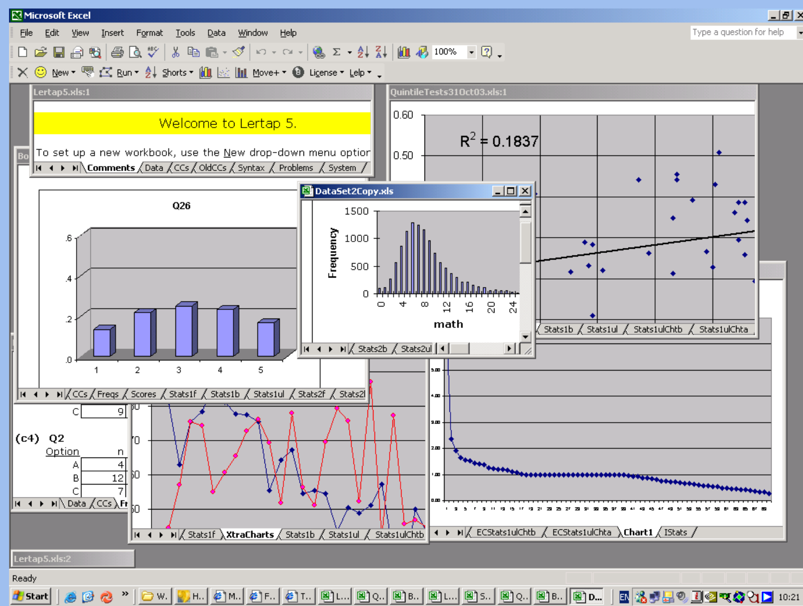


Datasets and Exercises

For use with Lertap 5

Interactive PDF version

Made in Western Australia

Table of Contents

1 Introduction

Lertap is a short name for the Laboratory of Educational Research Test Analysis Package.

The present Lertap, "**Lertap5**", is a system which works as an application running within Microsoft's **Excel** program.

> To get a quick idea of what Lertap5 does, move forward and select from our smorgasbord of samples.
>
> To have some immediate hands-on fun, jump right into the action part of the Cook's Tour.

The main purpose of this website is to provide paths to sample datasets for use with Lertap.

The "Cook's Tour" of Lertap5 is also found here, in this website. It has a description of the **Lertap5.xlsm** workbook, and its five worksheets: Comments, Data, CCs, System, and Syntax. The Cook's Tour also gets into an actual run of Lertap5, with steps to follow on your own computer to keep from falling asleep.

A bit of history, and some important links to other resources:

Lertap dates back to 1972, when it made its first appearance at the University of Colorado in the United States. It was written in the FORTRAN computer language, and ran on mainframe computers generally situated in a special, single-purpose building, often called the "Computing Centre". Mainframe computers were very large, usually occupying at least half of one of the building's floors.

When microcomputers became popular in the 1980s, versions of Lertap were made to run on them. Most of these were written in the BASIC computer language.

In the late 1990s a new genre, "**Lertap5**", was created to run as an **Excel** application. This version was made at Curtin University in Western Australia. It is written in **VBA**, Visual Basic for Applications.

Excel is a spreadsheet program created by Microsoft. It's a member of the suite of programs known as "Microsoft Office", a collection of programs which includes Word, PowerPoint, Outlook, Access, and, of course, Excel.

Numerous versions of Excel have emerged over the years. As of May 2020, editions of Lertap5 were available for Excel 2010 (Windows), Excel 2016 (Windows and Macintosh), and Excel 2019 (Windows and Macintosh). **Note**: Excel 365 is a version code used by Microsoft, applied to users who have taken out an annual subscription on Windows and macOS computers. These users automatically get frequent updates to ensure they have the very latest version of Excel.

The Excel 2010 version of Lertap5 was used to create the charts and reports in this document. Had one of the other editions been used, the results would have been much the same.

Here are some links to other Lertap resources:

1 A PDF copy of this website's topics. A CHM copy (compiled help file for Windows). An iBook copy, ready for reading on an iPad or an iPhone. A link to the website itself.

2 A small set of PowerPoint slides with a quick introduction to Lertap5. These are also available as a PDF file.

3 The main Lertap5 website. Has more examples and samples, with links to videos, the manual, and a variety of *riveting* technical papers (also known as "erudite epistles").

4 Some "Tips & Tricks" for users. Demonstrates the use of quintile plots; showcases selected Excel and Lertap features.

5 The online help system for Lertap5. A primary source for finding out how to obtain Lertap5, how to get it running, and understanding features added after the manual was printed.

6 The QUIA website, our developmental site. At times has a variety of special tidbits and morsels, especially for instructors and students. (No longer updated but may still be useful.)

7 The e-store for Lertap5, the place which sells licenses for Lertap5 users when they have more than 250 cases to process.

Please direct questions or comments to: lertap5@gmail.com

Last update: **2 March 2023**

1.1 How to get Lertap

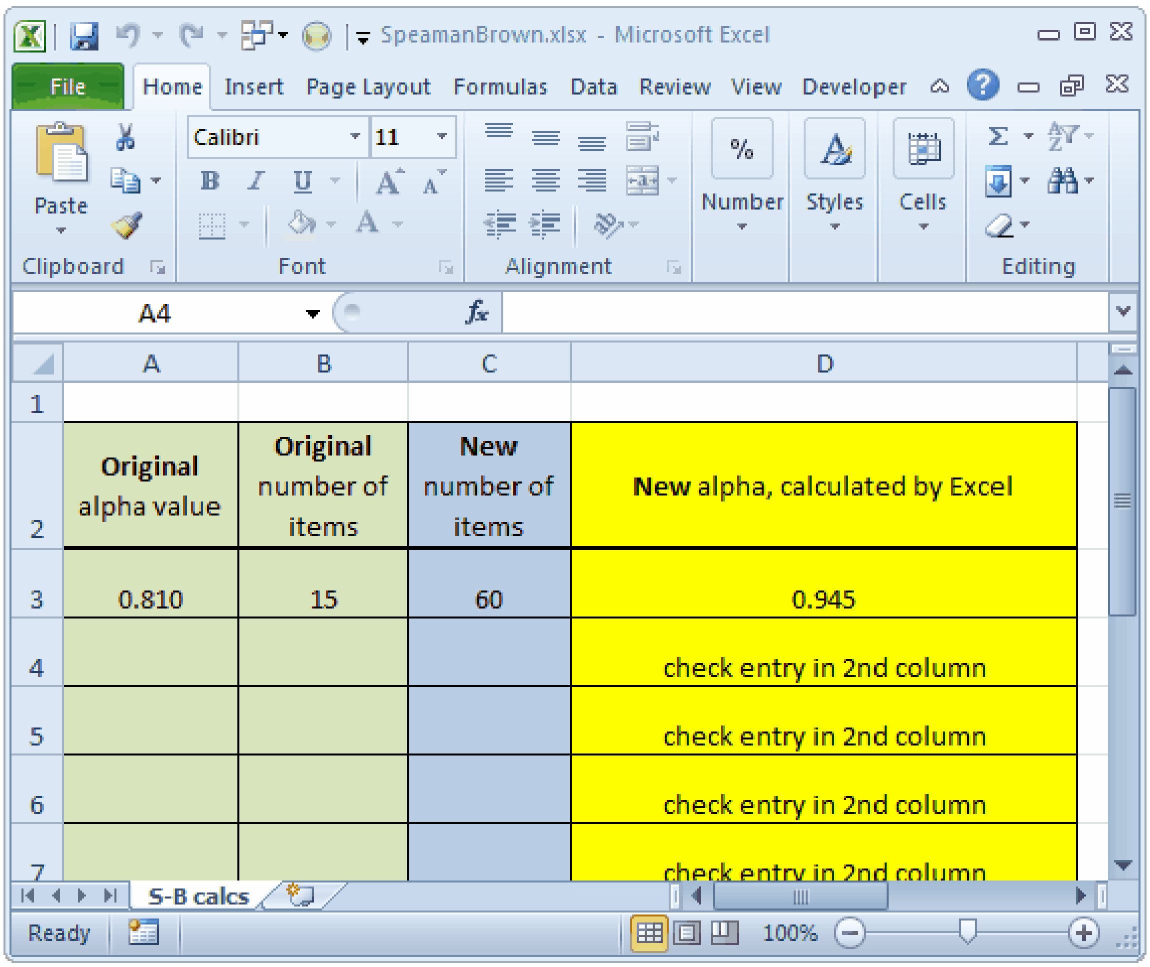
Our world renowned Lertap 5 help website, "**Lelp**", has all you need to know about how to get, install, and run Lertap 5. It also gets into such matters as upgrades, updates, and known problems.

Simply click here: to pay a visit to Lelp.

1.2 Spearman-Brown

The Spearman-Brown "prophecy formula" is applied in some of this site's samples. It's used to estimate how the reliability of a test would change if more items were added to the test. The "test" can be either a cognitive or an affective instrument.

An Excel workbook with this formula built in may be downloaded from here. (Note: in some browsers, you should right-click on this link, and then select the "Save as" option. In other browsers, an ordinary, plain old left-click will start the download. The workbook has been saved with an xlsx extension, ready to work with Excel 2007 or Excel 2010.)



To use this workbook, enter a reliability figure in the first column, such as the alpha coefficient found in Lertap's Stats1f report. This value should be greater than 0.00, and less than 1.00. In the example shown above, we've entered 0.810.

Enter the number of items in the present test (or subtest) in the next column. In the example shown here, the "present" test has 15 items; this number goes into the second (2nd) column, headed "original number of items".

The projected number of items in the "new" test is to be 60 in this example. In other words, what we'd like to know is what the reliability of our test would be were we to add 45 items similar in quality to the original 15, giving us a new, 60-item test.

The workbook will say "check entry in 2nd column" until suitable values are placed in the first and second columns. Then, after an entry is made in the third (3rd) column, the new reliability figure will appear in the yellow column, providing the new number of items is greater than the original number of items. In our example, the new test's reliability would be expected to approximate 0.945.

Search the internet for more about the Spearman-Brown formula (it is widely used). Wikipedia had a workable definition as of May, 2011.

2 Samples

We have a few samples for you to admire and play with.

All of these samples have links leading to downloadable workbooks ready to run with Lertap. One sample, the "Cooks Tour", has test and survey results which are not only Lertap-ready, but do not require a download of any sort.

A note for item and test analysis classes and instructors: two of the samples below might be particularly useful as they include actual test/survey items as well as Excel workbooks containing authentic data. They are No. 4 and No. 17. The other samples by and large have data only, an exception being "FIMS", No. 15, an example with items and data but involving considerable complexity.

In the table below, "Complexity" mostly refers to the set up of the workbook's **CCs** worksheet, where lines of Lertap's control language are housed.

No.	Name	Complexity	Subtests	Description
(1)	Chem Quiz	Simple	1 cognitive	A 10-item **quiz** given by a chemistry teacher.
(2)	CEQ	Simple	1 affective	A **questionnaire** with 10 Likert questions; some require reverse-scoring.
(3)	Maths Quiz	Simple	1 cognitive	A 15-item maths quiz for high school students. Has a mis-keyed item, but otherwise exhibits good item performance,

No.	Name	Complexity	Subtests	Description
				with very acceptable reliability, and interesting **quintile plots**.
(4)	Test13	Simple	1 cognitive	A set of 13 multiple choice maths items used to test elementary skills back before the advent of calculators. Includes actual test items, and allows for comparing results by four geographic zones. A recommended dataset for beginning Lertap users, and also for students new to item and test analysis topics.
(5)	M.Nursing	Simple	1 cognitive	Nursing exam with 60 items, scored on a **mastery** basis with a score of 70% required in order to pass. Has a faulty student record which should be culled from the Data worksheet. Looks at **classification consistency**. Includes a **breakout of results** from different countries.
(6)	Negocios	Simple	1 cognitive	A 60-item test from a business school, with some students caught **cheating**. (**RSA** analysis used to detect who they were. Note that Samples 12 and 13 below also apply RSA.)
(7)	LenguaBIg	Medium	3 cognitive	50-item English-language test for junior high school students in S.E. Asia country. Of the 50 items, 10 were **trial** (or "**pretest**") **items**.
(8)	LaFlorida	Simple	2 cognitive	Difficult university **screening test;** 25 mathematics and 25 physics questions, with cut-score at 50%. Poor reliability, but adequate to good **classification consistency**.
(9)	HalfTime	Medium	5 cognitive	A 100-item cognitive test, used to demonstrate how to set up Lertap so that it will calculate **split-half reliability** figures.

No.	Name	Complexity	Subtests	Description
(10)	Mente	Simple	2 affective	Trial items developed by a doctoral student to measure a person's perceived ability to change, and to succeed in spite of adversity. Looks at **scale reliability**; alludes to the relationship between **principal components** and **coefficient alpha**.
(11)	StuIQ	Complex	6 cognitive	Involves two forms of a 70-item aptitude test for senior high school students. Each form had a mixture of multiple-choice and **constructed-response** items. Issues investigated: **parallel-forms reliability**, correlations among M-C and C-R subtests, and **practice effects**.
(12)	Zmed	Simple	1 cognitive	Results from a medical school admissions test with 100 items, many of them using **multiple responses**.
(13)	Uni Class A	Simple	1 cognitive	Test results from a typical university instructor-made test. Has 30 items and 127 student responses. Used to show how Lertap's cheat checker, "**RSA**" is used, with results compared to those from another "similarity detector".
(14)	Uni Class B	Simple	1 cognitive	Very similar to the Uni Class A dataset; an additional example of the application of Lertap "**RSA**", response similarity analysis. Involves 34 items and 132 students.
(15)	FIMS	Complex	1 cognitive	Results from the First International Mathematics Study, "FIMS", with data from Japanese and Australian students -- has a mixture of **supply** (constructed response) and **multiple-choice** items, and also has **two nominal variables** for use in group comparisons, **gender** and **country**.

No.	Name	Complexity	Subtests	Description
(16)	BFI-25	Simple	5 affective	Has data for five five-item **personality scales** based on work from Berkeley University's Personality Lab.
(17)	BDI-21	Simple	1 affective	Sample data from an application of the **Beck Depression Inventory**.
(18)	Geology49	Simple	1 cognitive	Compare the performance of high school girls and boys on a geology test, with an eye for possible "**DIF**", differential item functioning.
(99)	Cook's Tour	Medium	1 cognitive 2 affective	An instrument given at the end of a **workshop**, with both cognitive and affective items. The cognitive questions covered the workshop's content. The affective questions collected feedback on how participants felt about the workshop.

Tidbit:

Lertap is controlled by the lines in the **CCs** worksheet. As you page ahead, you'll see a variety of CCs worksheets, with their corresponding lines of controls for Lertap. The best reference for these controls is the on-line help system. Another reference is the manual; although it is more dated than on-line help, the manual is still a good reference to turn to when you prefer to sit back in your armchair with a cuppa and a good, relaxing, off-line, book to read.

2.1 BDI-21

"**BDI**" = Beck's Depression Inventory

This well-regarded inventory was the focus of a lecture on instrument reliability and validity presented to Thai nursing students in 2018.

The inventory itself may be seen here.

A recommended read is this Wikipedia web page, active as of 20 September 2018.

A Lertap-ready workbook with sample data is available on the next page.

2.1.1 Downloads

The dataset, formatted as a Lertap 5 Excel workbook, may be downloaded from this link.

A "csv" file with item scores is available for downloading here.

The data are responses from 242 Australian university students as copied from the freely-available "**KernSmoothIRT**" package; refer to this reference.

2.2 BFI-25

This dataset is based on work from the University of Berkeley Personality Lab, under the direction of Oliver P. John; a link active in mid-July 2018 is here.

A Lertap-ready workbook for Excel, found under "Downloads", has been created by copying data from the BFI data file that accompanies the psych package developed by William Revelle at Northwestern University.

"BFI" means "Big Five Inventory". "BFI-25", as used here, refers to a 25-item subset of selected BFI items, having five items for each of five BFI scales: "Agreeableness", "Conscientiousness", "Extroversion", "Neuroticism", and "Openness".

A codebook for the data is available here.

Comments on scoring the items, as originally seen in documentation for the psych package, are here. Note that the Lertap-ready dataset, available via the next page, incorporates the scoring codes so that they're ready for use by Excel.

The manual for the psych package may be obtained from this link. An excellent introduction to the use of the psych package, *highly recommended*, is available via this link.

This document has sample output created by running one of the five-item scales, "Neuroticism", through Lertap, Xcalibre, and two R packages: psych, and ltm.

2.2.1 Downloads

The dataset, formatted as a Lertap 5 Excel workbook, may be downloaded from this link.

The dataset has 25 items and three demographic variables: gender, age in years, and education level. Data for 2,800 people are included.

If the demographic variables will be used with Lertap, for example, to get group- or item-breakouts, then the demographic data should be recoded in order to avoid a problem Lertap has with numeric variables when breakouts are requested. Lertap has a recoder which may be used for this very purpose. If you use it, you might do well to read the caveat on this page first.

Note that the data have been imported from the R psych package. If you are using R, and have installed the psych package, there wouldn't be any need to download the dataset.

2.3 ChemQuiz

A high school chemistry teacher developed a short quiz based the periodic table.

The quiz consisted of 10 multiple-choice questions. Each question had four options: A, B, C, and D.

The teacher gave the quiz to his class, and recorded the responses of each student in an Excel worksheet called "Data". These responses are captured here:

Data from the ChemQuiz sample.

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Anderson	D	B	B	C	D	D	A	C	A	B
Baker	B	B	B	B	D	B	C	C	B	B
Camberwell	B	B	A	B	D	A	B	C	B	D
Donaldson	B	B	C	B	D	C	B	C	C	B
Eggmont	B	B	A	B	B	C	C	C	B	B
Fredricksson	B	A	B	B	D	C	A	C	B	A
Graphner	D	D	A	B	D	C	A	C	B	A
Humphrey	B	B	B	D	D	C	C	D	B	B
Invererity	B	B	A	D	B	C	B	C	A	A
Johnson	B	D	B	C	A	B	C	C	B	B
Klein	D		B	A	D	C		C	A	A
Lampton	B		A	A		C		A	B	A
Mecurio	B		B	D	A	C	C	A	B	B
Nesbit	A	C	A	D	B	C		A	D	A
Oldfelt	A	D	A	A	A	C	A	C	B	A

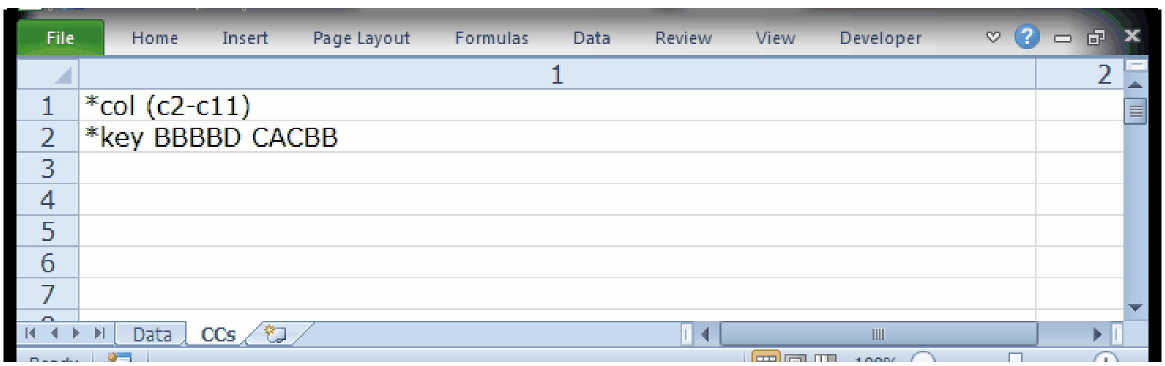
Four students did not answer some of the questions. For example, for some reason Klein did not answer two items, Q2 and Q7.

There is much information which the teacher could glean from the Data worksheet itself. There were so few students; he could, without much trouble, rather quickly summarise how many students got each item correct, and which students selected one of the "distractors" (wrong answers).

However, we want to show you what Lertap does, so let's move on, imagining the day when there will be many students.

Lertap requires two Excel worksheets before it'll do anything. Item responses have to be in a Data worksheet, with the first row used for a concise summary of the information in the worksheet, the second row for column headers, and then the actual student responses, starting in row 3.

We also have to have a "**CCs**" worksheet with instructions for Excel. The teacher knew this, and he entered these two lines of information in the syntax common to Lertap:



The purpose of the first CCs line, ***col**, is to tell Excel that item responses are found in columns 2 through 11 of the Data worksheet. The second CCs line, ***key**, gives Excel the correct answer to each question; these are referred to as "keys". There are 10 keys in the line, one for each of the 10 questions. The correct answer to the first item, Q1, is B. The correct answer, or the "key", for the 7th item is A.

(Note: the space between the 5th and the 6th key is <u>not</u> required. Grouping keys in sets of 5 makes it easier to quickly identify, for example, the keyed-correct answer to the seventh item, or the ninth, and so on. There's also a space before the first key, as there is before the opening parenthesis on the *col line; these spaces <u>are</u> required.)

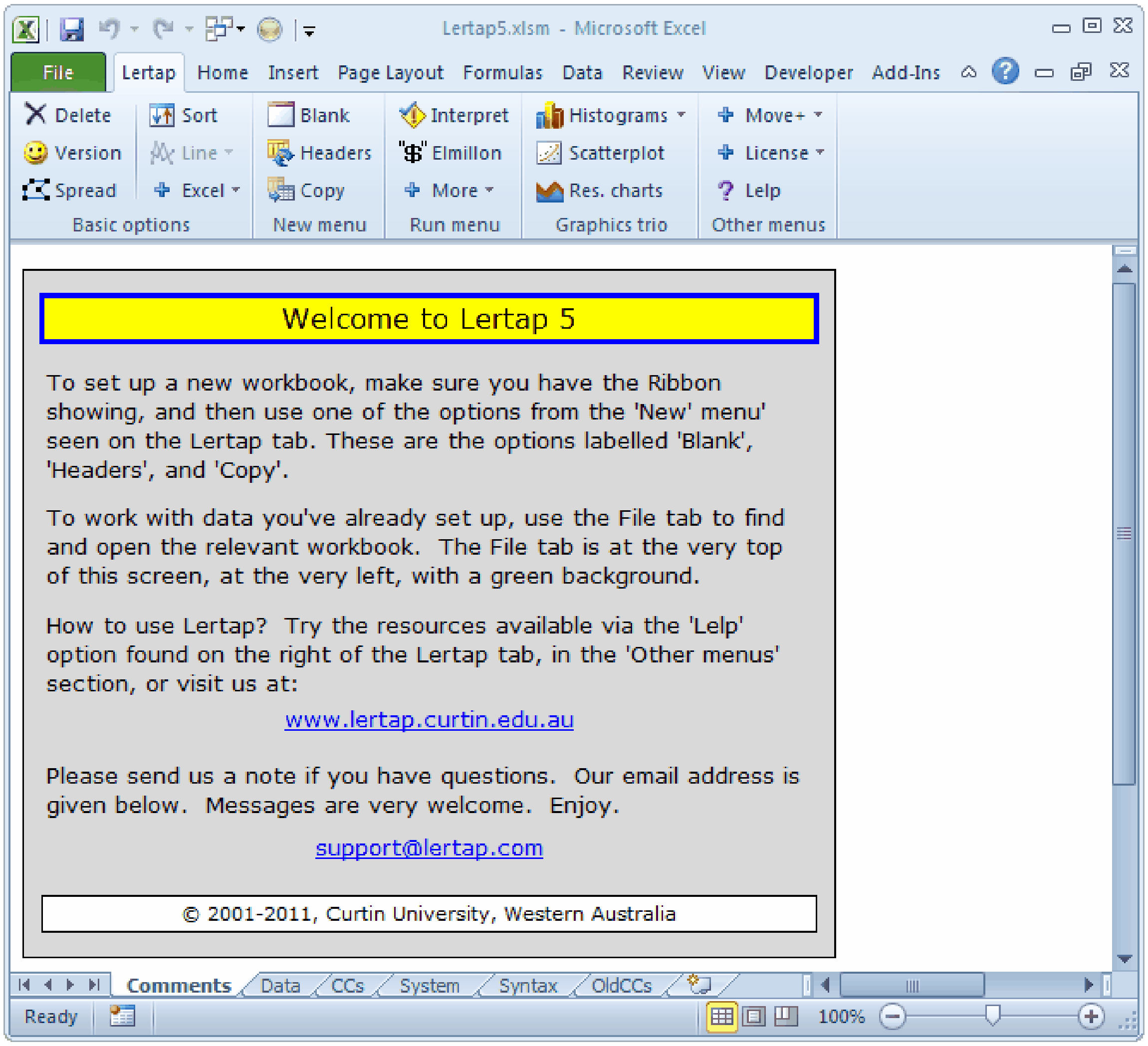
People sometimes ask what "**CCs**" stands for. The answer is: "Control Cards". The term dates back to the mainframe version of Lertap, to a time when "punch cards" were used to enter data, and to specify how a batch of punched cards was to be processed by the computer. (For more information about punch cards, search the internet.)

The next topic shows how Lertap can be used to process the students' responses.

