
b) the number and type of response codes used by the items of each subtest.
c) special treatment for residual (i.e. missing and invalid) item responses.

[48] This quiz covered Lertap 2. The *FMT card is no longer used, having been replaced by the *COL card in Lertap 5.

(6) In order to indicate the presence of an external criterion[49]

a) use the CODES/COLS specification on a *TST card.
b) use an E control letter on the first *FMT card.
c) employ the EXT control word on the *SUB card.

(7) Which of the following control letters is used on the *SUB card to indicate the presence of ID characters[50]?

a) A
b) I
c) C
d) P
e) none of the above.

(8) To inform LERTAP that a subtest is of the affective type, one must

a) use a *FMT card.
b) employ the AFF control word on the *SUB card.
c) indicate forward and reverse item weighting on the *POL.
d) multiply-weight every item with *MWS cards.

(9) With regard to precoded subtests, which of these statements is <u>not</u> true[51]?

a) they may be dichotomous variables.
b) they are always included in total test score computation.
c) they require both *FMT and *SUB control cards.
d) they may be variables with real or implied decimal points.
e) they are indicated by the presence of a P control letter on a *FMT card.
f) they are always plotted ("histogrammed") and included in the correlation matrix output by the program.

(10) *MWS control cards are used

a) for achievement items to multiply-weight the item (indicate that more than one response is to receive a non-zero weight), or to indicate that not all response codes are used by an item.
b) for affective items to alter the pattern of forward or reverse weighting established by preceding control cards, or to indicate that not all response codes are used by an item.
c) to override item control information entered on any or all previous control cards.
d) all of the above.

---

[49] External criteria were handled much differently in Lertap 2.
[50] This was a trick item, requiring careful reading. It has no relevance to Lertap 5.
[51] The term "precoded subtests" is no longer used in Lertap.

(11) What is the error in the following control card?

*FMT (A3,X,15I,T52,E4)

a) the A control letter has an invalid suffix.
b) the X control letter is in the wrong position.
c) the suffix to the E must be of the general form n.d.
d) a T control letter must always have a prefix and can never have a suffix greater than 1.
e) the format declaration is missing a P control letter.

(12) Which of the following is not a default action taken by LERTAP (a default action is that which always occurs unless you use a control word or letter to indicate otherwise)?

a) RES=(1,2,3,4,5)
b) each subtest score is the sum of the weights awarded each test respondent for his response to each subtest item.
c) WT=1.000
d) total test score is the sum of all subtest scores plus any external criterion value (if it exists).

Study the following control cards and then answer questions 13 - 18.

*FMT (10X,20I)
*KEY 54331 41642 34551 53222

(13) The subtest described by these control cards is of what type?

a) precoded.
b) achievement.
c) affective.

(14) These two control cards are sufficient to initiate an analysis

a) true
b) false

(15) How many response codes are used by the items of this subtest?

a) six
b) five
c) ten
d) impossible to tell.

(16) The twenty subtest items occupy which columns of the data cards?

a) 10 - 20
b) 11 - 30
c) 20 - 39
d) impossible to tell.

(17) The *KEY card has an error. What is it?

a) there is an insufficient number of keys.
b) blanks are not allowed to appear in the key string.
c) one item has an invalid key.

(18) Individual scores will be output for this subtest

a) yes
b) no
c) can't tell, need more information.

Study the following control cards, and then answer questions 19 - 20.

```
*FMT  (23I,T77,3A)
*SUB  AFF,  MDO,  RES=(A,B,C,D,E,F),  WT=.5
*POL  ++-++  -+++-  ---+-  ++-+-  +++
*ALT  33244  55555  55556  55455  566
*MWS  23,4,3,2,1,*,*
*MWS  21,*,1,2,3,4,5
```

(19) The 3A on the *FMT card

a) indicates the presence of ID characters on the data cards.
b) activates the request for individual scores output.
c) both (a) and (b).
d) should be A3 instead.

(20) Which of the following statement is <u>not</u> true?

a) the six control cards are in the proper order.
b) item 5 uses response codes (A,B,C,D) which will receive weights (4,3,2,1), respectively.
c) item 21 uses response codes (B,C,D,E,F) with respective weights (1,2,3,4,5).
d) item 23 used just four of the six response codes given in the RES specification.

(21) What is the problem with this control card?

*SUB NAME=(CEREAL PREFERENCE SCALE), MDO

a) no RES declaration is given.
b) the MDO control word may not appear without the AFF control word.
c) both of the above.
d) there is no problem.

(22) Which of the following is <u>not</u> true?

a) the *WGS card is used to multiply-weight items.
b) *WGS cards may not be used with affective subtests.
c) *MWS cards may be used instead of a *WGS card.
d) *MWS cards may be used with a *WGS card.
e) unless *WGS or *MWS cards specify otherwise, the keyed-correct response for each achievement item will receive a weight of 1.

(23) All but one of these control cards may be continued when there is insufficient space on one card to enter all control data. Which is it?

a) *KEY
b) *SUB
c) *ALT
d) *POL
e) *WGS

(24) The following statements concern the *TST card. One of them is not true. Which one is it?

a) the *TST card is required in multiple test runs.
b) *TST is used to specify selection criteria when not all data records are to be accepted for input.
c) *TST may be used to supply a test name which will label the output.
d) when an external criterion is present, it must be indicated on the *TST card.

(25) What is the problem with this set of control cards?

```
*TST  NAME=(PAEDIATRICS  503  EXAM,  1975)
*FMT  (15I,T78,2A)
*SUB  NAME=(INFANT PSYCHIATRY),  RES=(1,2,3,4)
*KEY  34322  23515  54332
*FMT  (14X,20I,T52,E4,T78,2A)
*SUB  NAME=(PRENATAL CARE)
*KEY  43321  14432  32455  13455
```

a) the control cards are not in the correct order.
b) the specification of the ID character is superfluous on all but the first *FMT card.
c) the external criterion specification does not appear on the first *FMT card.
d) one item belongs to both subtests.
e) no *WGS cards are present.
f) the user failed to specify the response codes used by the items of the second subtest.

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.	__	__	__	__	__

Finally, please respond to these two questions:

(36) Number of years I have been using computers or computer programs. _____

(37) Number of years I've been using tests in my research or my teaching. _____

Thank you.

References

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory.* Monterey, California: Brooks/Cole.

Berk, R.A. (1984). Selecting the index of reliability. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction.* Baltimore, Maryland: The Johns Hopkins Press.

Brennan, R.L. (1972). A generalized upper-lower discrimination index. *Educational and Psychological Measurement,* 32, 289-303.

Brennan, R.L. (1984). Estimating the dependability of the scores. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction.* Baltimore, Maryland: The Johns Hopkins Press.

Brennan, R.L. & Kane, M.T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement,* 14, 277-289.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

de la Harpe, B.I. (1998). *Design, implementation, and evaluation of an in-context learning support program for first year education students and its impact on educational outcomes.* Perth, Western Australia: unpublished doctoral dissertation, Curtin University of Technology.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of Educational Measurement* (4th ed.). Sydney: Prentice-Hall of Australia.

Frederiksen, N., Mislevy, R.J., & Bejar, I.I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Garrett, H.E. (1952). *Testing for teachers.* New York: American Book Company.

Glass, G.V & Stanley, J.C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Glass, G.V & Stanley, J.C. (1974). *Metodos estadisticos aplicados a las ciencias sociales.* London: Prentice-Hall Internacional.

Green, J. (1999). *Excel 2000 VBA programmer's reference*. Birmingham, England: Wrox Press.

Gronlund, N.E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Collier Macmillan Publishers.

Gulliksen, H. (1950). *Theory of mental test scores.* New York: John Wiley & Sons.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications.

Hays, W.L. (1973). *Statistics for the social sciences.* London: Holt, Rinehart and Winston.

Hills, J.R. (1976). *Measurement and evaluation in the classroom.* Columbus, Ohio: Charles E. Merrill.

Hopkins, K.D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Hopkins, K.D. & Glass, G.V (1978). *Basic statistics for the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.

Hopkins, K.D., Stanley, J.C., & Hopkins, B.R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kaplan, R.M. & Sacuzzo, D.P. (1993). *Psychological testing: principles, applications, and issues.* Pacific Grove, California: Brooks/Cole.

Kelly, T.L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Kerlinger, F.N. (1973). *Foundations of behavioral research* (2nd ed.). London: Holt, Rinehart, and Winston.

Levine, D.M., Berenson, M.L., & Stephan, D. (1999). *Statistics for managers using Microsoft Excel* (2nd ed.). London: Prentice-Hall International.

Lindeman, R.H. & Merenda, P.F. (1979). *Educational measurement* (2nd ed.). London: Scott, Foresman and Company.

Linn, R.L. & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley.

Magnusson, D. (1967). *Test theory.* London: Addison-Wesley.

Mehrens, W.A. & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). London: Holt, Rinehart and Winston.

Nelson, L.R. (1974). *Guide to LERTAP use and interpretation.* Dunedin, New Zealand: Department of Education, University of Otago.

Online Press, Inc. (1997). *Quick course in Microsoft Excel 97.* Redmond, Washington: Microsoft Press.

Oosterhof, A.C. (1990). *Classroom applications of educational measurement.* Columbus, Ohio: Merrill.

Pedhazur, E.J. & Schmelkin, L.P. (1991). *Measurement, design, and analysis: an integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pintrich, P.R., Smith, D.A.F., Garcia, T., & McKeachie, W.J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor, Michigan: the University of Michigan.

Popham, W.J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Sanders, D.H. (1981). *Computers in society.* New York: McGraw-Hill.

Subkoviak, M.J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265-276.

Subkoviak, M.J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction.* Baltimore, Maryland: The Johns Hopkins Press.

Wiersma, W. & Jurs, S.G. (1990). *Educational measurement and testing* (2nd ed.). Boston: Allyn & Bacon.

Index