2.3.1 Results

The teacher used the Excel 2010 version of Lertap 5 to find out how many students answered each question, or item, correctly.

This is how Lertap 5 looked when he processed his data early in 2011:

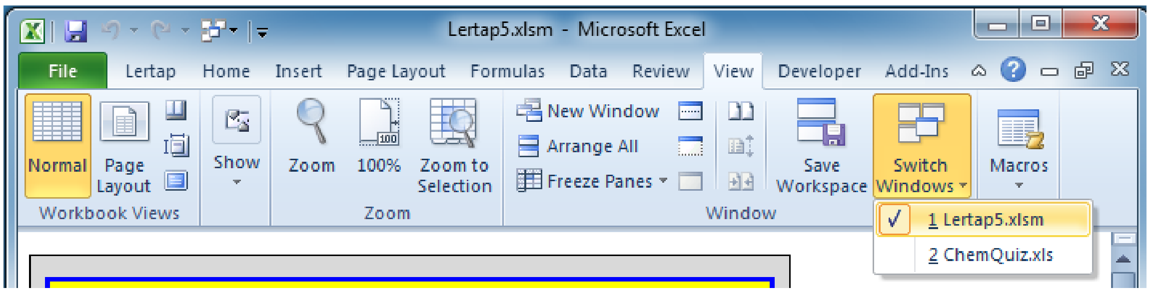


In order to get Lertap to produce results for his quiz, the teacher has to follow three steps.

Step 1: switch the view to his workbook.

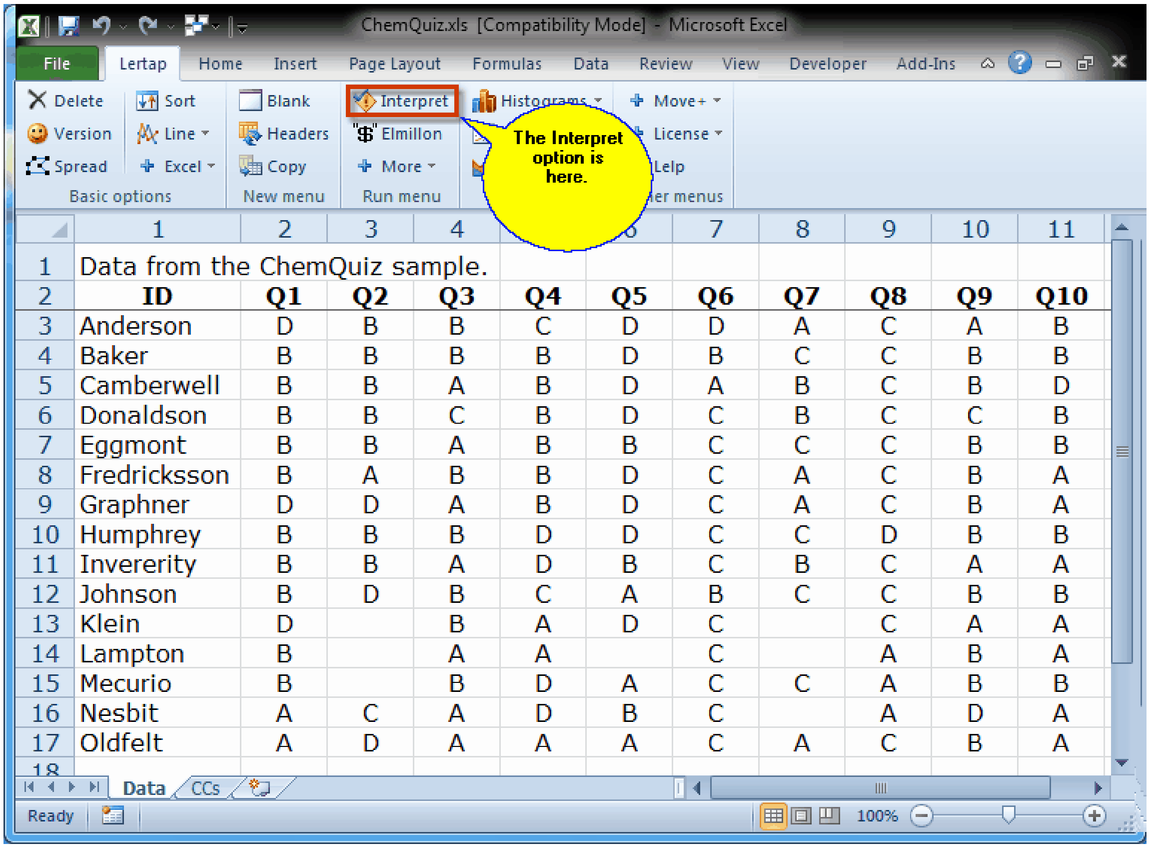
The name of the teacher's Excel workbook was "ChemQuiz.xls". The workbook in focus above is "Lertap5.xlsm"; it's the "active workbook". Before we can work with ChemQuiz.xls, we have to bring it to the fore, making it the active one. There are several ways to do this -- holding the Alt key and then pressing the Tab key is one way, and easy. In Windows Vista, each open document has a "tab" on the task bar at the bottom of the screen, and switching from one workbook to the other is done by clicking on the tabs. In Windows 7, the task bar at the bottom of the screen also has "tabs" for each running application, but the links to the documents in use by the applications are all collapsed; they reveal themselves when you hover above the tab.

Another way to switch from Lertap5.xlsm to ChemQuiz.xls is to go to the View tab, select Switch Windows, and then click on ChemQuiz1.xls.



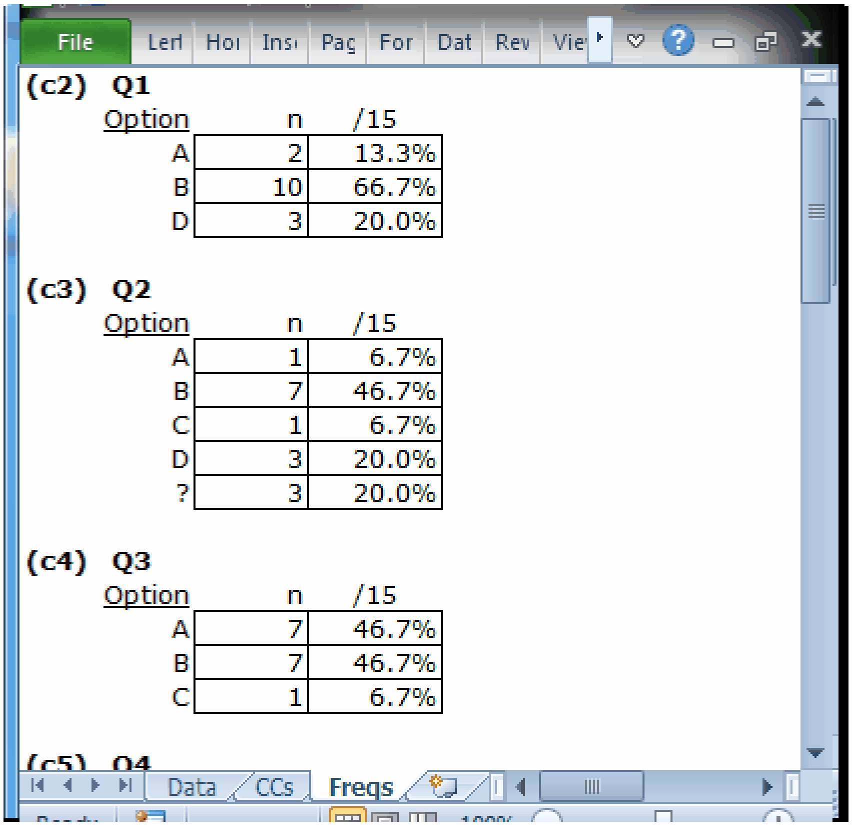
Step 2: from the Lertap tab, select Interpret in the Run menu section.

Lertap action begins with a click on the **Interpret** option:



	1	2	3	4	5	6	7	8	9	10	11
1	Data from the ChemQuiz sample.										
2	**ID**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
3	Anderson	D	B	B	C	D	D	A	C	A	B
4	Baker	B	B	B	B	D	B	C	C	B	B
5	Camberwell	B	B	A	B	D	A	B	C	B	D
6	Donaldson	B	B	C	B	D	C	B	C	C	B
7	Eggmont	B	B	A	B	B	C	C	C	B	B
8	Fredricksson	B	A	B	B	D	C	A	C	B	A
9	Graphner	D	D	A	B	D	C	A	C	B	A
10	Humphrey	B	B	B	D	D	C	C	D	B	B
11	Invererity	B	B	A	D	B	C	B	C	A	A
12	Johnson	B	D	B	C	A	B	C	C	B	B
13	Klein	D		B	A	D	C		C	A	A
14	Lampton	B		A	A		C		A	B	A
15	Mecurio	B		B	D	A	C	C	A	B	B
16	Nesbit	A	C	A	D	B	C		A	D	A
17	Oldfelt	A	D	A	A	A	C	A	C	B	A

The Interpret option actually does a considerable number of things. It checks to see that it can understand, or "interpret", the lines in the CCs worksheet. It sets up the crucial "**Sub**" worksheet for Lertap's internal usage, and then, after hiding the Sub worksheet, goes on to make a new worksheet, "**Freqs**", which looks like this:



(c2) Q1

Option	n	/15
A	2	13.3%
B	10	66.7%
D	3	20.0%

(c3) Q2

Option	n	/15
A	1	6.7%
B	7	46.7%
C	1	6.7%
D	3	20.0%
?	3	20.0%

(c4) Q3

Option	n	/15
A	7	46.7%
B	7	46.7%
C	1	6.7%

(c5) Q4

The Freqs, or ***frequencies***, report is truly a simple one. One of its primary purposes is to provide a check on data processing. In this case, all of the items, or questions, used A, B, C, and D as possible responses. So, were we to see an E or an F in the Freqs report, we might be alarmed (more likely, perhaps, would be a lower-case a, b, c, or d, which might also alarm us).

If all items used A, B, C, and D, why is there no C for the little Q1 table? Because no student selected option C on this item -- if you scroll up to the view of the Data sheet above, you can see there's no C in c2, the column used for Q1. (We might say that this option was a no-C, eh?)

What's that little **?** mark doing at the bottom of Q2's table? It's pointing out that there were three responses in c3, column 3, which were not A, B, C, or D.

Another glance back at the Data worksheet shows that c3 has three blanks, three places where students did not answer Q2. These were the students named Klien, Lampton, and Mercurio.

Having said these things, we should definitely point out that Freqs has some useful content, a message or two for us to take in. For example, Q1, whose responses are located in c2, the second column, had an option, C, which no-one chose. Later on we'll have the chance to say that, in some cases, having an option which nobody selects can be an unwanted outcome at times.

Freqs also indicates that students were "all over the place" on Q2; their responses were spread out over all of the item's four options. This may have been an item which confused the students, causing three of them to refrain from offering any answer at all.

What if Freqs did indicate an error? What if, for example, it showed this:

(c2) Q1

Option	n	/15
A	2	13.3%
B	10	66.7%
D	2	13.3%
x	1	6.7%

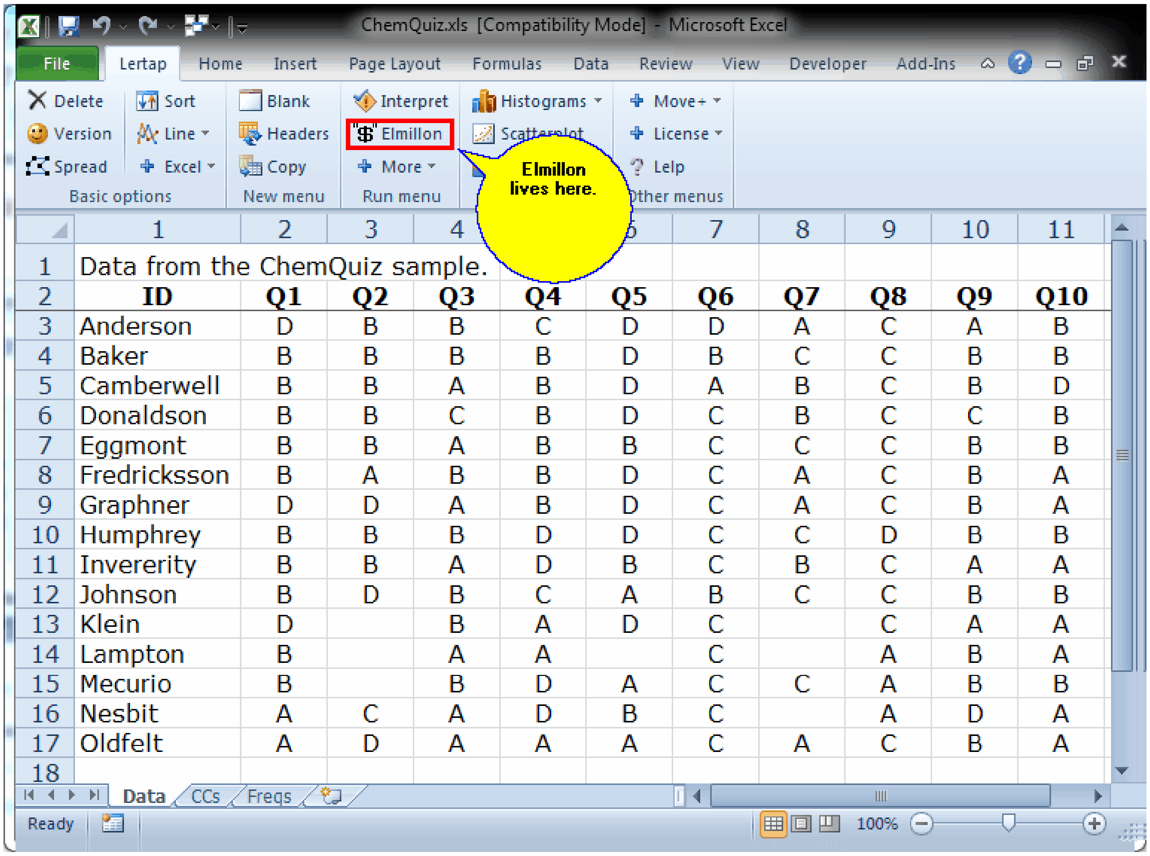
Of the 15 records of student Q1 responses in the Data worksheet, one has an "x" in the second column, c2. Highly unexpected, and unwanted. Before asking Lertap to make any more reports of results, we would want to go back, find the error, and correct it. In this case that would be easy to do as there are only 15 students. In a larger data set, with thousands of students, finding the x is something we'd ask Excel to do by, perhaps, simply using a filter on c2, or, perhaps, by searching c2 for the row which has that x. Excel has heaps of power and options, and its help system isn't too bad either.

What does our chemistry reacher know after using the Interpret option? How the students responded to each question. This may be useful information on its own, but most times it's likely that the teacher would go on to Step 3.

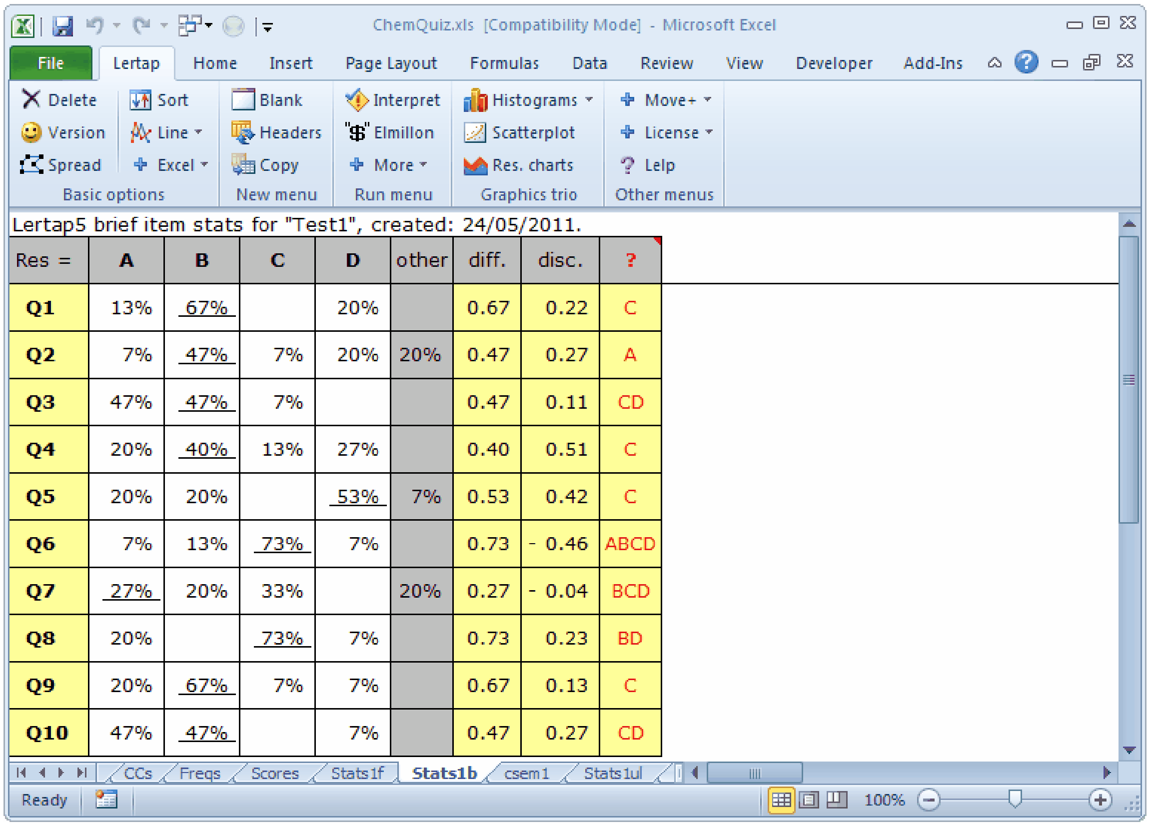
Step 3: from the Lertap tab, select Elmillon in the Run menu section.

"**Elmillon**" was the name of the first version of Lertap, developed for the Venezuelan Ministry of Education in 1971. When it was finished, and installed in the Ministry's research division, the head of the division said it was worth a million Bolivares to the division, "*un millon de Bolivares*", just as Americans might say "*Hey!, it's worth a million bucks to us!*". In English, Elmillon might be pronounced as "el me own".

Elmillon is the real engine room of Lertap.



After clicking on the Elmillon option, the teacher's screen looked something like this:



Lertap5 brief item stats for "Test1", created: 24/05/2011.

Res =	A	B	C	D	other	diff.	disc.	?
Q1	13%	67%		20%		0.67	0.22	C
Q2	7%	47%	7%	20%	20%	0.47	0.27	A
Q3	47%	47%	7%			0.47	0.11	CD
Q4	20%	40%	13%	27%		0.40	0.51	C
Q5	20%	20%		53%	7%	0.53	0.42	C
Q6	7%	13%	73%	7%		0.73	- 0.46	ABCD
Q7	27%	20%	33%		20%	0.27	- 0.04	BCD
Q8	20%		73%	7%		0.73	0.23	BD
Q9	20%	67%	7%	7%		0.67	0.13	C
Q10	47%	47%		7%		0.47	0.27	CD

Elmillon has added five (5) new worksheets to the workbook: Scores, Stats1f, Stats1b, csem1, and Stats1ul, and has placed the focus on Stats1b.

Our chemistry teacher was not an expert on the theory of tests and measurement. He simply wanted to know how many students were able to answer each question correctly. The **Stats1b** report suits his needs: the columns labeled **A B C D** display the percentage of students who selected each option, with the correct answer, the item key, underlined. (The "other" column has an entry when some students do not answer an item.)

On several questions, more than two thirds of the class knew the right answer. These were Q1, Q6, Q8, and Q9. While this was pleasing to the teacher, he was concerned with some of the other questions. For example, he noted that only 27% got Q7 correct, an item which had to do with the rare earth elements. He made a note to review this topic with the students.

It bears repeating that the teacher in this example had a simple, practical need, one which related to feedback on his teaching, and the possible need to review topics with his class. Of Lertap's numerous reports, he used only Stats1b, and, even then, only part of it. The **diff.**, **disc.**, and **?** columns in Stats1b went unused in this case. They refer to item difficulty, item discrimination, and to item options which may not have performed in a statistically-optimal manner; these indicators were of no interest to our teacher.

It's also for this reason that the teacher did not care to look at the other reports made by Lertap, such as Scores, Stats1f, csem1, and Stats1ul. Stats1b had all the information he needed in this case. The other reports come to the fore when test scores are of interest, a matter which leads directly into a determination of test quality. The MathsQuiz sample is our first example of how the other reports are used to look at into such issues; the M.Nursing sample gets into the case of mastery testing, and also shows how Lertap and Excel can team up to look at how item responses, and test scores, may differ among groups of test takers (such as males and females, or various geographic regions). The Negocios sample touches on how Lertap may be used to see if students may have colluded when answering test items (that is to say: cheated).

Tidbits:

Users can turn off the creation of some of Lertap's reports. The production of the Stats1ul sheet, in particular, is quite an intensive activity for Excel to engage in, taking time, memory, and temporary disc space for scratch work. If the teacher had no interest in Stats1ul, he could use the settings in Lertap's **System** worksheet to turn off its production.

(The Stats1ul report is extensively covered in the printed manual. An electronic version of the manual, a file called "Lertap5UserManual.pdf", may be downloaded from here.)

2.3.2 Download

The ChemQuiz.xls Excel workbook may be downloaded here.

A data processing / Lertap exercise based on ChemQuiz may be downloaded here. This exercise is a doc file which contains student answer sheets, one per student, and some (optional) instructions for processing the responses with Lertap. The purpose of the exercise is to provide a practical data-entry and data-analysis experience, using Excel and Lertap.

(If you do the exercise, here's a tip that will save you time when you have to enter the item responses: use The Spreader.)

From a measurement and psychometric point of view, this quiz had poor properties. Its reliability was low, and some of the items had unacceptably low discrimination. This was deemed not relevant for this introductory section of the Guide. (It would sort of ruin the story line, but this is not to say, not for a moment, that the example is not real. Using the Stats1b report in the manner exemplified here is entirely legitimate, of course. Users do not need to know about the intricacies of all of Lertap's output, which can indeed be excessive, even unnecessary, for numerous users, our teacher among them.)

2.4 CEQ survey

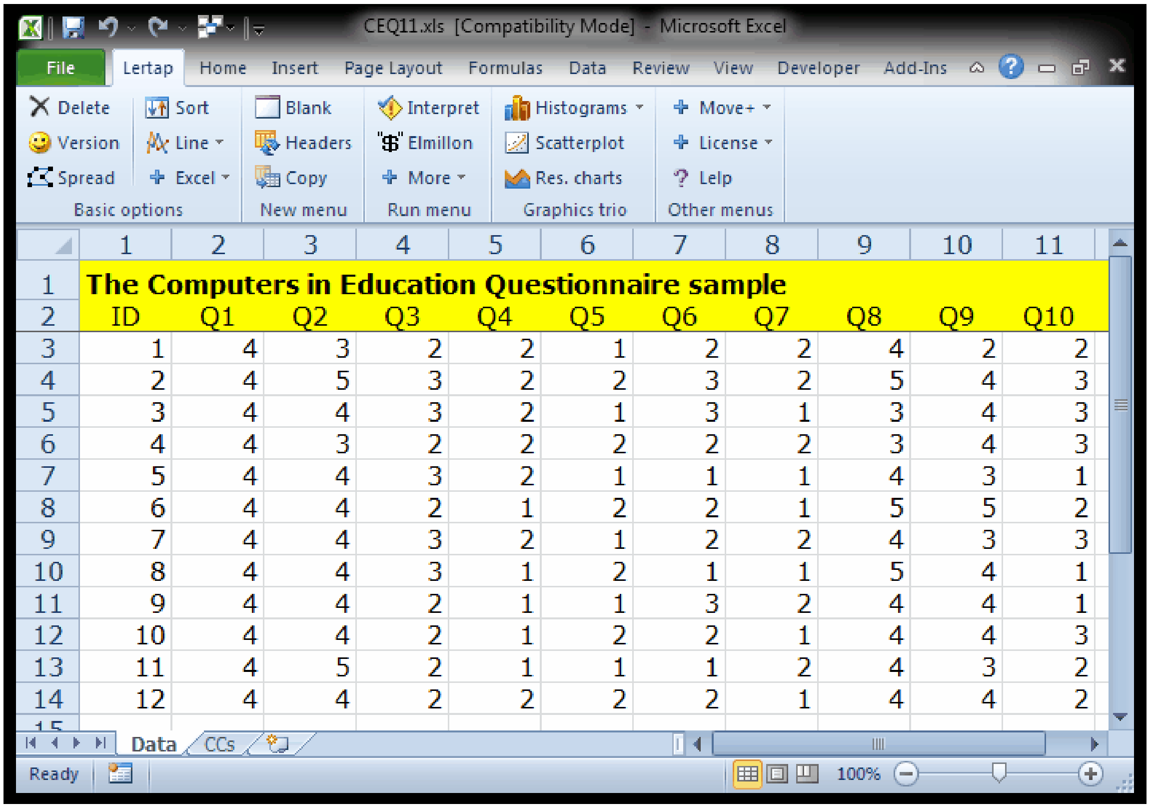
An Otago University honours student used a short survey to assess the attitudes of selected Dunedin high school principals regarding the incorporation of microcomputers in classrooms. This she did in 1985, when personal computers were starting to make their way into education. At that time some people said that all students should be "computer literate"; some educators maintained that computers could also be used as very effective learning and teaching aids. There was a cry to get them into classrooms.

The questionnaire's items are shown below. Principals responded on a five-point **Likert** scale, going from *Strongly Disagree* on the left, to *Disagree*, to *Undecided*, to *Agree*, to *Strongly Agree* on the right. When their responses were coded into Excel, digits were used, from 1 (Strongly Disagree) to 5 (Strongly Agree).

---

1. Computers are valuable tools which can be used to improve the quality of education.
2. Computers should be used by schools more than they are now.
3. A school system should buy all other educational materials before purchasing computers.
4. A computer is an unnecessary luxury in most schools.
5. Computers are of little value in education because they can be used to teach just one or two subjects.
6. Computers are a danger because they dehumanize teaching.
7. Computers are of little value in the classroom because they are too difficult to use.
8. Computers provide motivation for students to learn.
9. All students should learn how to use computers.
10. Computers in schools have an adverse effect on students.

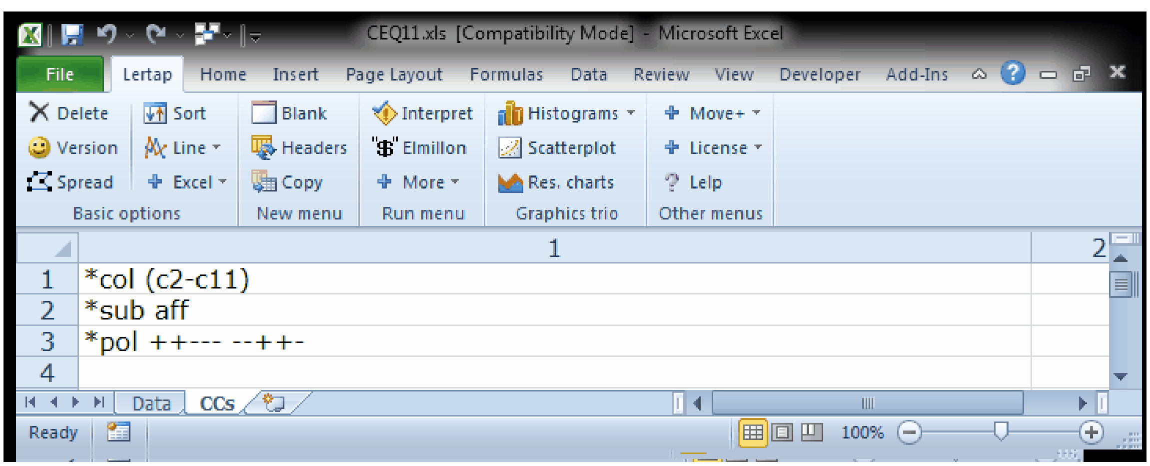
When the responses were entered into Excel, they looked like this:



	1	2	3	4	5	6	7	8	9	10	11
1	The Computers in Education Questionnaire sample										
2	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
3	1	4	3	2	2	1	2	2	4	2	2
4	2	4	5	3	2	2	3	2	5	4	3
5	3	4	4	3	2	1	3	1	3	4	3
6	4	4	3	2	2	2	2	2	3	4	3
7	5	4	4	3	2	1	1	1	4	3	1
8	6	4	4	2	1	2	2	1	5	5	2
9	7	4	4	3	2	1	2	2	4	3	3
10	8	4	4	3	1	2	1	1	5	4	1
11	9	4	4	2	1	1	3	2	4	4	1
12	10	4	4	2	1	2	2	1	4	4	3
13	11	4	5	2	1	1	1	2	4	3	2
14	12	4	4	2	2	2	2	1	4	4	2

The honours student was interested in the responses on a question-by-question basis. Importantly, she also wanted to add the item responses together so that she would have a scale score, which she proposed to refer to as the "CEQ scale".

If Lertap is to be used to process results, a CCs worksheet is required in order to tell Excel how to do the job. The following CCs sheet was created by the student.



	1
1	*col (c2-c11)
2	*sub aff
3	*pol ++--- --++-
4	

The first CCs line, ***col**, tells Excel that responses are to be found starting in column 2 (c2) of the Data worksheet, ending in column 11.

The second line, ***sub**, is required whenever the data are from a survey or questionnaire. "aff" is a Lertap control "word" which means "affective" -- surveys and questionnaires are said to be affective instruments, working in the "affective domain" (as opposed to the "cognitive domain"; search the internet for definitions of these terms).

The third line, ***pol**, has four + signs, and six - signs, one sign for each of the 10 items. These indicate the "polarity" of each item: if an item is a positive statement, such as that seen in the first question, its entry on the *pol line will be (naturally) a +. On the other hand, negative statements, such as that seen in the last question, have a - sign (you guessed it!). *pol lines are not always required; if all questions have the same "polarity", *pol is not needed. If there is no desire to add the item responses together to get a scale score, then *pol is not needed.

*pol lines effectively **reverse-score** negative questions. In this example, a response of *Strongly Disagree* to an item is coded as a 1 in the Data worksheet. If this item has a - entry in the *pol line, then the 1 will be converted to a 5 before the item response is added in as part of the scale score. If a principal strongly agrees with the first question, and strongly disagrees with the last, he, or she, will get 5 + 5 scoring points. Had the last item not been reversed, then the scoring points would be 5 + 1.

Scoring the item responses in this manner is meant to produce a CEQ scale score with high scores indicating a positive attitude towards the use of computers in schools. A principal who strongly agrees to the positive items, and strongly disagrees to the negative items, will end up with a score of 50. A principal who strongly disagrees to the positive items, and strongly agrees to the negative items, will end up with a score of 10.

The next topic gets into getting Lertap to produce some results.

2.4.1 Results

To get results, the student followed the three steps introduced in an earlier topic.

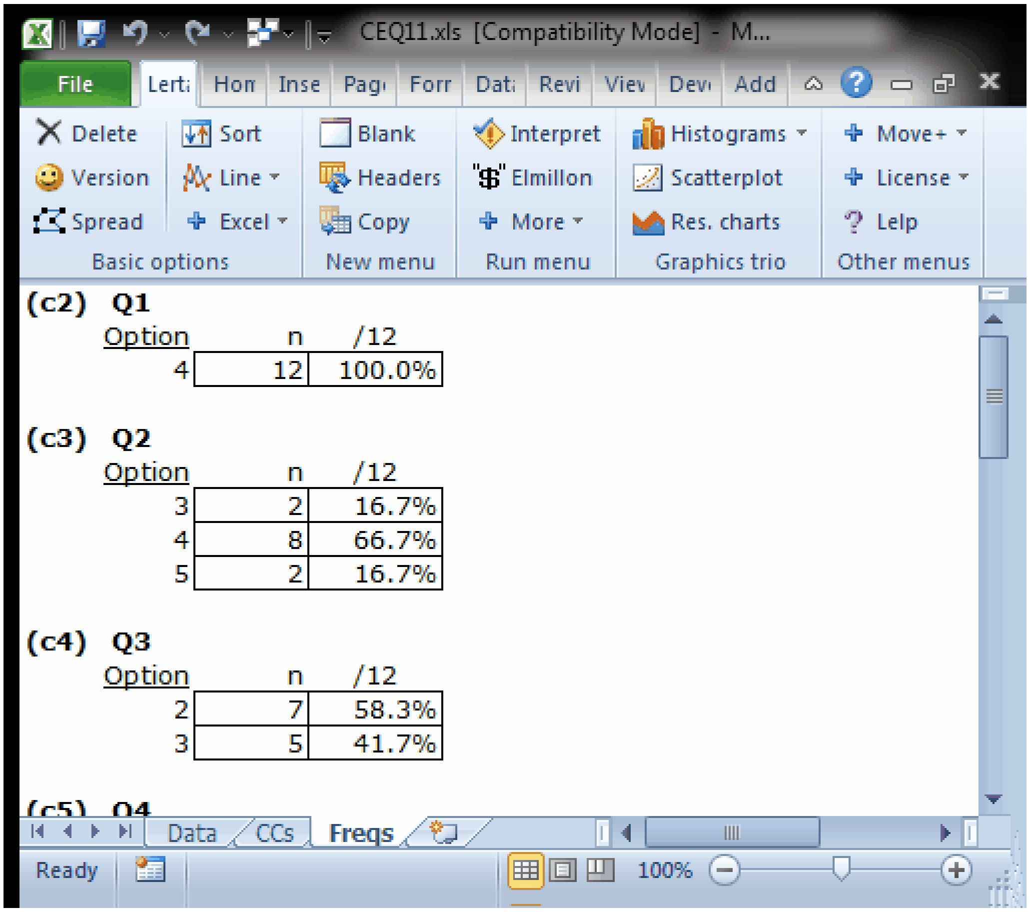
The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

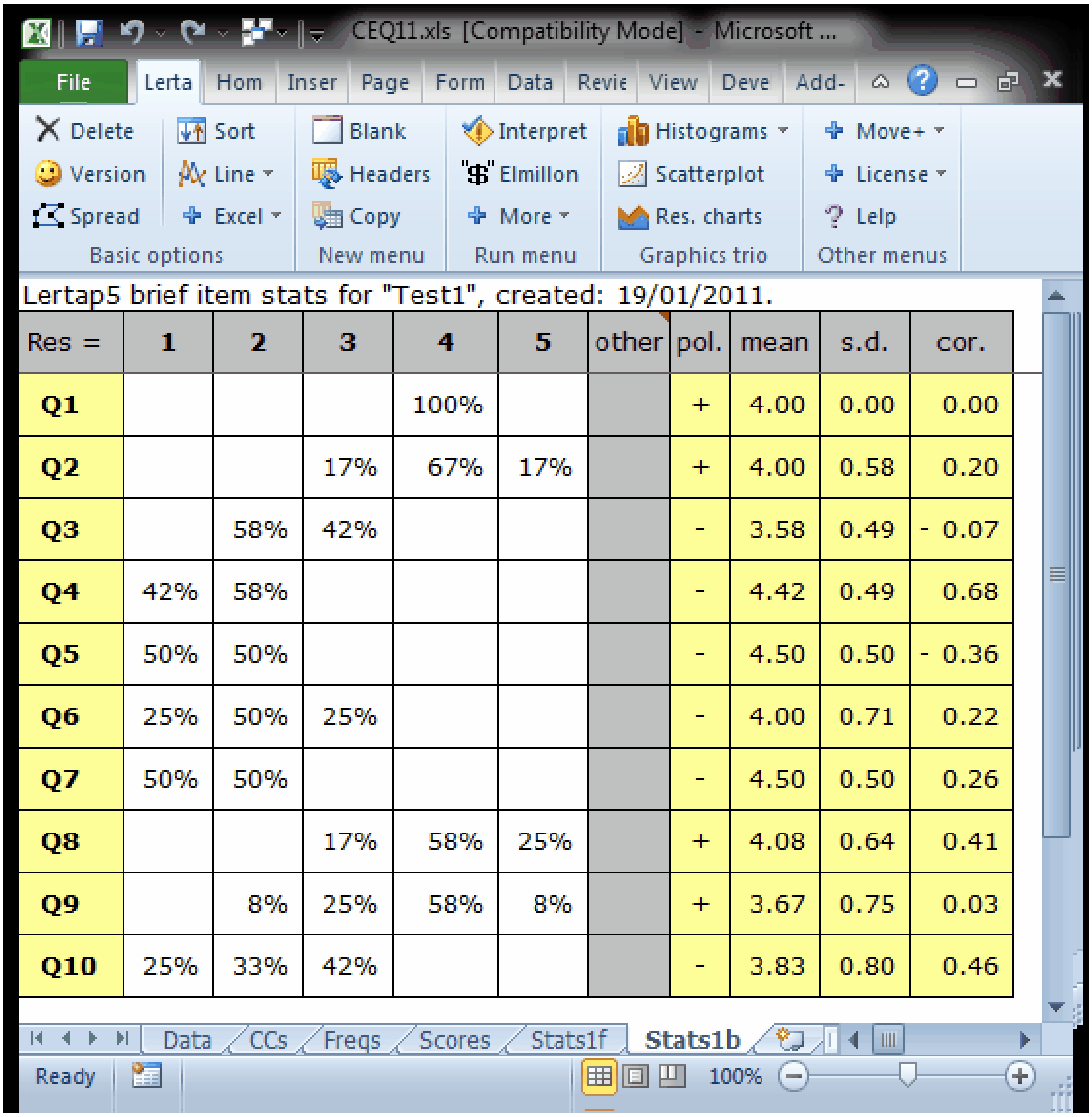
3) Use the **Elmillon** option on the Lertap tab.

The Interpret option produced the "Freqs" worksheet seen below:



The Freqs report was examined to see if there were any unexpected responses, such as, for example, a 7, which would represent a data processing error (responses were supposed to be coded as 1, 2, 3, 4, or 5). The student found no errors of this sort, so she went on to use Elmillon.

Elmillon always finishes by placing the focus on the Stats1b report.



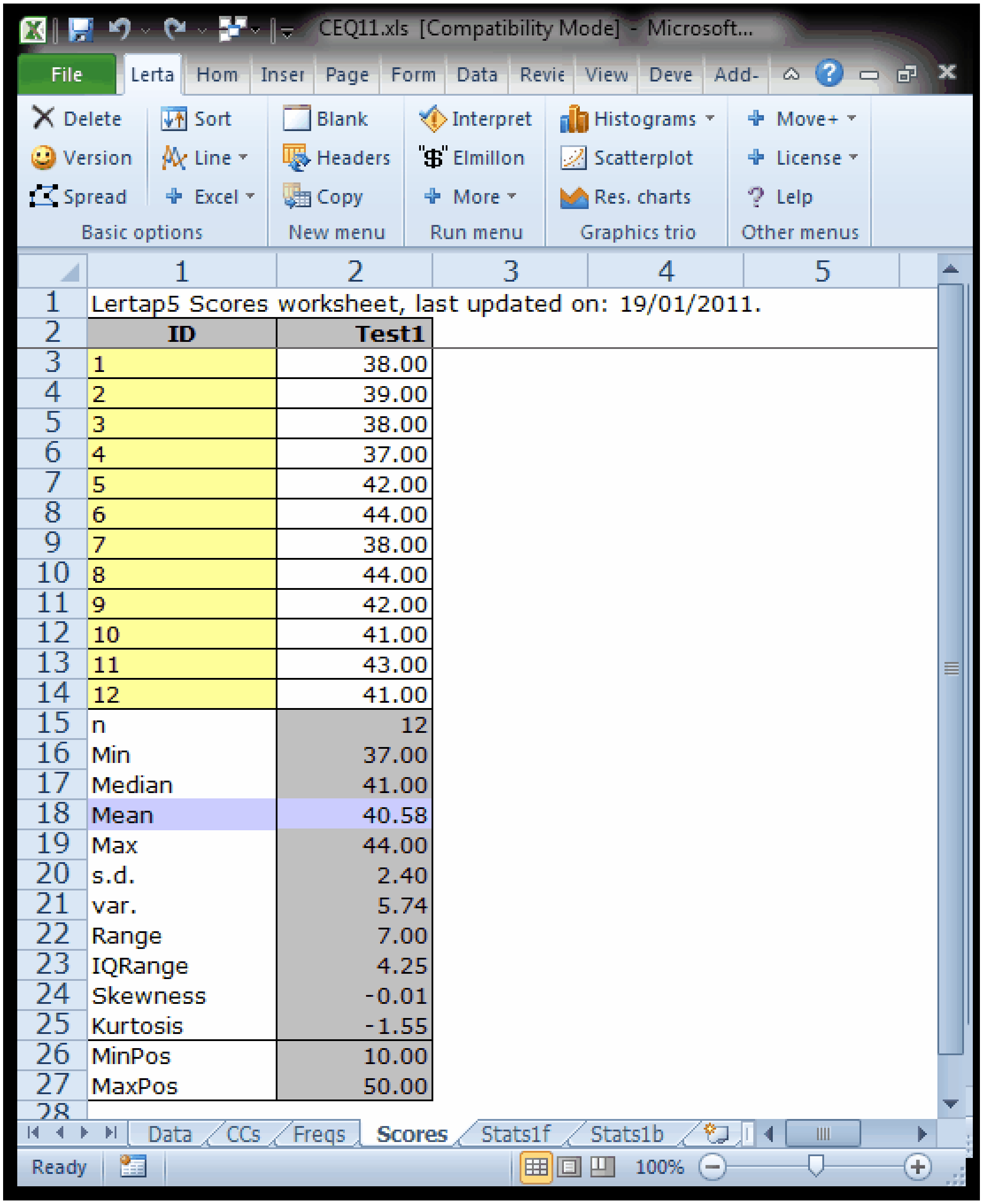
Lertap5 brief item stats for "Test1", created: 19/01/2011.

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q1				100%			+	4.00	0.00	0.00
Q2			17%	67%	17%		+	4.00	0.58	0.20
Q3		58%	42%				-	3.58	0.49	- 0.07
Q4	42%	58%					-	4.42	0.49	0.68
Q5	50%	50%					-	4.50	0.50	- 0.36
Q6	25%	50%	25%				-	4.00	0.71	0.22
Q7	50%	50%					-	4.50	0.50	0.26
Q8			17%	58%	25%		+	4.08	0.64	0.41
Q9		8%	25%	58%	8%		+	3.67	0.75	0.03
Q10	25%	33%	42%				-	3.83	0.80	0.46

The column headed "mean" gives the average response score for each item, with negative items reverse scored. The highest possible mean for any item is 5.00; an item with a mean of 3.00 would indicate that, on average, the principals were undecided about the item's statement. In this case, seven of the 10 items have means which are at or above the "agree" score of 4. (This is equivalent to "disagree" on the negative items.)

The Stats1b report gives the honours student the information needed to summarise the principals' responses on an item by item basis. They had the most positive response, on average, to items 5 and 7. Read up a bit more about Stas1b and its stats by a click.

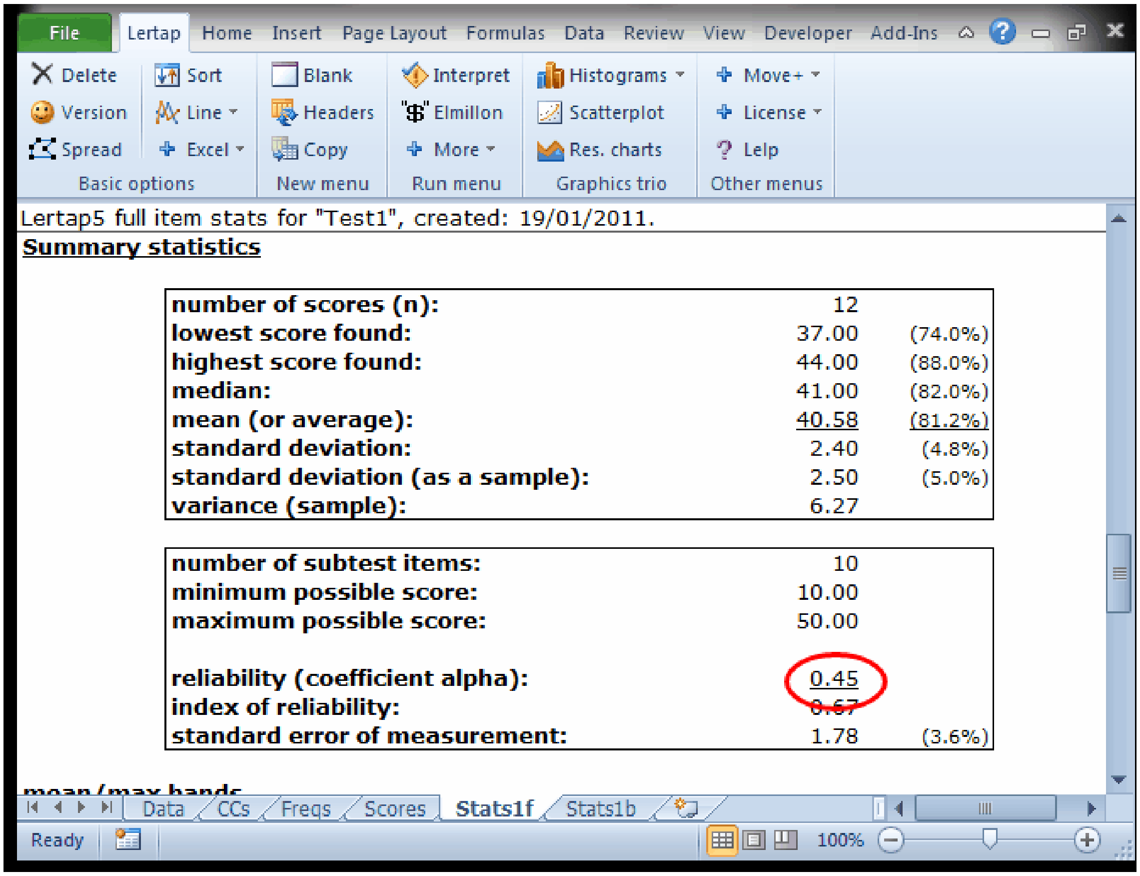
But the student had wanted to do more with the results. She had hoped to use the scale scores found in the Scores report for a portion of her dissertation:



Lertap5 Scores worksheet, last updated on: 19/01/2011.

ID	Test1
1	38.00
2	39.00
3	38.00
4	37.00
5	42.00
6	44.00
7	38.00
8	44.00
9	42.00
10	41.00
11	43.00
12	41.00
n	12
Min	37.00
Median	41.00
Mean	40.58
Max	44.00
s.d.	2.40
var.	5.74
Range	7.00
IQRange	4.25
Skewness	-0.01
Kurtosis	-1.55
MinPos	10.00
MaxPos	50.00

Scale reliability is an important issue when using measures like the CEQ. It is often said that an affective scale should have a reliability of at least 0.70. Did this one, when given to the 12 Dunedin principals? The answer is found in the Stats1f report:

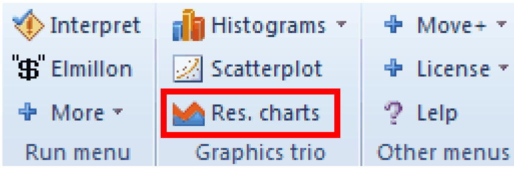


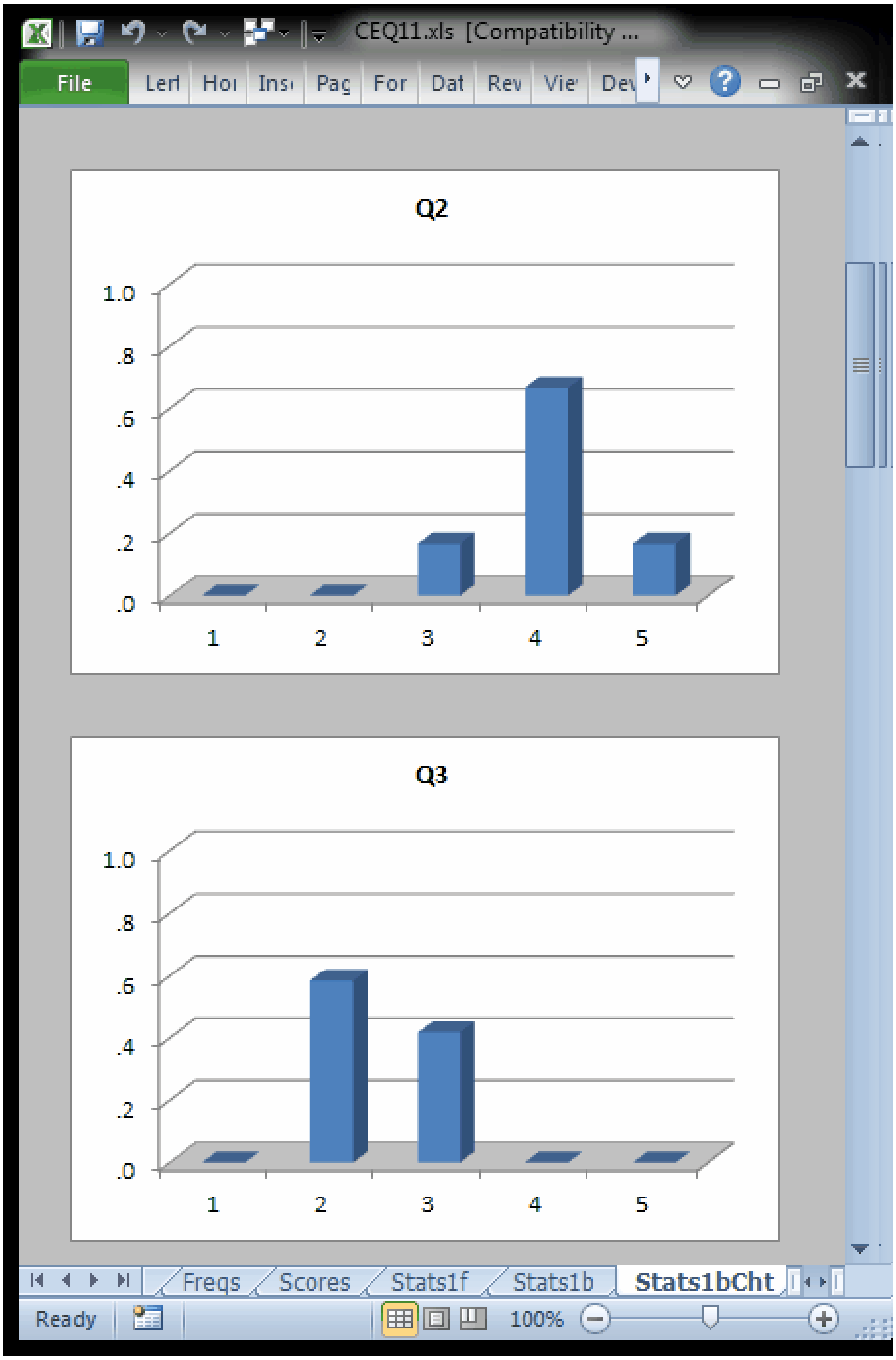
In this case the reliability was 0.45, quite less than the affective scale benchmark minimum of 0.70. This does not imply that the questionnaire was inadequate. What it means is that the scale scores cannot be used with any confidence: low reliability means high measurement error. The honours student had to get by with reporting her results using item data only, such as those given in the Stats1b report. This was by no means a bad outcome at all. The student wanted to know how principals felt about computers in schools, and their responses to each of the 10 items, combined with selected response charts (see the next topic), plus follow-up interviews with some principals, provided sufficient fuel for her honours dissertation.

(It was found that many of the principals had been under some pressure from parents who felt that their children might be left behind if not introduced to computers. Another great factor was financial: unless the school's budget could be increased, spending money on computers meant that other areas of the budget would have to be cut back. Some of the parent organisations started to organise cake drives to help raise funds for the schools.)

2.4.2 Charts

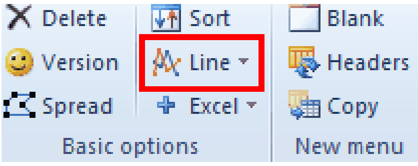
Excel has good capabilities when it comes to making charts. For example, when a Stats1b report has the focus (is the active worksheet), the Res. charts option will make a response bar chart for each item.



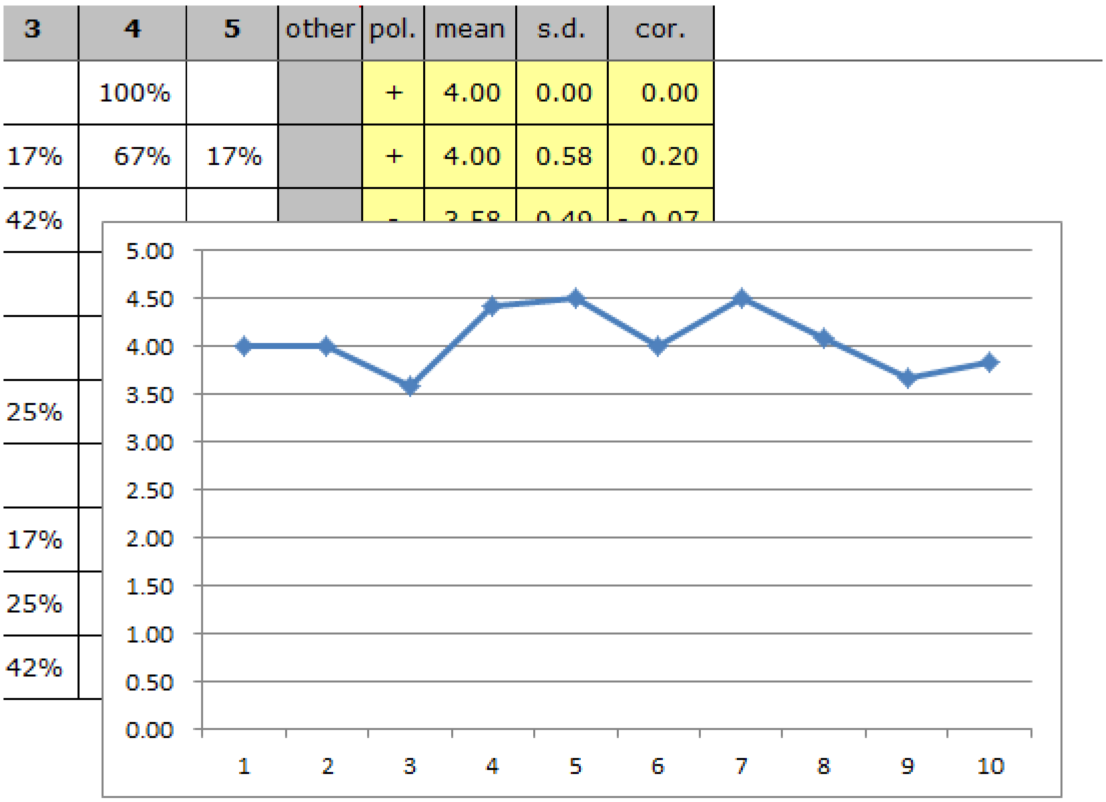


It is possible to quickly change these charts. A click in the chart area with the right mouse button will open up a menu of options. The "ChartChanger1" macro might also be used when it is desired to change all charts at once.

Another Lertap option provides a quick way to make an Excel line graph. For example, to get a graph of the item means, the student would select the values under the "mean" column of Stats1b, and then click here:



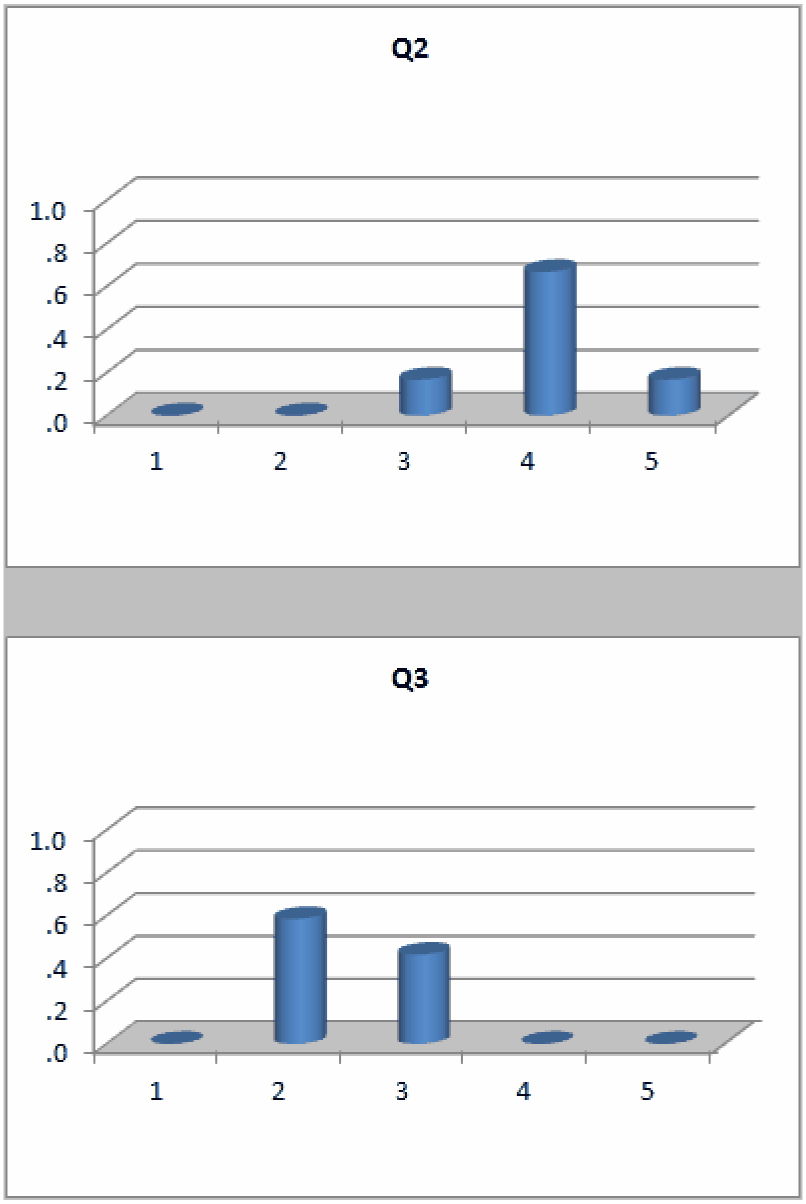
This rapidly results in a graph such as the following. Once again this graph, or chart, can be quickly changed by right-clicking on top of it, and then choosing from the options presented by Excel.



Tidbit:

The ChartChanger1 macro may be used to quickly change many of the charts seen on a Lertap report.

For example, it was used to change the response charts above from "3-D Clustered Column" to "3-D Cylinder", producing the examples seen below:



2.4.3 Download

The CEQ11.xls Excel workbook may be downloaded here.

A data processing / Lertap exercise based on the CEQ questionnaire may be downloaded here. This exercise is a doc file which contains questionnaire answer sheets, one per principal, and some (optional) instructions for processing the responses with Lertap. The purpose of the exercise is to provide a practical data entry and data analysis experience, using Excel.

(If you do the exercise, here's a tip that will save you time when you have to enter the item responses: use The Spreader.)

Lertap is not required to process survey results of the type seen in this example. Some researchers might use **SPSS** instead; its "Reliability Analysis" routine will produce results similar to those made by Lertap. However, reverse scoring survey questions in SPSS is not (in our opinion) straightforward; there is no equivalent to Lertap's *pol card. Of course, many readers will realise that SPSS does much more than reliability analysis—it's an extremely versatile, powerful data analysis system.

Most general-purpose data analysis systems, such as SPSS, make it fairly easy to exchange data with Excel. Consequently, entering survey results into Excel by no means closes the door to subsequently using SPSS, or **SAS**, or something similar.

The survey exemplified in this example was, we think, first used by "**MECC**", the Minnesota Educational Computing Consortium. As far as we know, these items may be used without copyright violations. (MECC was purchased by The Learning Company in the 1980s.)

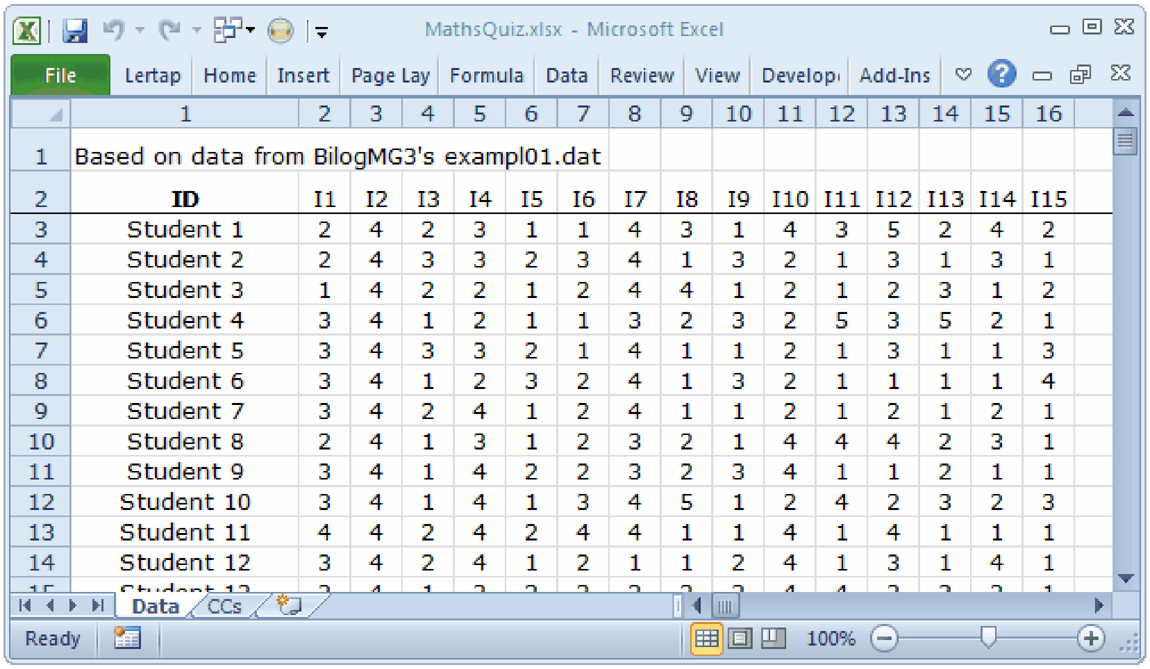
2.5 MathsQuiz

This little example is a nifty one.

About 1,000 high-school students responded to a short maths quiz, a test with just 15 items.

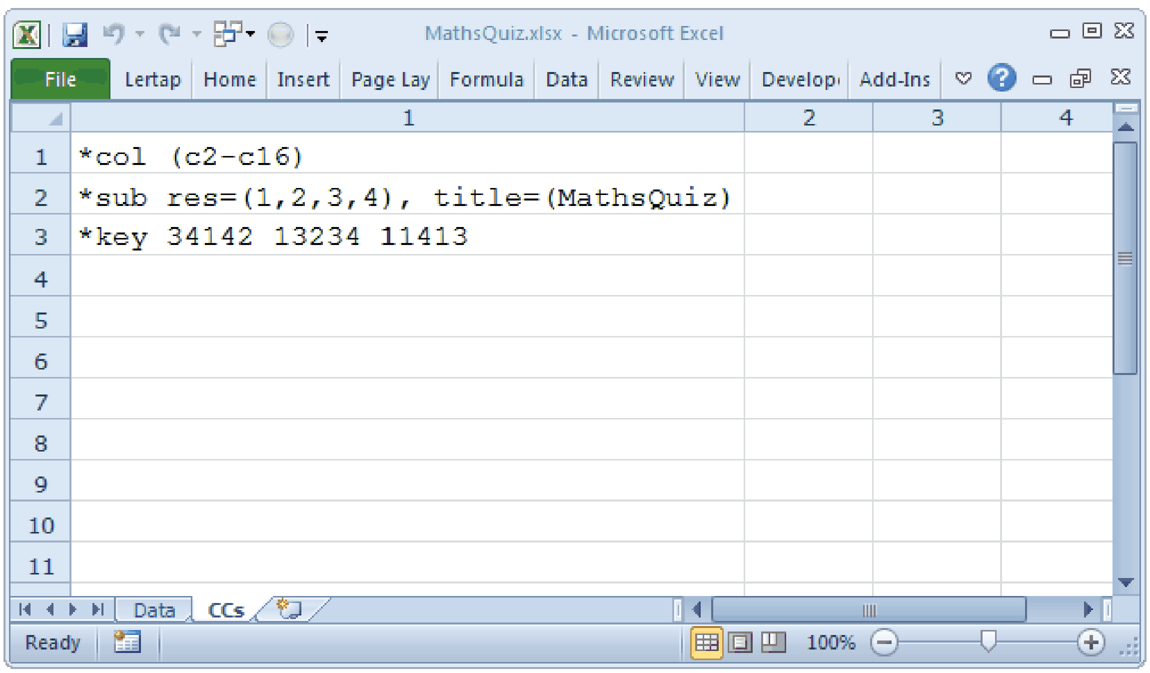
Each item used four responses, 1 2 3 4.

Snippets from the Data and CCs worksheets are shown below:



Based on data from BilogMG3's exampl01.dat

ID	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15
Student 1	2	4	2	3	1	1	4	3	1	4	3	5	2	4	2
Student 2	2	4	3	3	2	3	4	1	3	2	1	3	1	3	1
Student 3	1	4	2	2	1	2	4	4	1	2	1	2	3	1	2
Student 4	3	4	1	2	1	1	3	2	3	2	5	3	5	2	1
Student 5	3	4	3	3	2	1	4	1	1	2	1	3	1	1	3
Student 6	3	4	1	2	3	2	4	1	3	2	1	1	1	1	4
Student 7	3	4	2	4	1	2	4	1	1	2	1	2	1	2	1
Student 8	2	4	1	3	1	2	3	2	1	4	4	4	2	3	1
Student 9	3	4	1	4	2	2	3	2	3	4	1	1	2	1	1
Student 10	3	4	1	4	1	3	4	5	1	2	4	2	3	2	3
Student 11	4	4	2	4	2	4	4	1	1	4	1	4	1	1	1
Student 12	3	4	2	4	1	2	1	1	2	4	1	3	1	4	1



```
*col (c2-c16)
*sub res=(1,2,3,4), title=(MathsQuiz)
*key 34142 13234 11413
```

Did you notice the *sub line above? When the options for a cognitive test, called "**response codes**" in Lertap, are not A B C D (the default response codes for cognitive items), then an **res=** specification has to be made, and the *sub line is where it goes. Compare this *sub line with that **not** found in the ChemQuiz example. ChemQuiz does not need res= as its items use the default response codes for cognitive items.

(For affective items, the default response codes are 1 2 3 4 5.)

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

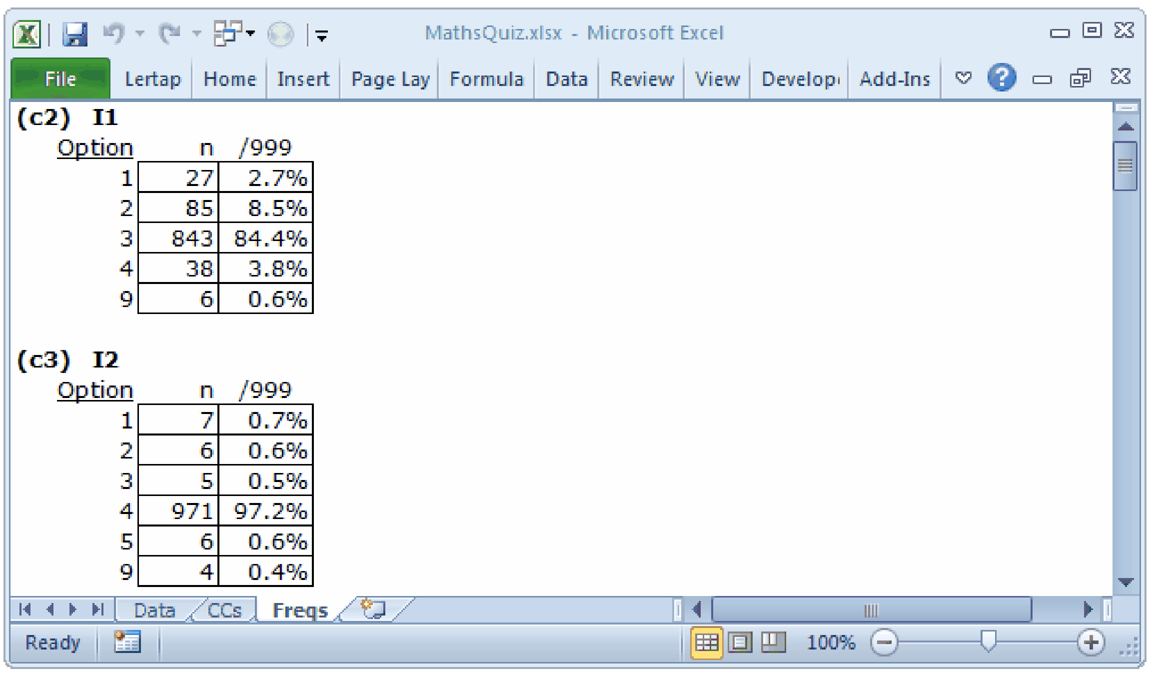
3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

Results were obtained by completing the three basic Lertap steps. Contents of selected "reports" (worksheets) are shown below:



(c2) I1

Option	n	/999
1	27	2.7%
2	85	8.5%
3	843	84.4%
4	38	3.8%
9	6	0.6%

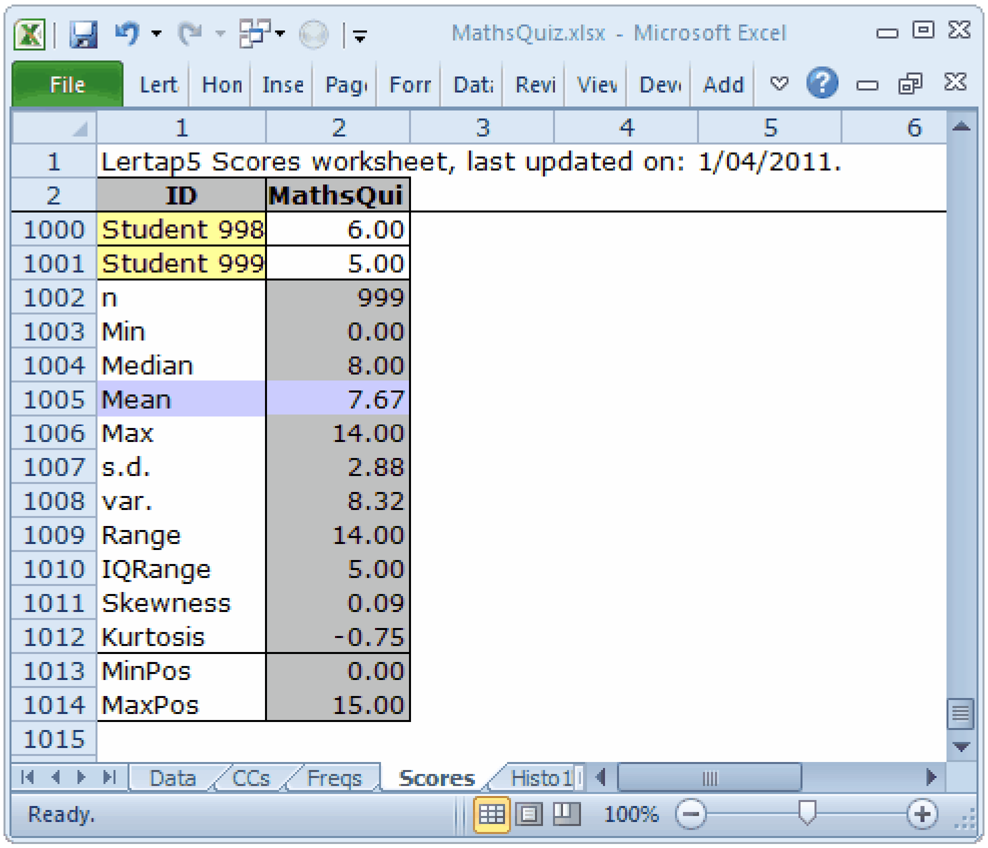
(c3) I2

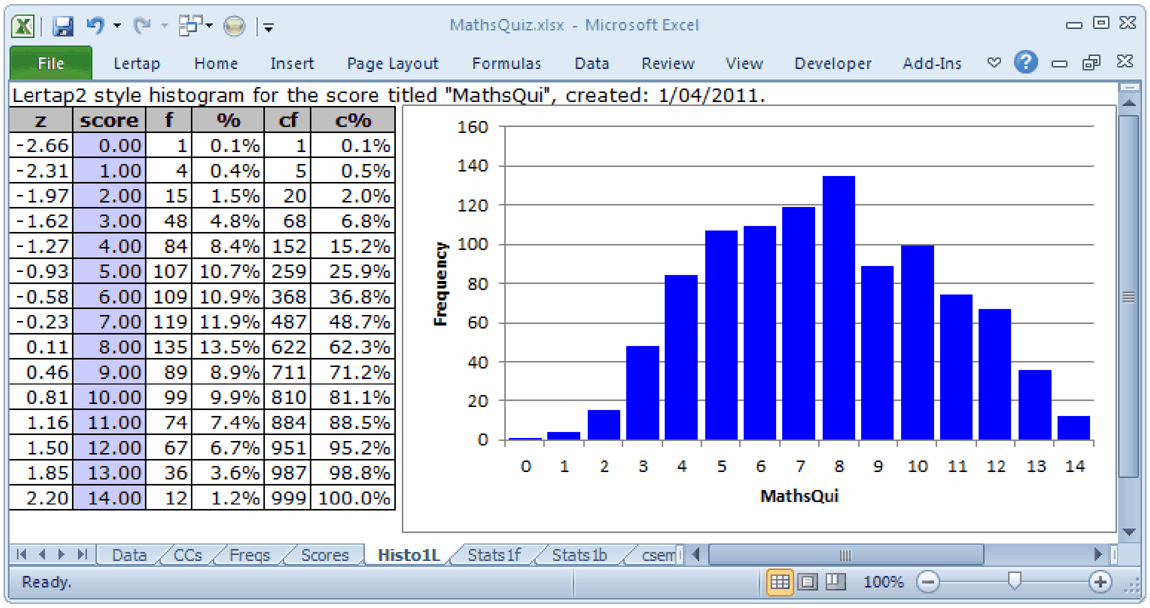
Option	n	/999
1	7	0.7%
2	6	0.6%
3	5	0.5%
4	971	97.2%
5	6	0.6%
9	4	0.4%

The data in this sample are from one of the datasets distributed with the Bilog-MG3 system; they're probably quite dated -- it's likely to be the case that the data were processed many moons ago.

So it is that we can only guess at the meaning of a "response" of 9. Almost all of the items had 9s; they probably indicate missing or bad data. But what about a "response" of 5? The results in the **Freqs** report indicate that 6 students appeared to select option 5 on the second item, I2, whose responses are found in column 3 (c3) of the **Data** worksheet.

Of course, there wasn't really a fifth option. It could be that mark-sense answer sheets were used to collect student responses, and then scanned. It could be that the mark-sense form allowed for five responses per item, even though only four were used on the actual test. This being the case, a response of 5 would be an error, and, were that true, 5s would be the same as 9s.





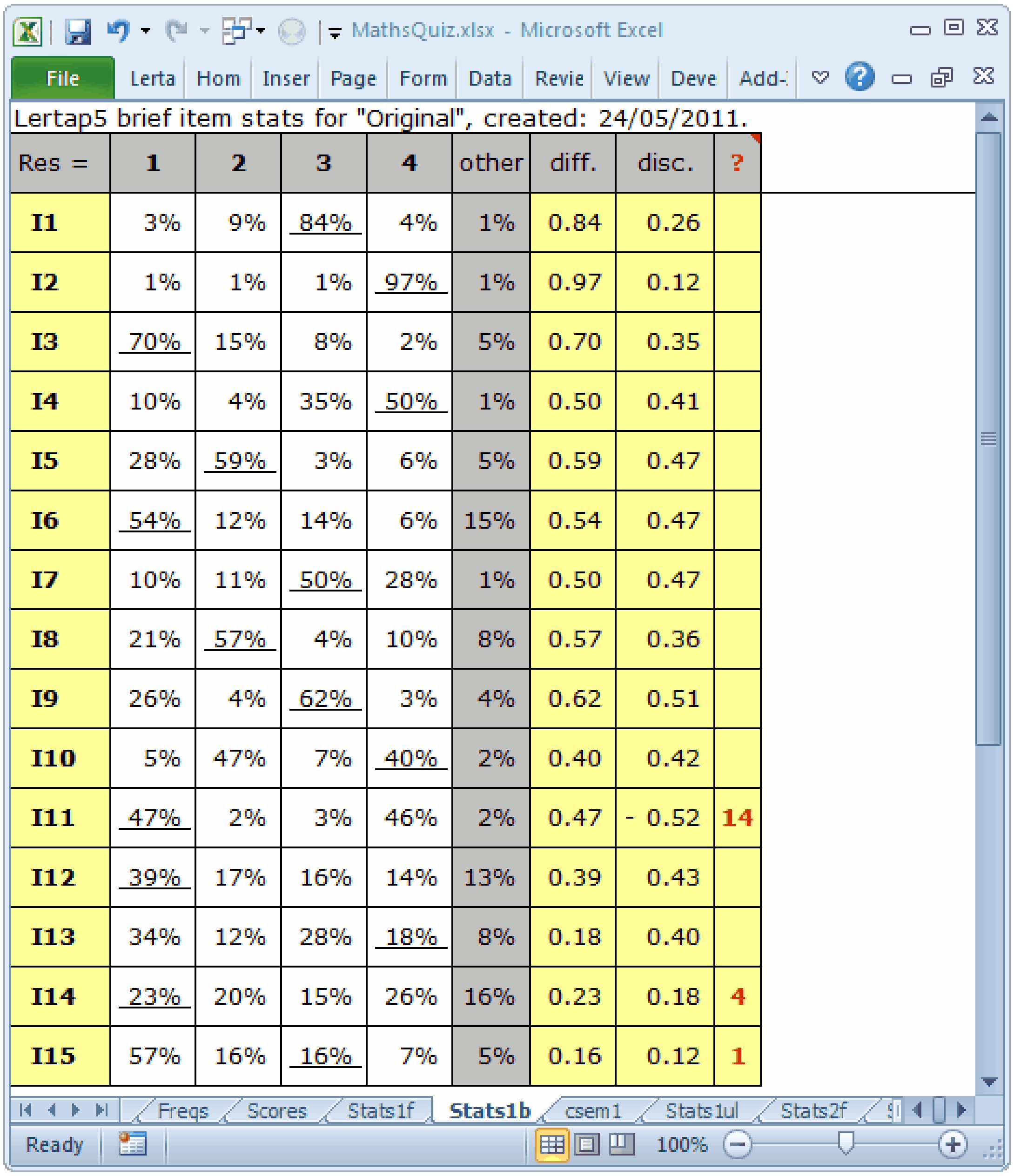
z	score	f	%	cf	c%
-2.66	0.00	1	0.1%	1	0.1%
-2.31	1.00	4	0.4%	5	0.5%
-1.97	2.00	15	1.5%	20	2.0%
-1.62	3.00	48	4.8%	68	6.8%
-1.27	4.00	84	8.4%	152	15.2%
-0.93	5.00	107	10.7%	259	25.9%
-0.58	6.00	109	10.9%	368	36.8%
-0.23	7.00	119	11.9%	487	48.7%
0.11	8.00	135	13.5%	622	62.3%
0.46	9.00	89	8.9%	711	71.2%
0.81	10.00	99	9.9%	810	81.1%
1.16	11.00	74	7.4%	884	88.5%
1.50	12.00	67	6.7%	951	95.2%
1.85	13.00	36	3.6%	987	98.8%
2.20	14.00	12	1.2%	999	100.0%

The students' scores were quite spread out, as seen above. "**MaxPos**", the maximum possible score, was 15, but "**Max**", the highest test score achieved was 14, as shown in the Scores report. At this point it seems that no-one got a perfect score.

Now, in answer to those 12 people who raised their hands to ask about the score's title showing as "MathsQui", when, according to the CCs sheet it's "MathsQuiz": the reason for the truncated title is that titles are limited to a length of just 8 characters. If **title=** is not used on the *sub line, then Lertap invents a title; usually it will be "Test1".

Even more people are asking about the Histo1L worksheet whose contents are portrayed immediately above. They're right to ask, too, as a **Histo1L** "report" does not result from following "The 3 steps". It's an extra. The way to get it is mentioned here.

Okay? What to look at next? Actually, after the 3 steps have completed, Lertap automatically gets Excel to set its focus on the **Stats1b** report. It does this because we (at Lertap HQ) reckon that the Stats1b report is the one most users are most anxious to see. So, let's look at Stats1b. Here's the top part of Stats1b for this little quiz:



Lertap5 brief item stats for "Original", created: 24/05/2011.

Res =	1	2	3	4	other	diff.	disc.	?
I1	3%	9%	84%	4%	1%	0.84	0.26	
I2	1%	1%	1%	97%	1%	0.97	0.12	
I3	70%	15%	8%	2%	5%	0.70	0.35	
I4	10%	4%	35%	50%	1%	0.50	0.41	
I5	28%	59%	3%	6%	5%	0.59	0.47	
I6	54%	12%	14%	6%	15%	0.54	0.47	
I7	10%	11%	50%	28%	1%	0.50	0.47	
I8	21%	57%	4%	10%	8%	0.57	0.36	
I9	26%	4%	62%	3%	4%	0.62	0.51	
I10	5%	47%	7%	40%	2%	0.40	0.42	
I11	47%	2%	3%	46%	2%	0.47	- 0.52	14
I12	39%	17%	16%	14%	13%	0.39	0.43	
I13	34%	12%	28%	18%	8%	0.18	0.40	
I14	23%	20%	15%	26%	16%	0.23	0.18	4
I15	57%	16%	16%	7%	5%	0.16	0.12	1

How to make use of these results? The answer depends on why we created and used the test to begin with.

If we're one of the maths teachers who used this little test, we might simply want to see where the students had a tough time. We might, for example, look for items where less than 50% of the students were able to identify the right answer. Then we'd focus on the subject content targeted by these items, and review it in class.

In this case, that would be I10 (40% correct), I11 (47%), I12 (39%), I13 (18%!), I14 (23%), and I15 (16%!).

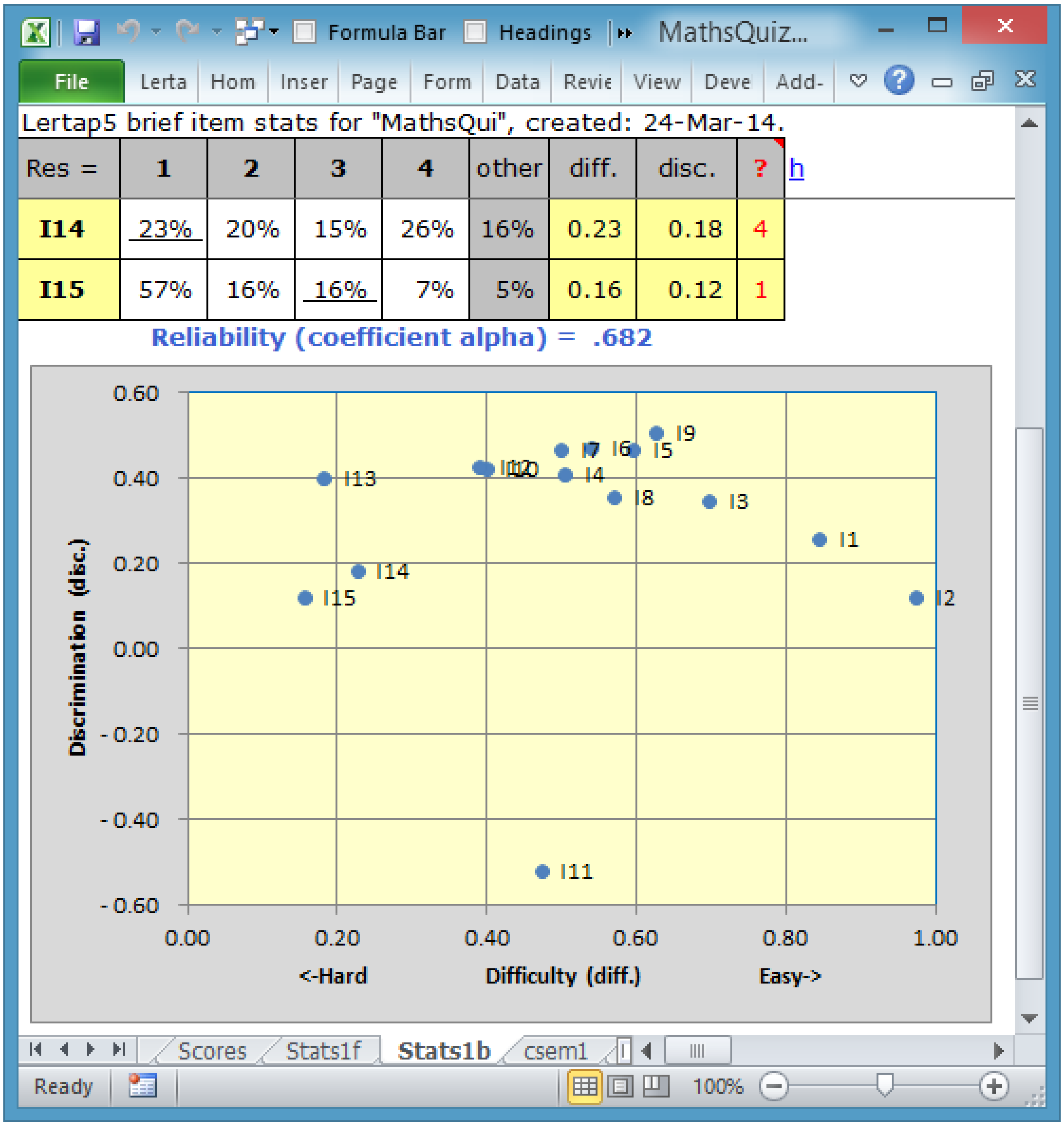
We might also take into account items where quite a few students had an "other" response, that is, did not choose one of the four options, 1 2 3 4. This would add I6 (15% other) to the mix.

Fair enough. The only potentially naive problem with this idea is that it assumes we have scored the items correctly, which is the same as saying that we didn't make an error when we entered the correct answers on the *key line in the CCs worksheet.

The **?** column will often (but not always) let us know if an item has been incorrectly keyed. Three items have an entry in the **?** column in this case: I11, I14, and I15. Of these, I11 was indeed incorrectly keyed: the right answer to this item was 4, not 1.

We will have more to say about the **?** column below.

Here's the bottom part of the Stats1b report:



The information in this scatterplot comes directly from the "diff." and "disc." columns of the Stats1b report. I15 has a diff. value of 16%, or, as a proportion, 0.16. Its disc. value is 0.12.

"**diff.**" means item difficulty. "**disc.**" means item discrimination.

"diff." is easy to interpret: it's the proportion of students who correctly answered the item. Easy items have diff. values above 0.80, which would be I1 and I2 here. Difficult items will have diff. values less than 0.20, items I13 and I15 in this example.

Many many hands are now waving in the air: "You mean to say that difficult items have low diff., while easy items have high diff.?"

Yes. Unfortunately diff. is a misnomer. Some authors have suggested that diff. should be called the facility index, but they have not held sway.

The maximum possible value for diff. is 1.00, corresponding to a percentage of 100%, something which happens when all students get an item right.

What about "disc."? It's a measure of how an item discriminates. If the most capable students get the item correct, while the least capable do not, the item is said to be discriminating. Such items are wanted when the purpose of the test is to identify the "best" students.

There are different ways to index item discrimination. A thorough discussion may be found in Chapter 7 of the Lertap manual (Nelson, 2000). Assuming you've read that chapter, you know that the disc. value displayed in Stats1b is the correlation between the item score and the criterion score, which is usually an internal one, being the number of the other items which the student got correct (it is possible to use an external criterion too). In this example, a common one, the items are scored on a right-wrong basis, with one point for the correct answer, zero points otherwise. This is called "dichotomous" scoring. The correlation between a dichotomous variable and an interval variable is generally referred to as a "point-biserial" correlation, abbreviated in Lertap Stats1f reports as "pb(r)", and referred to as "disc." in Stats1b reports.

Okay? Back to the **? column**: one or more of an item's options will appear in the **?** column whenever one of these conditions is met: no-one selects the option, the option is the correct answer but was selected by students with below-average criterion scores, or the option corresponds to a distractor (an incorrect option) selected by students with average or above-average criterion scores. The idea here is (basically) that we want an item's correct option to be selected only by the strongest students, while each of the distractors is selected by less than average students. Note that it is often the case that there will be many items with entries in the **?** column when the number of students who take the test is small, say less than 50 or so. When there aren't many students, an item with four or more options will very frequently have "dead" distractors, wrong answers which were not selected by anyone; these will show up in the **?** column.

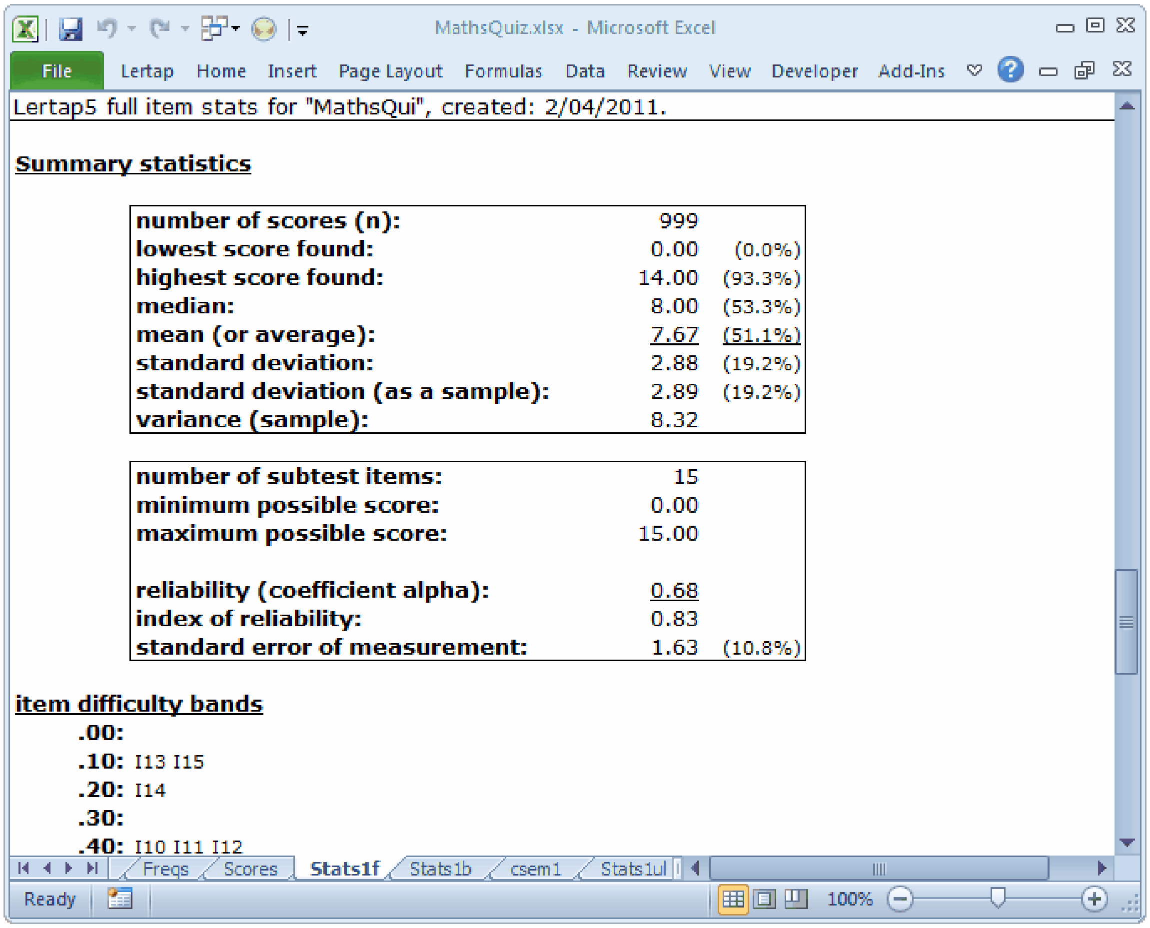
If a test is meant to discriminate, its items will have very few entries in the **?** column, and no items with negative disc. figures. Such tests will more often than not have good reliability, a subject which leads to the next topic: the Stats1f report.

2.5.1 Stats1f reports

"Stats1f" means full statistics, the big shebang. If the CCs worksheet refers to more than one subtest, as is the case with the CooksTour sample, there will be one "f" report for each subtest. Their names will be Stats1f, for the first subtest, Stats2f for the second, and so on.

A Stats1f report is a long one. It has complete, or "full", item statistics at the top, reliability information in the middle, and then some handy summary info towards the bottom.

When Lertap creates a Stats1f report, it automatically scrolls to the middle section, where the **reliability** (coefficient alpha) figure is found:



A test meant to discriminate, to pick out the strongest students, should have good reliability.

Lertap uses Cronbach's coefficient alpha index for its reliability figure; for details, refer to Chapter 7 of the Lertap manual.

This little test does *not* have good reliability. A figure of 0.68 is lousy if we want a test which discriminates.

Keep in mind that there will be times when Lertap users do not care to look at Stats1f results. The ChemQuiz sample is an example. There are certainly times when teachers, and others, can, and do, make do with just the Stats1b report.

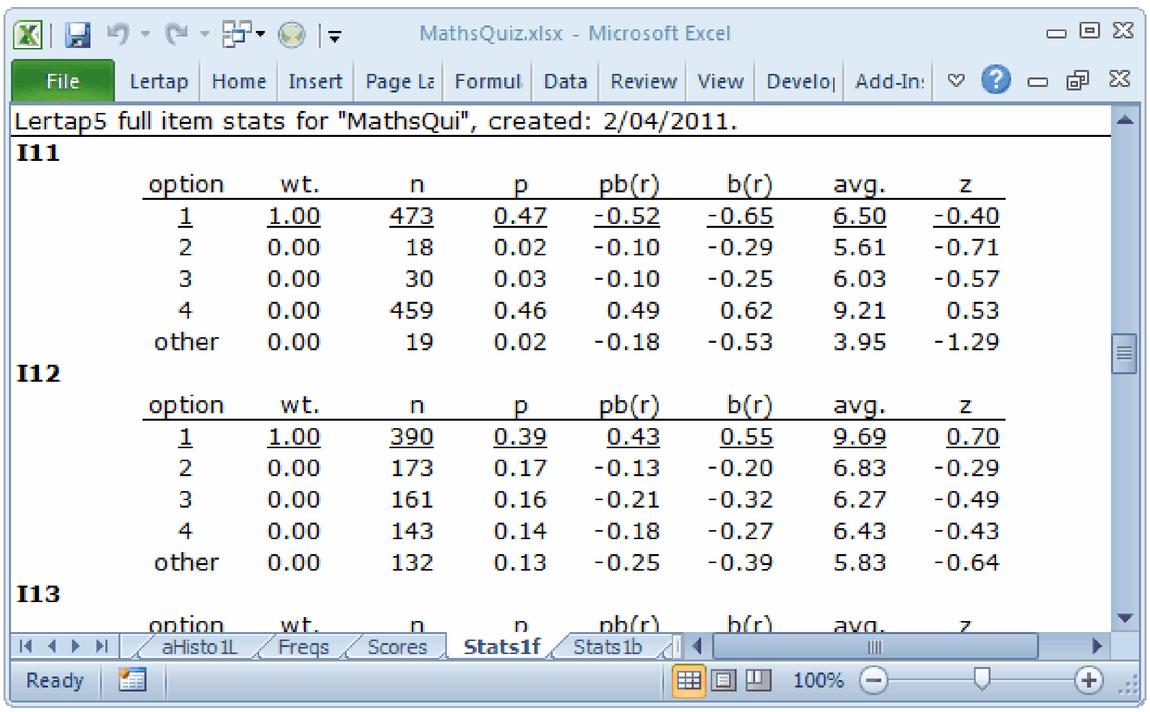
But in this case Discrimination R Us. We have a quiz whose purpose is to help teachers identify not just the strong students, but also the weak ones, perhaps with the goal of taking the poor performers into a special class session or two for some make-up work.

To do this we'd like a test with a reliability of 1.00. And, while we're at it, why don't we go for a new car, plus a first-class, around-the-world ticket on the Queen Mary? All of these are unlikely events. If we can get a test with an alpha reliability of 0.90 we're doing well; 0.95 is not unheard of, but not common.

For a 15-item quiz, many might be happy with an alpha of 0.80 or thereabouts. This is a short test, and short tests usually struggle to get into the really high reliability area (say, greater than 0.90). But 0.68 for a test meant to discriminate? Nope.

So now we ask if there might be something wrong with some of the 15 items. And yes, there is. That **? column** in the Stats1b report dobbed in three items: I11, I14, and I15.

Scroll the Stats1f report up to get the full stats for I11:



Lertap5 full item stats for "MathsQui", created: 2/04/2011.

I11

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	473	0.47	-0.52	-0.65	6.50	-0.40
2	0.00	18	0.02	-0.10	-0.29	5.61	-0.71
3	0.00	30	0.03	-0.10	-0.25	6.03	-0.57
4	0.00	459	0.46	0.49	0.62	9.21	0.53
other	0.00	19	0.02	-0.18	-0.53	3.95	-1.29

I12

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	390	0.39	0.43	0.55	9.69	0.70
2	0.00	173	0.17	-0.13	-0.20	6.83	-0.29
3	0.00	161	0.16	-0.21	-0.32	6.27	-0.49
4	0.00	143	0.14	-0.18	-0.27	6.43	-0.43
other	0.00	132	0.13	-0.25	-0.39	5.83	-0.64

I13

option	wt.	n	p	pb(r)	b(r)	avg.	z

The screen snapshot above is a picture of Lertap's full item stats. Complete details on interpreting these statistics are in that Chapter 7 mentioned above; there's also a bit about them right here, in this document.

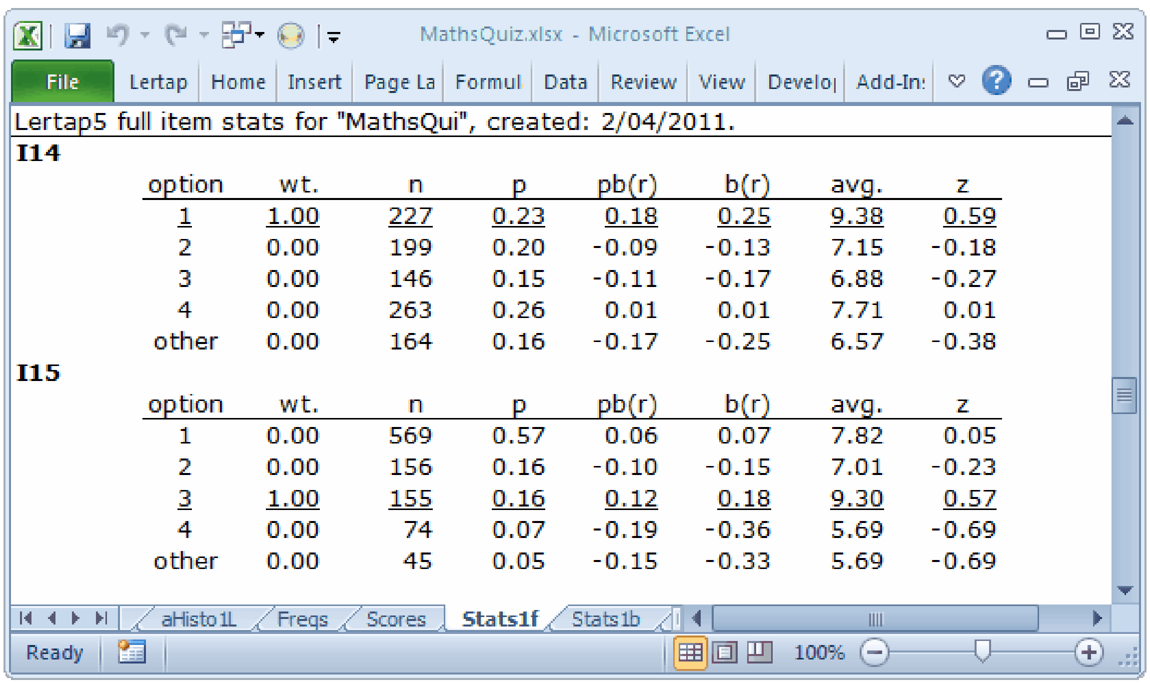
What's wrong with I11 is that the 473 students who selected the first option, the keyed-correct option, had an average test score of 6.50, which was below the overall test average of 7.67. I11's **pb(r)** value is -0.52, which is "way bad". You can see this "way-badness" in the little scatterplot at the base of the Stats1b report, shown in the previous topic. I11's blip in the scatterplot is far removed from the others.

This item was mis-keyed. The correct answer was 4, not 1.

Have a gander at the statistics for I11's option 4: it was selected by 459 students with an average test score of 9.21. These were strong students.

Looking down the "**avg.**" column, what we'd like to see for a discriminating item is that the highest value is underlined, meaning that those students who got the correct answer (underlined) were the strongest. We'd also want to look the **z** column in the eye, and find all but one of the z's negative in value; this would mean that the people who took the distractors had average test scores below the mean (that is, below the overall test average, in this case 7.67), while those who got the item right had above average test scores. I12 is a good example of the patterns we want to see. (Note that 132 students didn't answer I12, representing 13% of the test takers. This figure is high. Also note that their average test score of 5.83 was the weakest of the lot for I12; the implication here is that a fair portion of the weaker students were unwilling to even take a guess on this item.)

What about I14 and I15, two other items flagged in the ? column in Stats1b? Stats1f says this:



Lertap5 full item stats for "MathsQui", created: 2/04/2011.

I14

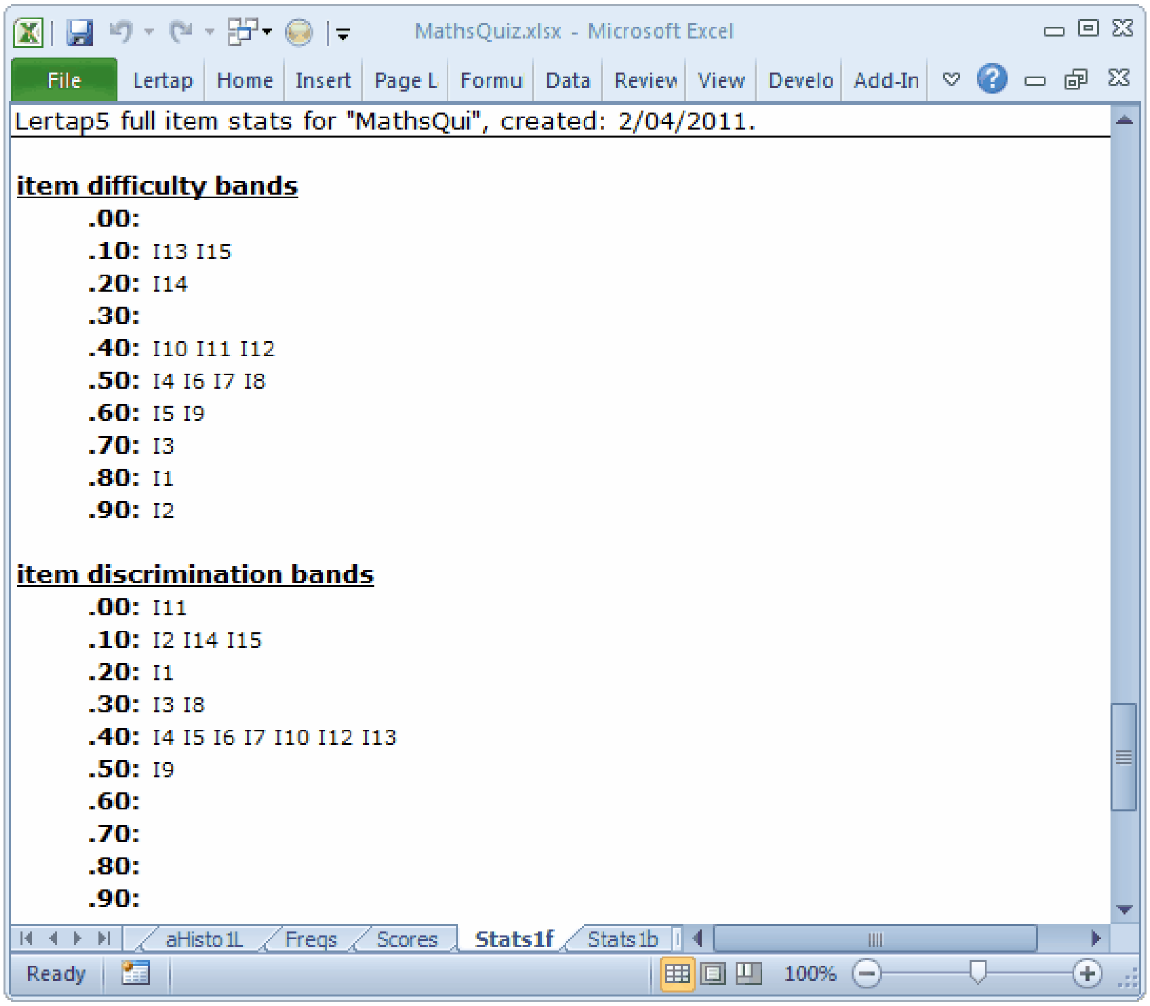
option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	227	0.23	0.18	0.25	9.38	0.59
2	0.00	199	0.20	-0.09	-0.13	7.15	-0.18
3	0.00	146	0.15	-0.11	-0.17	6.88	-0.27
4	0.00	263	0.26	0.01	0.01	7.71	0.01
other	0.00	164	0.16	-0.17	-0.25	6.57	-0.38

I15

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	569	0.57	0.06	0.07	7.82	0.05
2	0.00	156	0.16	-0.10	-0.15	7.01	-0.23
3	1.00	155	0.16	0.12	0.18	9.30	0.57
4	0.00	74	0.07	-0.19	-0.36	5.69	-0.69
other	0.00	45	0.05	-0.15	-0.33	5.69	-0.69

What's happened here is that each of these items has a distractor selected by students whose average test scores were above the test mean of 7.67. In the case of I14, it's option 4. For I15 it's option 1.

These are undesirable outcomes; these two items were difficult for the students -- that's not bad at all, but the fact some capable (that is, above average) students opted for a distractor would indicate a good reason to review these items. There may well be some ambiguity in the items, a bit of wording, perhaps, abundantly clear in the item writer's head, but confusing to the students. Often a good way to uncover ambiguity of this sort is to ask the students why they thought the distractor was the best answer.

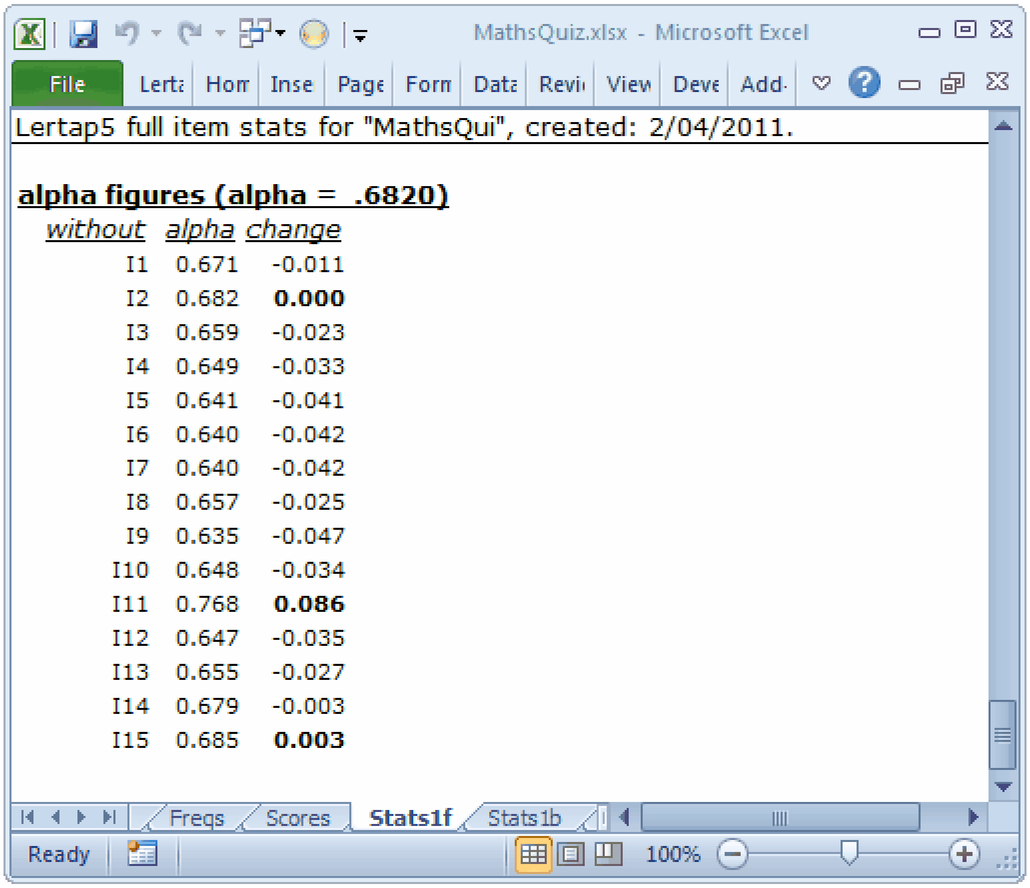


If you thought you heard music playing whilst looking at Stats1f, no doubt it was from the bands found towards the bottom of the report.

The information in these two bands is very similar to the information provided in the scatterplot at the bottom of Stats1b.

I13 and I15 had difficulties between 0.10 and 0.19. Two items, I1 and I2, were easy, having difficulties at or above 0.80.

I9 had the highest discrimination figure. I11 had the lowest, but there's a problem here: the discrimination bands are incapable of hitting really low notes. I11's discrimination figure was -0.52. This is a truly low figure. The little band shown above does not indicate exactly how low I11 dipped. In this regard, the scatterplot found at the base of the Stats1b report is better -- when item discriminations go negative, the scatterplot will so indicate -- but the Stats1f band will not.


MathsQuiz.xlsx - Microsoft Excel

Lertap5 full item stats for "MathsQui", created: 2/04/2011.

alpha figures (alpha = .6820)

without	alpha	change
I1	0.671	-0.011
I2	0.682	**0.000**
I3	0.659	-0.023
I4	0.649	-0.033
I5	0.641	-0.041
I6	0.640	-0.042
I7	0.640	-0.042
I8	0.657	-0.025
I9	0.635	-0.047
I10	0.648	-0.034
I11	0.768	**0.086**
I12	0.647	-0.035
I13	0.655	-0.027
I14	0.679	-0.003
I15	0.685	**0.003**

Freqs Scores Stats1f Stats1b

Ready 100%

Finally, at the bottom of the bottom of Stats1f reports, there's a wee table which shows what would happen to the test's alpha figure if an item were to be totally removed from the test.

Here we start with an incoming total-test alpha of .6820. Take out the first item, I1, and alpha would fall to 0.671 (so, we wouldn't want to take out I1).

Take out I11, on the other hand, and alpha goes up to 0.768. Any item with a positive value in the "change" column would be one we might want to re-examine. Why? Because a positive change means that alpha would actually increase if the item were excluded from the test.

Summary remarks

The reliability figure for this administration of the test is considerably lower than what would be desired. In large part this is due to one item, I11, being incorrectly keyed. Were we to put in the correct key for this item, changing the 11th entry in the *key line on the CCs worksheet from 1 to 4, coefficient alpha would get a nice boost, a bit more than the 0.768 figure just mentioned. We leave it to you to confirm this: page ahead to the download topic, get a copy of the workbook, make the change in the *key line, and see what happens.

The matter of having two items, I14 and I15, with distractors selected by above-average students might be a cause of some concern. It could be suggested that the students who selected these options have been disadvantaged, maybe even "penalized". We'd be tempted to think this if, on reviewing the items and maybe speaking with students, it became apparent that the distractors could well be seen as correct answers. We could use an ***mws** line to double-key each of these items, an action which will tell Excel that these items have two correct answers. For each of these "correct answers" we will give one point:

*col (c2-c16)
*sub res=(1,2,3,4), title=(MathsQuiz)
*key 34142 13234 **1**1413
*mws c15, 1, 0, 0, 1
*mws c16, 1, 0, 1, 0

The first *mws line refers to the item found in column 15 (c15) of the Data worksheet, which is I14. It says give one point for the first and fourth options, zero points for the second and third options. For I15, located in c16 of the Data worksheet, the second *mws line will give one point to those who selected the first option, zero points to those going for the second option, one point for the third option, and zero points for the last option.

These changes will impact the test scores, and they might very well also affect coefficient alpha. How? See for yourself: page ahead to the download topic, get a copy of the workbook, add the *mws lines, and see what results.

Finally, a suggestion: the following CCs lines add a new subtest, making it possible to retain the original reports. If you try these lines, you'll see that the Scores worksheet will include three columns, one for MathsQui, one for MathsQ2, and a new one, Total. Don't want the Total? Change each *sub line so that there's Wt=0 at the end. For example: `*sub res=(1,2,3,4), title=(MathsQuiz), Wt=0`

*col (c2-c16)
*sub res=(1,2,3,4), title=(MathsQuiz)
*key 34142 13234 **1**1413

```
*col (c2-c16)
*sub res=(1,2,3,4), title=(MathsQ2)
*key 34142 13234 41413
*mws c15, 1, 0, 0, 1
*mws c16, 1, 0, 1, 0
```

Tidbits:

The matter of improving coefficient alpha by eliminating test items is further discussed in a Lertap erudite epistle. It's the "*Coefficient alpha, eigenvalues, scree tests*" paper. Download this PDF technical marvel. Read about Lertap and **eigenvalues** here.

There's much more about *mws in the manual, and here (see Example C10 and following discussion).

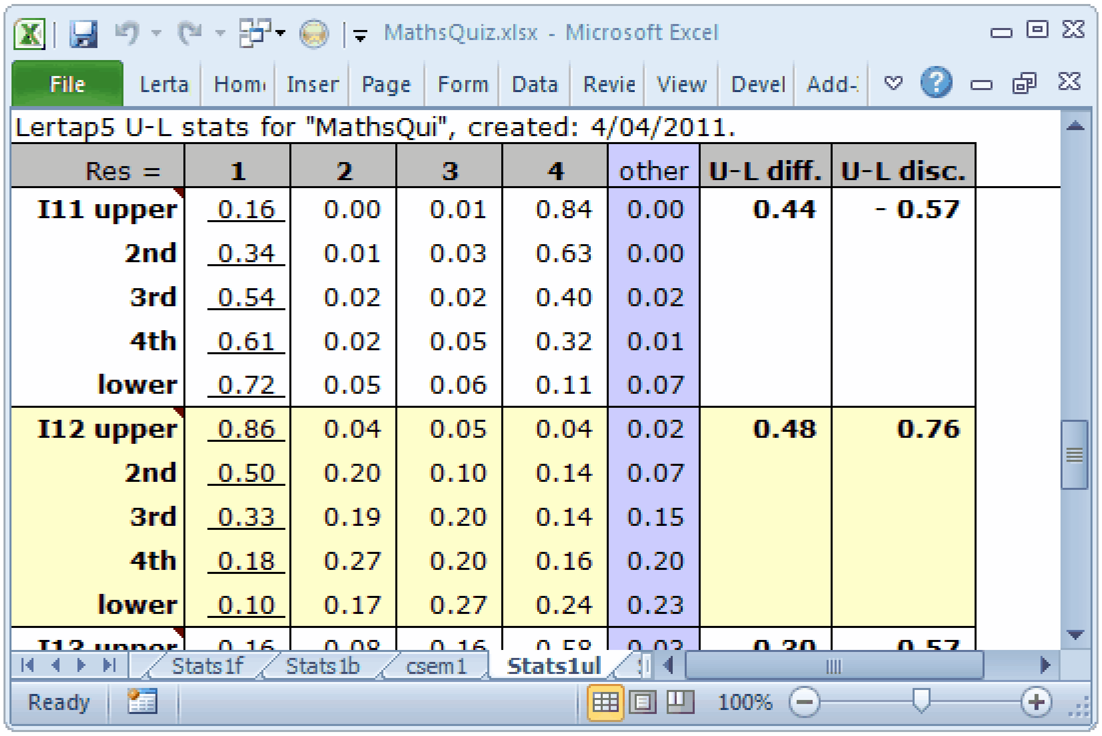
2.5.2 Stats1ul

"Stats1ul" means upper-lower statistics. If the CCs worksheet refers to more than one subtest, as is the case with the CooksTour sample, there will be one "ul" report for each subtest. Their names will be Stats1ul, for the first subtest, Stats2ul for the second, and so on.

The Stats1b and Stats1f reports use correlation coefficients to index how items are discriminating. Stats1ul reports use quite a different method. They're based on a procedure used before we had personal computers. A suggested method for determining if an item was discriminating involved these steps: (1) score all of the tests (by hand, of course); (2) take out the top 27% of the test papers, calling this the upper group; (3) take out the bottom 27% of the test papers, calling this the lower group; (4) for each item, find the proportion of students in each group who got the item correct, call these proportions Prop(upper) and Prop(lower); (5) finish by subtracting Prop(lower) from Prop(upper).

Why 27%? Garret (1952, footnote on p.215), wrote: "*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 ....*". The Lertap manual has more about this in Chapter 10; Chapters 3 and 7 of the manual also discuss U-L, the upper-lower method.

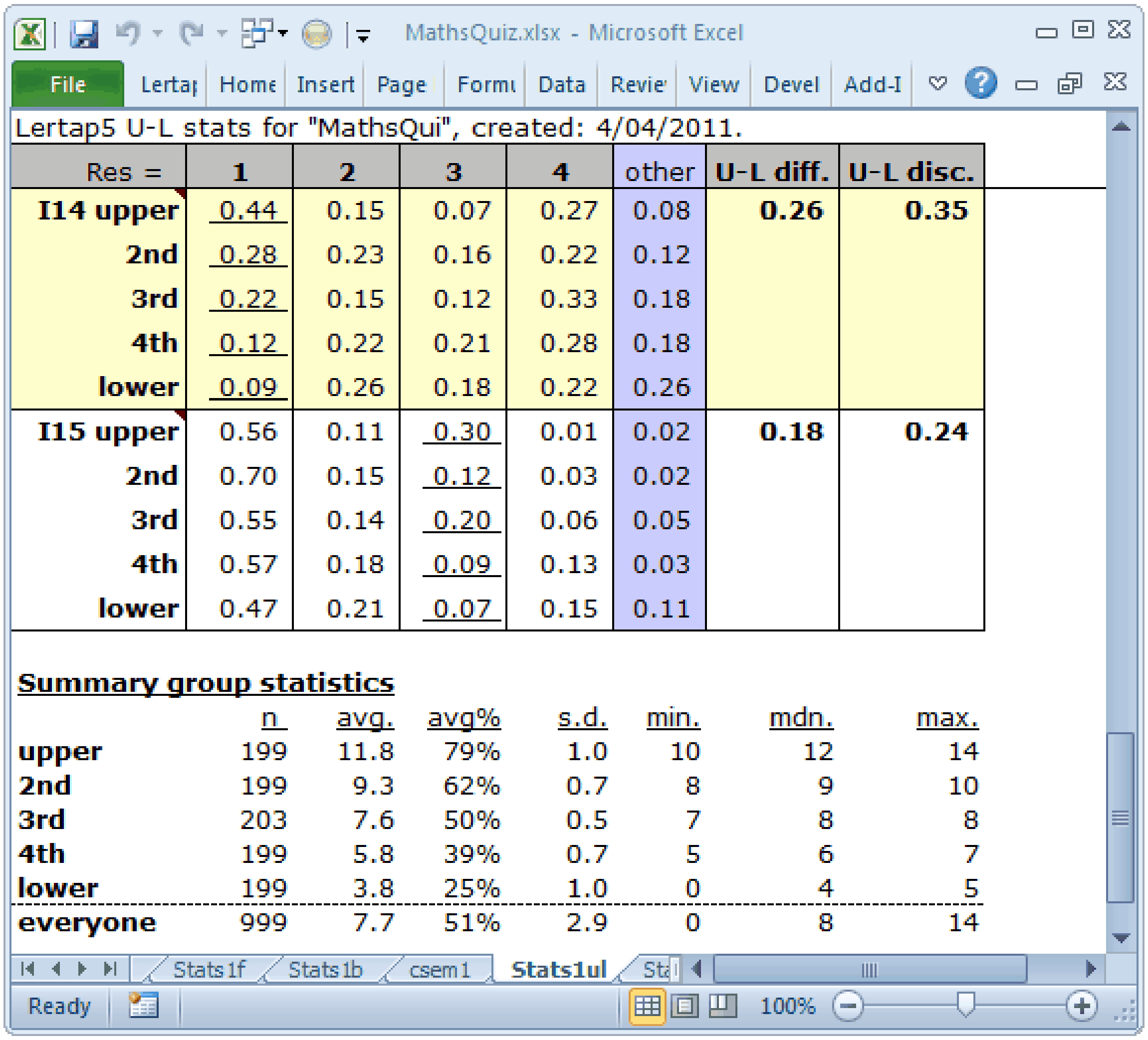
Despite this well-regarded historical antecedent, when you get Lertap at the store, take it home and unwrap it, you will find that Lertap's U-L report looks like this:



Lertap5 U-L stats for "MathsQui", created: 4/04/2011.

Res =	1	2	3	4	other	U-L diff.	U-L disc.
I11 upper	0.16	0.00	0.01	0.84	0.00	**0.44**	**- 0.57**
2nd	0.34	0.01	0.03	0.63	0.00		
3rd	0.54	0.02	0.02	0.40	0.02		
4th	0.61	0.02	0.05	0.32	0.01		
lower	0.72	0.05	0.06	0.11	0.07		
I12 upper	0.86	0.04	0.05	0.04	0.02	**0.48**	**0.76**
2nd	0.50	0.20	0.10	0.14	0.07		
3rd	0.33	0.19	0.20	0.14	0.15		
4th	0.18	0.27	0.20	0.16	0.20		
lower	0.10	0.17	0.27	0.24	0.23		
I13 upper	0.16	0.08	0.16	0.58	0.02	**0.30**	**0.57**

Lertap's default U-L activity is to separate the test takers into quintiles, five groups, with about 20% in each. There's always a small table at the bottom of a Stats1ul report which summarises this:



MathsQuiz.xlsx - Microsoft Excel

Lertap5 U-L stats for "MathsQui", created: 4/04/2011.

Res =	1	2	3	4	other	U-L diff.	U-L disc.
I14 upper	0.44	0.15	0.07	0.27	0.08	**0.26**	**0.35**
2nd	0.28	0.23	0.16	0.22	0.12		
3rd	0.22	0.15	0.12	0.33	0.18		
4th	0.12	0.22	0.21	0.28	0.18		
lower	0.09	0.26	0.18	0.22	0.26		
I15 upper	0.56	0.11	0.30	0.01	0.02	**0.18**	**0.24**
2nd	0.70	0.15	0.12	0.03	0.02		
3rd	0.55	0.14	0.20	0.06	0.05		
4th	0.57	0.18	0.09	0.13	0.03		
lower	0.47	0.21	0.07	0.15	0.11		

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	199	11.8	79%	1.0	10	12	14
2nd	199	9.3	62%	0.7	8	9	10
3rd	203	7.6	50%	0.5	7	8	8
4th	199	5.8	39%	0.7	5	6	7
lower	199	3.8	25%	1.0	0	4	5
everyone	999	7.7	51%	2.9	0	8	14

Stats1f | Stats1b | csem1 | Stats1ul | Sta

Ready 100%

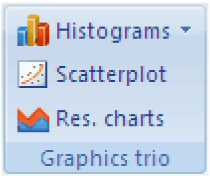
The 199 "upper" students, those with the top test scores, had an average of 11.8, or 79% of 15, the maximum possible score. The lower quintile, housing those 199 scores at the lowest end of the score distribution, had an average of just 3.8, or 25%.

We can use these little tables to re-visit the items discussed in the previous topic.

I11 is clearly whomperjawed. The keyed-correct option, 1, was selected by just 16% of those in the upper group, compared with 72% in the lower group. On the other hand, look at the response pattern for I11's option 4: 84% endorsement in the top quintile, only 11% in the lower. We mentioned before that this item was incorrectly keyed, and here we can see it readily. If we want to have a discriminating item, it should be the strongest students, the upper quintile, who get the item right.

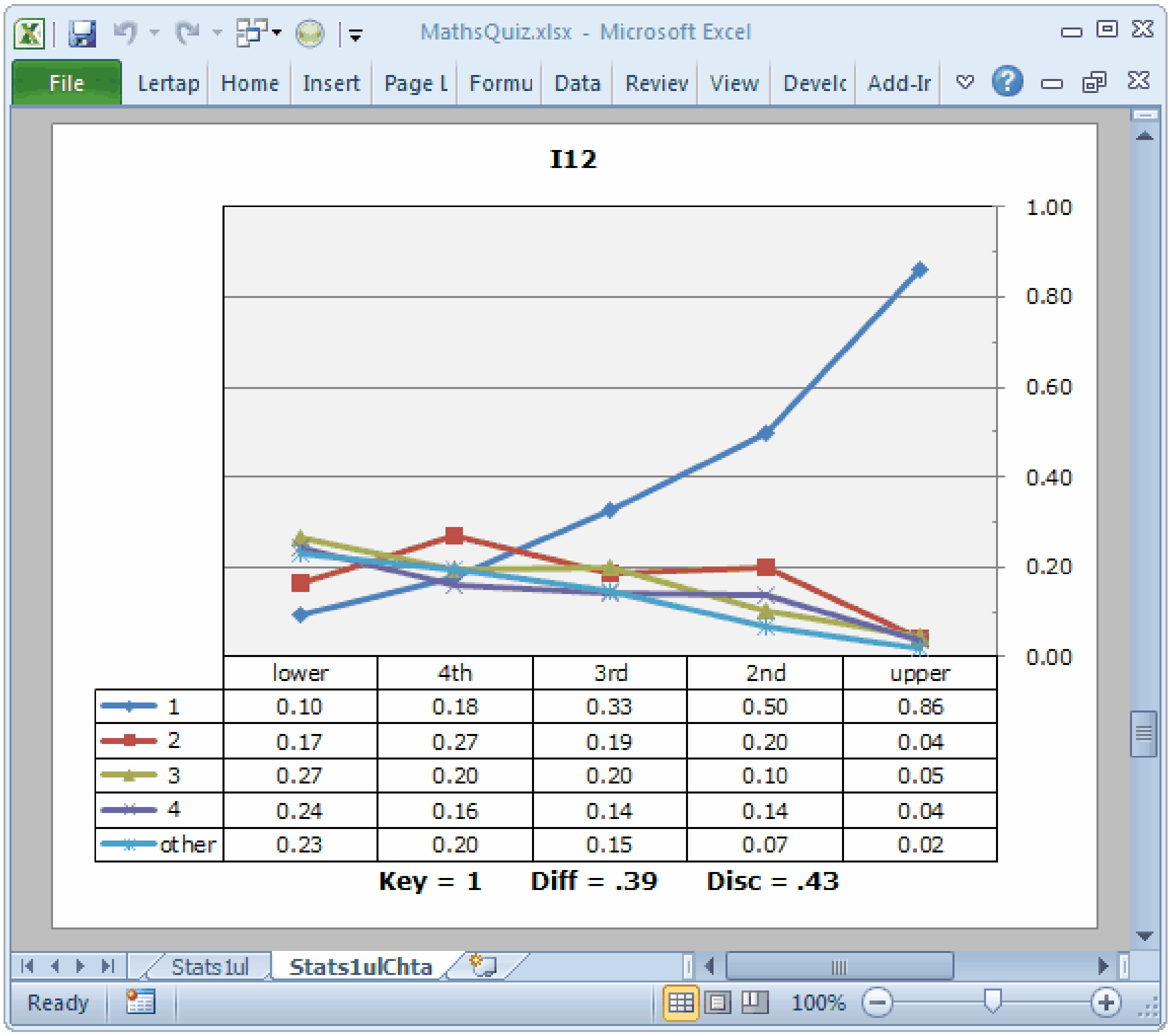
I14 and I15 have positive U-L disc. values, but note: were the correct answer to I14 the fourth option, the U-L disc. would still be positive (0.27 minus 0.22). Were the correct option on I15 the first option, the U-L disc. would be again be positive (0.09). These results run counter to what we want to see. For a distractor, we want the proportion selecting it to be lowest in the upper group, highest in the lower group.

Would it not be nice to have a picture to go along with these tables? Can do? You bet. We'll get some "quintile plots".



While looking at a Stats1ul report, go to the "Graphics Trio" part of the Lertap tab, and select the last option, "Res. charts".

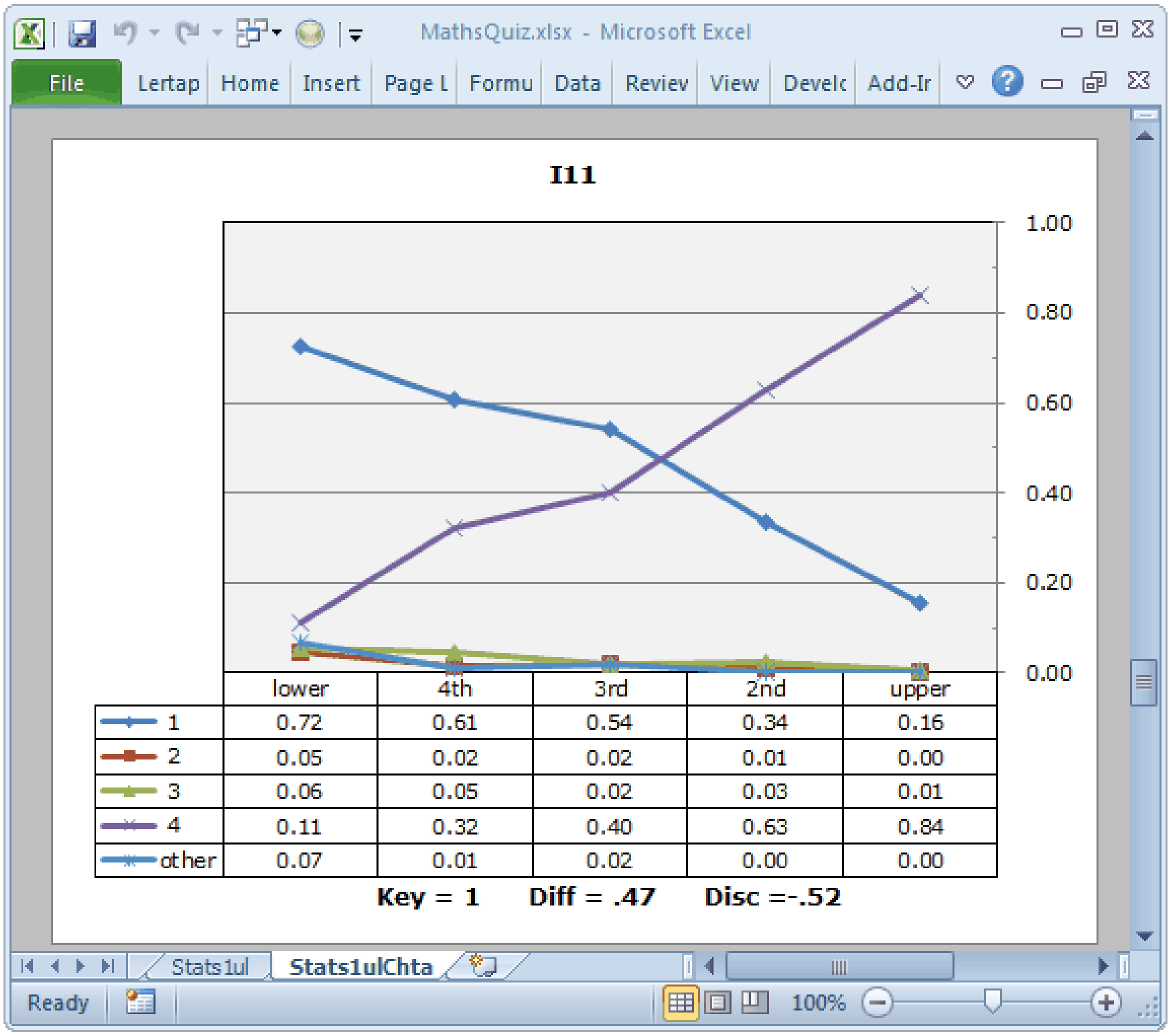
Lertap and Excel will team up to make pictures which will surely rival those from your last holiday. Behold:



This "picture" is a quintile plot.

Look at the blue line above. It corresponds to the keyed-correct answer for I12. As the students gain in proficiency, their ability to identify the correct option increases. That is, as we go from left to right, from the lower group to the upper one, the line in the plot which traces the proportions for the correct answer, Key=1 in the table, rises, from 0.10 on the left, to 0.86 on the right. This is a good plot. It's what we want to see for a discriminating item.

(The Diff. and Disc. values seen at the bottom of the plot come from the Stats1b report.)

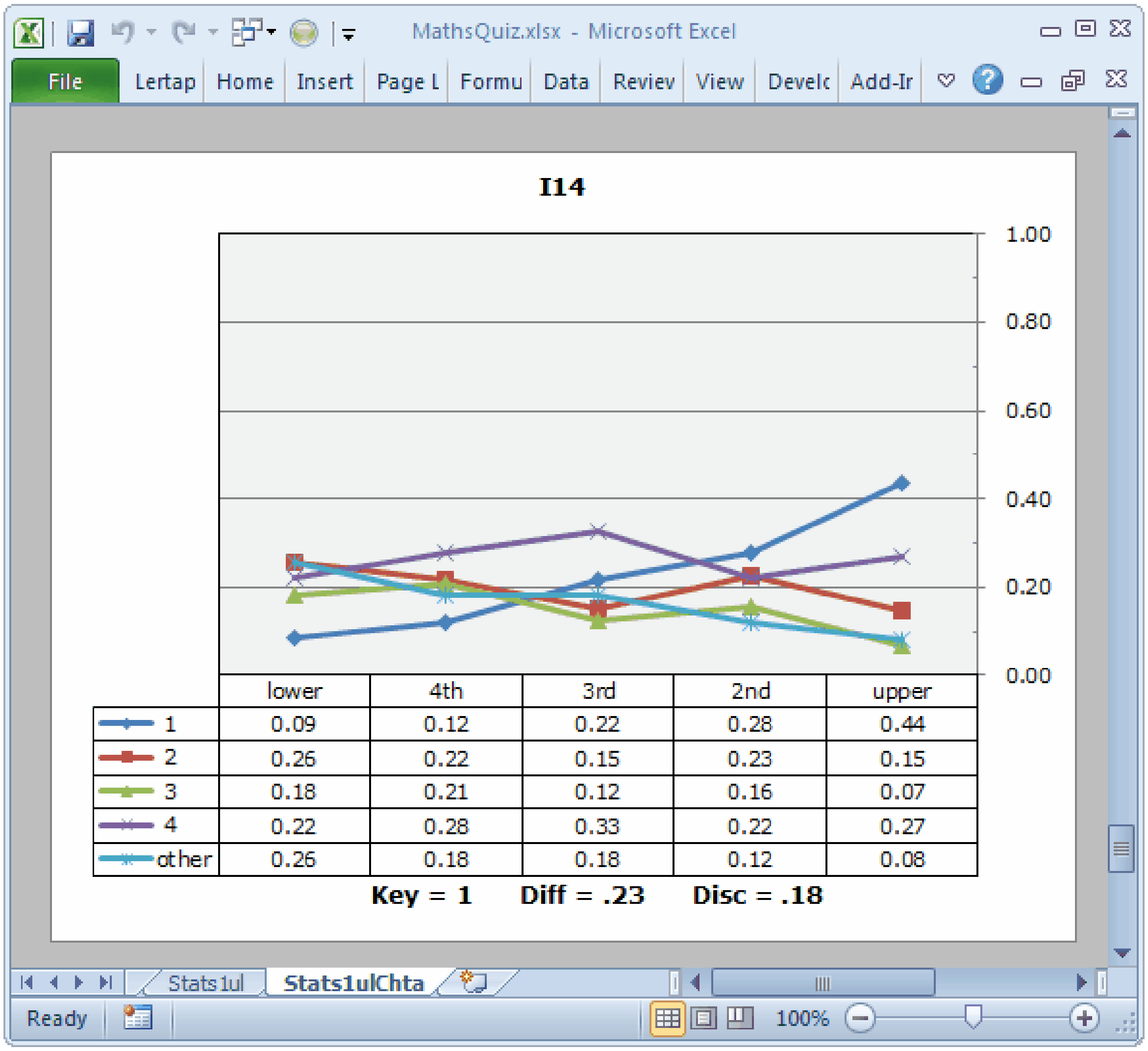


One of the first things to note about I11's plot is that only two of the item's options, 1 and 4, attracted any students to speak of. This isn't uncommon; it's not ideal, but it certainly happens.

The next thing to note is that the line which rises from left to right corresponds to a distractor, option 4. The line for the keyed-correct option, 1, has the pattern expected of a distractor. This is the mis-keyed item. Were we to correct the error, the plot would look the same, but the Disc. value at the base of the plot would be positive, and we'd see Key=4 instead of Key=1.

If you take on the challenge of downloading the MathsQuiz workbook, and repeating the various things we've done here, you will notice that most of the test's items have a pattern similar to I12's. We know people who base their item analysis very largely on an examination of these quintile plots, using the plots to visually identify good and bad items. While this is certainly a solid approach to item analysis, some care has to be taken: just scrolling down the plots without looking at the Disc. values can lead to problems -- for example, a simple scroll of this test's quintile plots would not uncover the error in I11, an item whose plot appears to be okay in that it has one trace line which rises, and almost dramatically, from left to right. The problem this item has is not apparent without looking at the Disc. value, or the row in the table corresponding to Key=1.

Items which may have some ambiguity in them will generally have a plot similar to I14's, as shown below:



This is a hard item, having a Diff. of just 0.23. The Disc. value is positive, and the trace line for the keyed-correct option, 1, does indeed rise as we move from left to right. However, the traces for the distractors do not all decrease from left to right. One of them, corresponding to option 4, actually rises. Another distractor, option 2, remains fairly strong in the upper group, with an endorsement of 15%. These figures imply that the top students were split on this item, with more than a quarter of them selecting option 4, a distractor.

Forward!

Tidbits:

The criterion score used to form the U-L groups is generally an internal one, the test score itself. However, it is possible to use an external criterion.

While the default U-L number of groups is 5, it can be set to a smaller number by changing an entry in the System worksheet. And, it's possible to reverse the rows and columns in the tables which underpin these plots, getting another type of picture (neat). Read about such morsels here.

The matter of using quintile plots as a primary means of identifying good and not-so-good items is further discussed in a Lertap "erudite epistle". It's the "*Visual item analysis with quintile plots*" paper. Download this PDF.

2.5.3 csem1

csem stands for "conditional standard error of measurement".

A test administration whose reliability estimate is less than 1.00, less than perfect, has measurement error. Were we able to test the same students again, with the same items, trusting there to be no memory effect (students will not remember having seen the items before), student scores can be expected to change. Our testing is fallible. If it weren't, the reliability coefficient would be 1.00. (Ours is a paltry 0.68 in this example.)

Let's say we took a single student, and tested him or her repeatedly (an almost infinite number of times, say) with the same test. Again assuming there to be no memory effect, and acknowledging that the test is fallible, there will be a distribution of test scores for our student. Each of these test scores is referred to as an "**observed score**".

In classical test theory, **CTT**, the mean of the observed scores is the "**true score**", the score which would be expected if the test administration were perfectly reliable.

The standard deviation of the observed scores is **SEM**, the **standard error of measurement**.

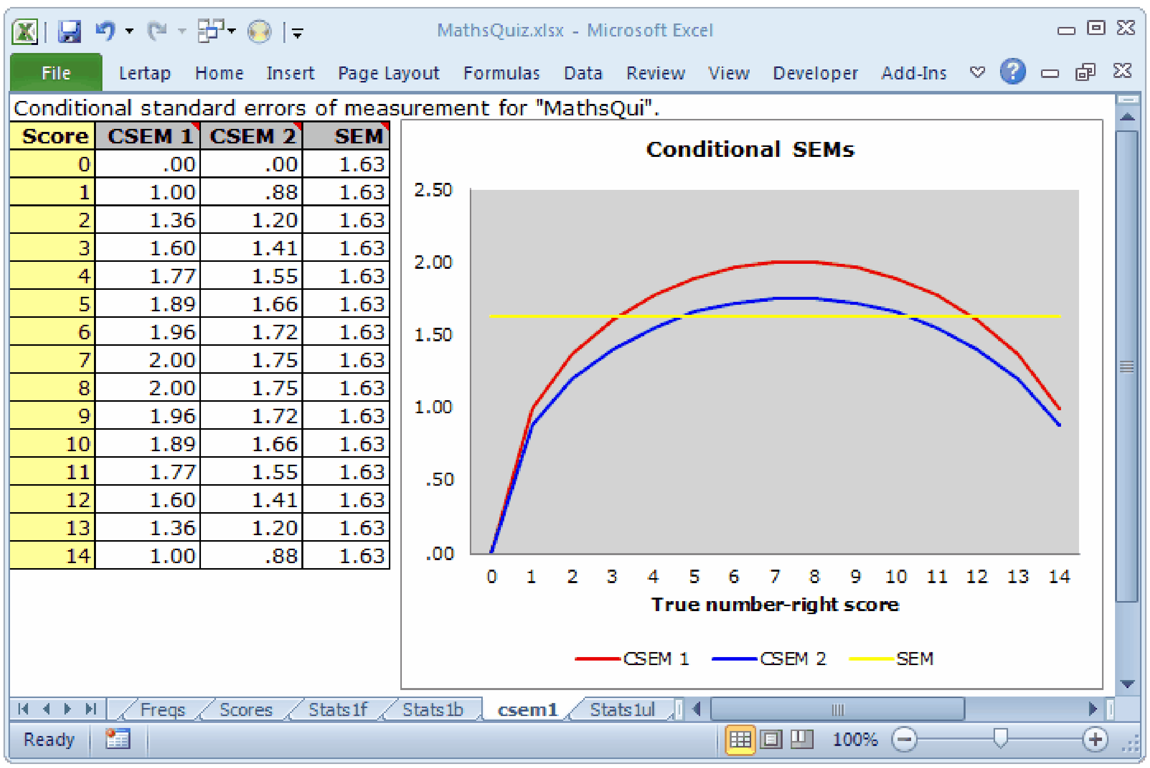
For our little MathsQuiz, Lertap's Stats1f report says that the standard error of measurement is 1.63, or 10.8 on a percentage-correct scale.

Your read of the Lertap manual will have made you full-bottle on the usual standard error of measurement, SEM, and the invaluable role it has to play in the interpretation of cognitive test scores. (See especially Chapter 7 of the manual; also see the tidbit below.)

Now, it has long been known that standard errors of measurement vary by ability. As we go from test takers with little proficiency, to those with average proficiency, and then on to the strongest students, respective SEMs change. The SEM commonly used in CTT is but an average figure, and, until 2006, it was the only measurement error estimate provided by Lertap.

These days Lertap computes SEMs for various score levels using methods from a paper by Lord (1984). More exactly, Lertap employs Lord's Method III, the binomial error model, and also the adjustment to Method III estimates, known as Method IV. These SEMs are called "conditional" SEMs as they vary, depending on true score levels.

Have a look at an example of Lertap output for CSEMs:


Conditional standard errors of measurement for "MathsQui".

Score	CSEM 1	CSEM 2	SEM
0	.00	.00	1.63
1	1.00	.88	1.63
2	1.36	1.20	1.63
3	1.60	1.41	1.63
4	1.77	1.55	1.63
5	1.89	1.66	1.63
6	1.96	1.72	1.63
7	2.00	1.75	1.63
8	2.00	1.75	1.63
9	1.96	1.72	1.63
10	1.89	1.66	1.63
11	1.77	1.55	1.63
12	1.60	1.41	1.63
13	1.36	1.20	1.63
14	1.00	.88	1.63

Charts such as this one will be found on worksheets with catchy labels of "csem1", "csem2", and so on. They are automatically computed whenever a cognitive subtest is found to have used number-right scoring, which is the most common type of scoring.

There are always three lines in these graphs. The flat one, the horizontal one, corresponds to the usual SEM value, as seen in Lertap's Stats1f reports. The SEM value is a constant; it's the same for all test takers, no matter their ability.

The top line, a curve, corresponds to standard errors computed using the binomial error model, labeled as Method III in **Lord** (1984). The lower curve is what results when Method III estimates are adjusted using Method IV; Method IV estimates will be lower than Method III's whenever the items used in the subtest have different difficulties ("diff" values, to use the parlance of Lertap's Statsb reports).

So: what's the graph tell us?

Consider those applicants with a true number-right score of 7 (just under 50% of the maximum possible test score of 15). The usual SEM at this score level was 1.63, as it was for all other levels. But the conditional standard errors of measurement at this point were greater: 2.00 using Method IV, and 1.75 using Method III; these CSEMs are a better reflection of the true state of affairs in the center of the score range -- errors of measurement are highest here, falling away as we move left or right to the extremes.

CSEMs will have special relevance when cut scores are used, for example, as in setting pass-fail points. For the test pictured above, we'd want to acknowledge greater measurement imprecision for scores near the middle, while at the same time having reason to believe that we have less error when it comes to testing the best students, and the weakest. (To be noted is the relationship between reliability and SEM. As reliability increases, SEM decreases. It would be interesting, would it not?, to see what would happen to these SEMs were we to correct I11's miskey; this is left as an exercise for you.)

Page forward.

Tidbits:

There's a paper which has lots more about measurement error, and the use of cut scores, with several compelling examples. Read all about it.

CSEM curves have an inverted U shape, opposite to the curve found in IRT, item response theory. In IRT, measurement precision is generally greater around the average score, and much less at the extremes, giving rise to a curve which dips in the middle, and goes up at either end (a U shape). Take in another paper by clicking here.

2.5.4 Download

The MathsQuiz.xlsx workbook may be downloaded by clicking here.

This workbook has an error in the CCs worksheet, in the *key line.

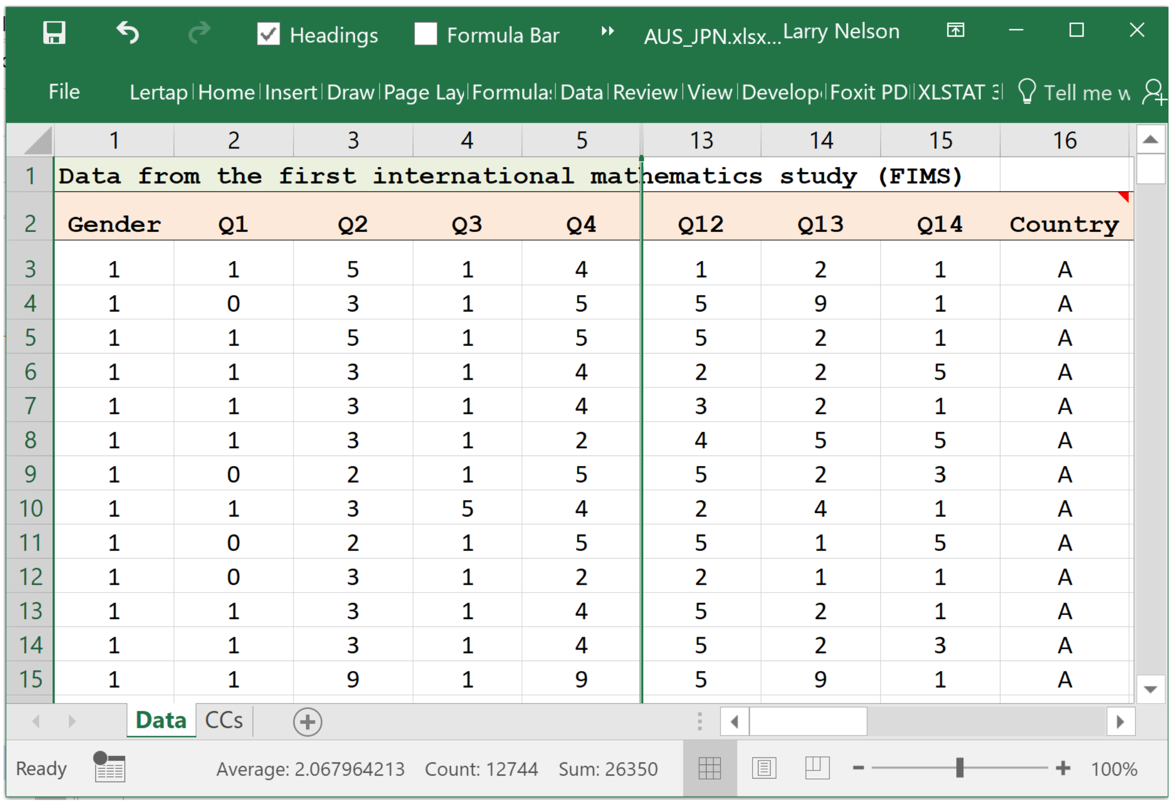
The key for item 11 should be 4, not 1.

2.6 FIMS

This dataset involves a 14-item mathematics test given as part of the First International Mathematics Study, "FIMS".

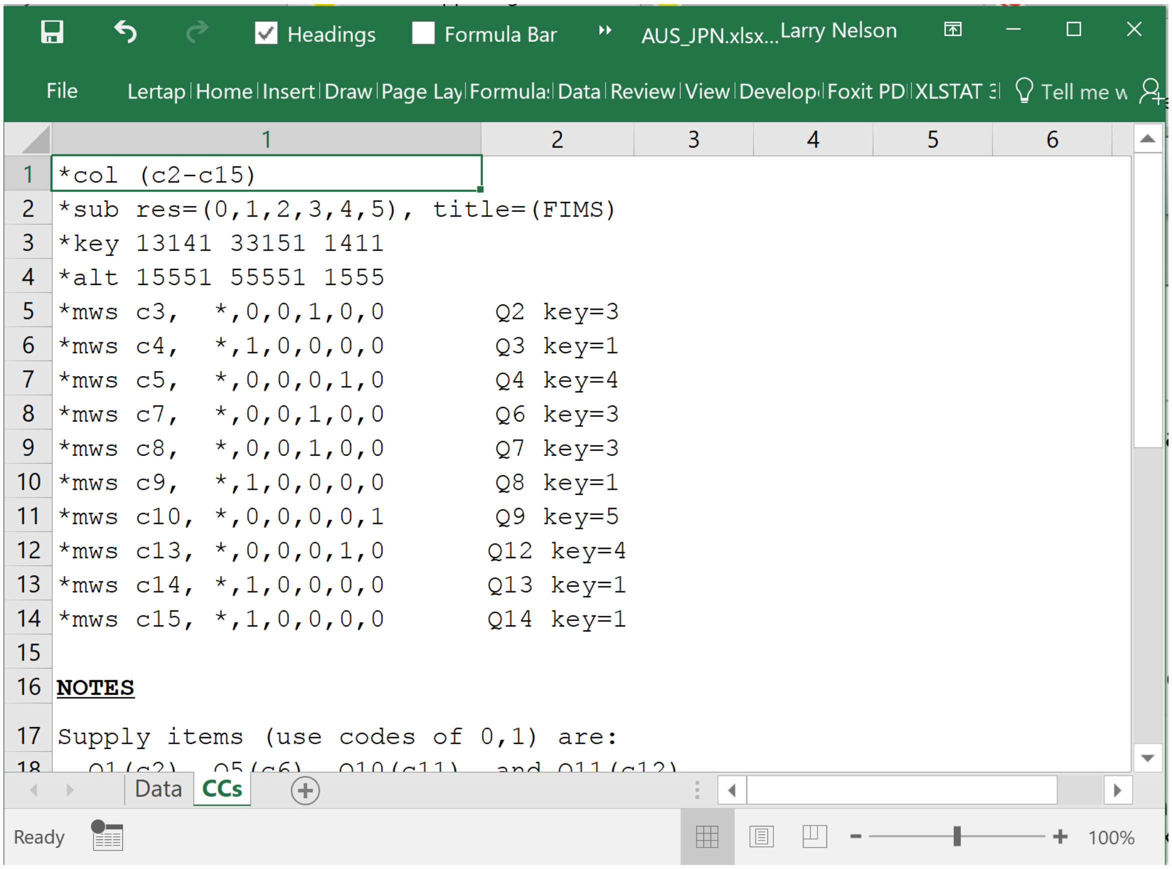
The items included in the dataset are a subset of the questions found on the original FIMS test. They are available as a docx file here. This study's items are similar to those found in an international numeracy study in that they use a mixture of item types, supply and multiple choice.

Responses from over 6,000 students are provided -- about two-thirds of them were from Australia with the rest coming from students in Japan. The top rows of the Data sheet are shown below; some of the columns have been hidden so that it is more obvious that two nominal variables were involved: Gender and Country.



	1	2	3	4	5	13	14	15	16
1	Data from the first international mathematics study (FIMS)								
2	Gender	Q1	Q2	Q3	Q4	Q12	Q13	Q14	Country
3	1	1	5	1	4	1	2	1	A
4	1	0	3	1	5	5	9	1	A
5	1	1	5	1	5	5	2	1	A
6	1	1	3	1	4	2	2	5	A
7	1	1	3	1	4	3	2	1	A
8	1	1	3	1	2	4	5	5	A
9	1	0	2	1	5	5	2	3	A
10	1	1	3	5	4	2	4	1	A
11	1	0	2	1	5	5	1	5	A
12	1	0	3	1	2	2	1	1	A
13	1	1	3	1	4	5	2	1	A
14	1	1	3	1	4	5	2	3	A
15	1	1	9	1	9	5	9	1	A

The top lines in the corresponding CCs sheet are pictured below:



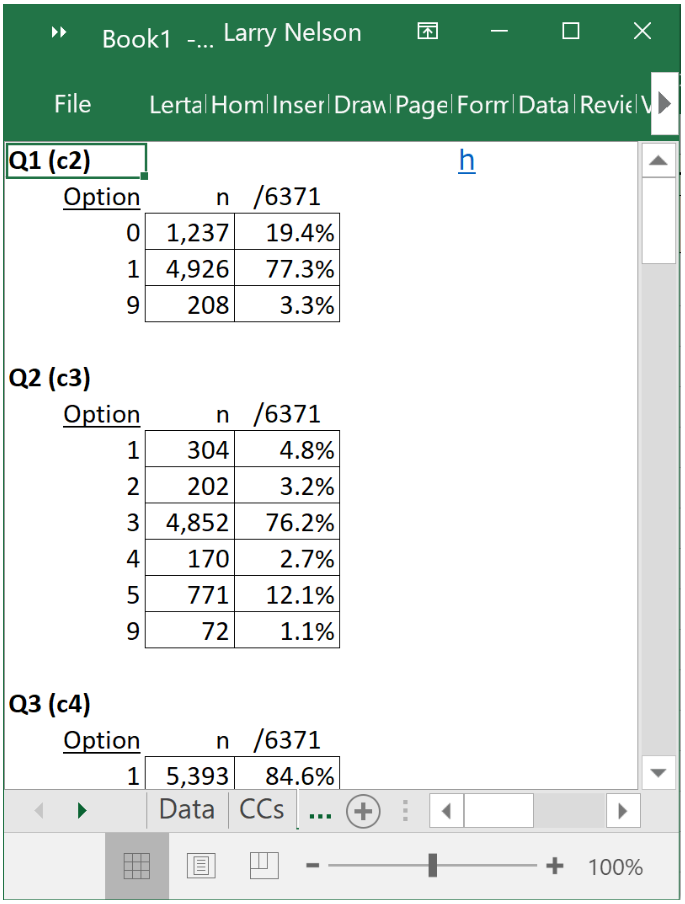
The lines in this CCs sheet are very similar to those used in the numeracy study mentioned above.

The "Interpret" and "Elmillon" options are used to get Lertap 5 to create its numerous reports.

Page ahead to see some of the output created by Lertap 5 and Excel.

2.6.1 Reports

A sample of the output seen in Lertap's Freqs report is pictured here:



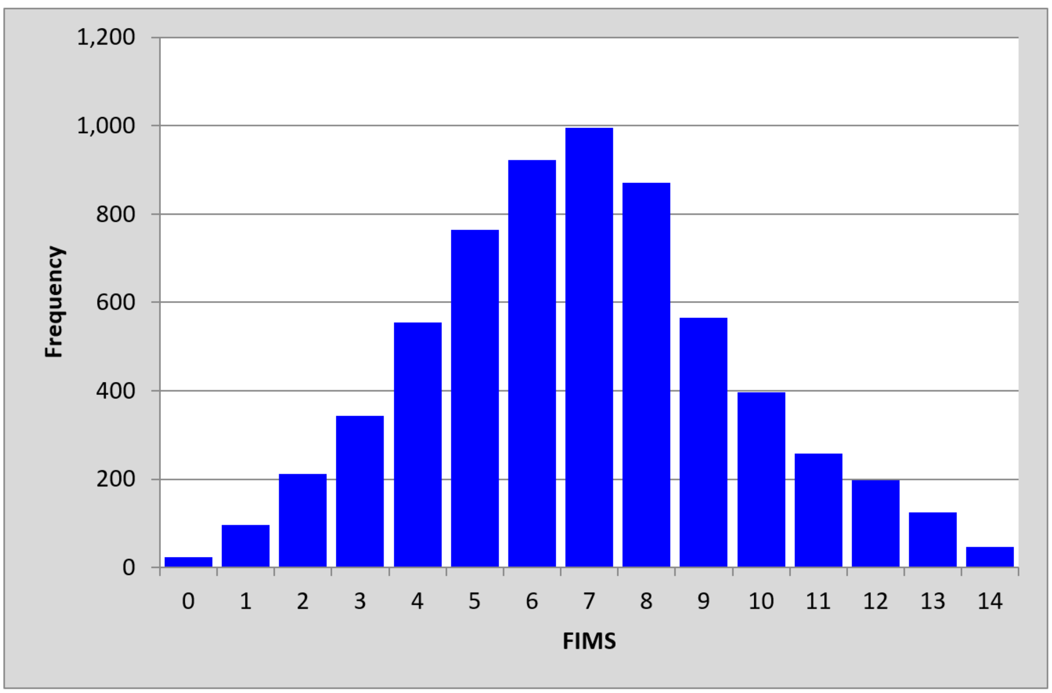
The first question, Q1, was a supply item (also known as a "constructed response" item). A response code of 0 (zero) indicates an incorrect answer, 1 was the code used for a correct answer, 9 was used to code non responses: 208 of the students failed to answer Q1.

The second question, Q2, was multiple-choice, with 72 students failing to answer it. It is curious that almost three times as many students failed to answer Q1 when compared to Q2 unless we consider guessing as a factor (Q1 was a much easier item but not multiple choice).

The Stats1b snippet below displays results for the initial five items; Q5 was very difficult for the students, with 9% (over 500 students) of them deciding to leave the question unanswered.

Options->	0	1	2	3	4	5	other	Difficulty	Discrimination
Q1	19%	77%					3%	0.77	0.26
Q2		5%	3%	76%	3%	12%	1%	0.76	0.40
Q3		85%	3%	3%	3%	5%	1%	0.85	0.29
Q4		2%	18%	2%	57%	20%	1%	0.57	0.41
Q5	75%	16%					9%	0.16	0.42
Q6		4%	5%	80%	3%	5%	3%	0.80	0.35

A histogram of the scores is shown below:

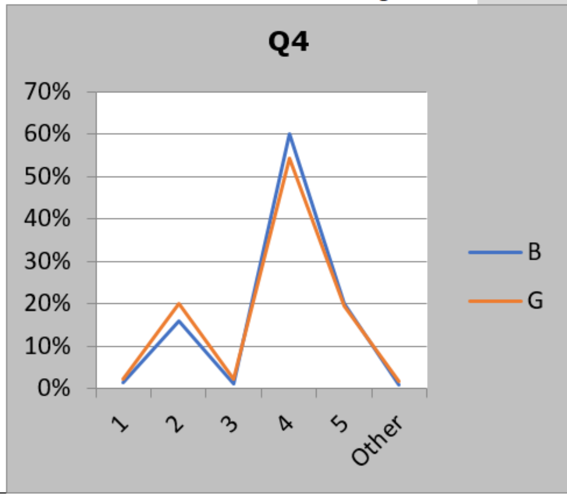


A variety of interesting research questions might be begging for answers.

Were there gender differences? This might be answered at two levels: the overall score level, and the item level. The plot below compares boy and girl responses on Q4. The boys did better on this item with 60% getting it correct compared to 54% of the girls -- these percentage figures are underlined to indicate that they correspond to the keyed-correct answer (which was option C in the original test).

Note that all students from both countries are included in the plot below. The next page provides the opportunity to look for gender differences within each country.

Q4	1	2	3	<u>4</u>	5	Other		n	mean	s.d.
B	1%	16%	1%	**<u>60%</u>**	20%	1%		3052	0.60	0.49
G	2%	20%	2%	**<u>54%</u>**	20%	2%		3319	0.54	0.50
						F, sig, eta^2:		*22.64*	*0.00*	*0.00*



Note: there's an anomaly in Excel and Lertap 5 which requires that nominal variables use non-numeric codes. This made it necessary to recode the data in column 1 of the Data sheet, mapping an original code of 1 to a new code of G, while the original code of 2 was mapped to B. This is easy to do using a special Lertap 5 option. The recoded gender column is called "GenderA".

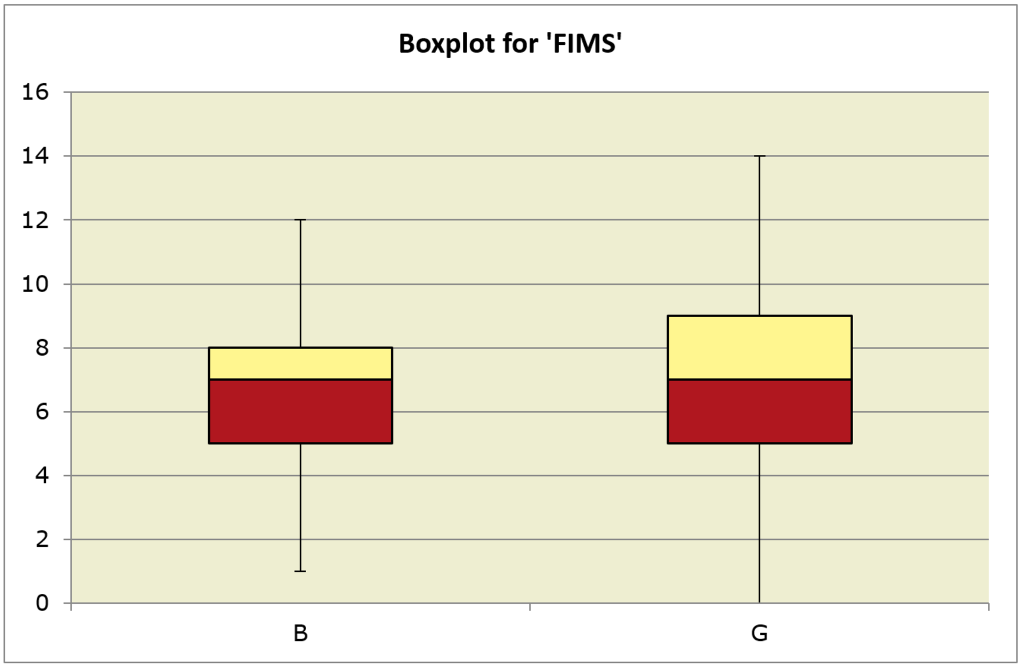
Quite a number of students thought that the second option, B (coded as 2), was the best answer on Q4. It would be best to check the Stats1f report to get the complete lowdown on this item:

Q4 (c5)

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	114	0.02	-0.12	-0.38	4.34	-0.92
2	0.00	1,152	0.18	-0.34	-0.49	4.89	-0.72
3	0.00	116	0.02	-0.14	-0.42	4.06	-1.03
<u>4</u>	<u>1.00</u>	<u>3,639</u>	<u>0.57</u>	<u>0.41</u>	<u>0.52</u>	<u>8.15</u>	<u>0.48</u>
5	0.00	1,266	0.20	-0.23	-0.33	5.60	-0.46
other	0.00	84	0.01	-0.16	-0.54	3.12	-1.38

The Stats1f summary for Q4 indicates that the second option, B, was a functioning distractor. The 1,152 students who chose this option were weak students: their overall test score average (avg.) was 4.89, much less than the average of 8.15 rendered by those who selected the fourth option, D. We could (and would) say that Q4 was a "discriminating item", one which performed quite well in this administration of the test.

The box plot below indicates that the boys and girls had the same median test score (7), with the girls' scores showing a greater range, and a more symmetric distribution.



What about country-level differences? Over years of international testing, Japanese students have often had superior mathematics test scores. Was that the case here with the FIMS data? Try to find out by downloading the dataset using the next page. (The country variable is coded in two ways in the Data sheet; use column 16 to answer this question.)

2.6.2 Download

The items used in this study may be seen in this docx file.

The Excel workbook for all students, Australian and Japanese, may be downloaded from this link.

A workbook with data for only the Australian students is available here, while a workbook for only the students from Japan may be found here.

Possible research questions are posed on the previous page.

Note1: the workbooks have an extra column with "GenderA", making in unnecessary to recode the Gender column (the Gender column uses codes of 1 and 2; the GenderA column uses codes of G and B and is the column to use when looking for gender differences).

Note2: these files have been adapted from those found at the TAM tutorials website. (Thanks to Professor Margaret Wu for assistance.)

2.7 Geology49

This dataset contains authentic test results from a multiple-choice test given to several thousand high school students in the United States.

It may be used to demonstrate and investigate how to use Lertap 5 to look for group differences; in this case there were two groups, girls and boys. Test developers were interested in seeing if, among other things, the items were gender fair, that is, of equal challenge (and difficulty) to both genders. This involved a look at "**DIF**", differential item functioning.

Read all about it here.

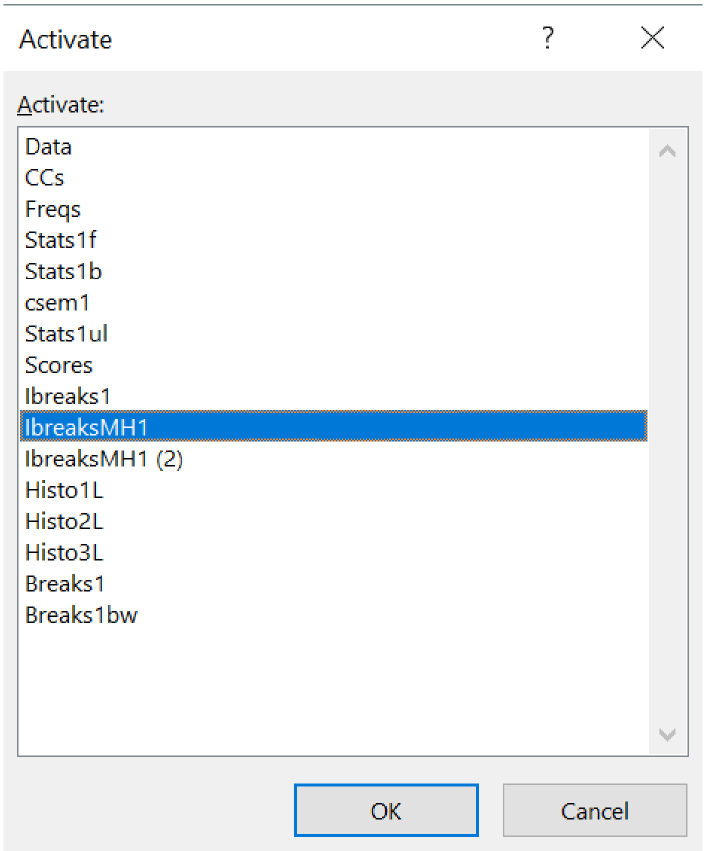
The actual data may be downloaded using links on the next page.

2.7.1 Downloads

The original dataset may be downloaded from this link as an Excel workbook, ready to use with Lertap 5.

Actual results found by using many of Lertap's options may be downloaded and admired using this link. Many of these results are discussed in this paper.

Note: Excel workbooks will often have many worksheets when Lertap options are used. It is possible to see a list of all of them as seen in the screen snapshot below; in this case, the IbreaksMH1 worksheet has been selected and will become the "active" worksheet:



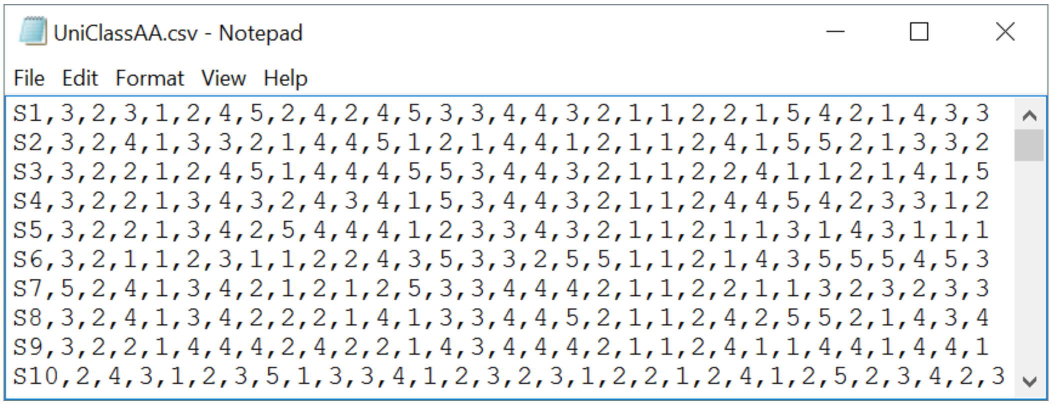
This webpage mentions how to get the list to appear. It's easy.

2.8 Uni Class A

This example is from a large North American university. It's based on a 30-item multiple-choice test taken by 127 students and will be used to exemplify the application of Lertap's response similarity analysis capability, RSA, to see if there might be some suggested evidence of possible cheating.

The original data are found in one of the sample files distributed with Wesolowsky's **SCheck** program; SCheck is short for "Similarity Checker".

A link to a "csv" (comma-separated values) file with original item responses is here; a snapshot of some of the lines in this file is shown below:



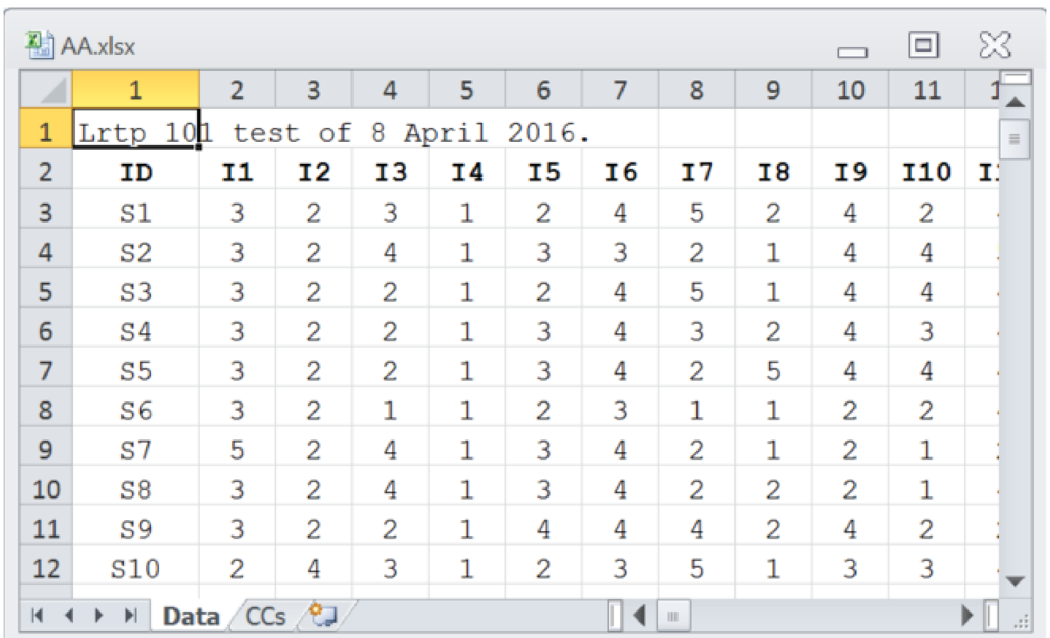
UniClassAA.csv - Notepad

File Edit Format View Help

```
S1,3,2,3,1,2,4,5,2,4,2,4,5,3,3,4,4,3,2,1,1,2,2,1,5,4,2,1,4,3,3
S2,3,2,4,1,3,3,2,1,4,4,5,1,2,1,4,4,1,2,1,1,2,4,1,5,5,2,1,3,3,2
S3,3,2,2,1,2,4,5,1,4,4,4,5,5,3,4,4,3,2,1,1,2,2,4,1,1,2,1,4,1,5
S4,3,2,2,1,3,4,3,2,4,3,4,1,5,3,4,4,3,2,1,1,2,4,4,5,4,2,3,3,1,2
S5,3,2,2,1,3,4,2,5,4,4,4,1,2,3,3,4,3,2,1,1,2,1,1,3,1,4,3,1,1,1
S6,3,2,1,1,2,3,1,1,2,2,4,3,5,3,3,2,5,5,1,1,2,1,4,3,5,5,5,4,5,3
S7,5,2,4,1,3,4,2,1,2,1,2,5,3,3,4,4,4,2,1,1,2,2,1,1,3,2,3,2,3,3
S8,3,2,4,1,3,4,2,2,2,1,4,1,3,3,4,4,5,2,1,1,2,4,2,5,5,2,1,4,3,4
S9,3,2,2,1,4,4,4,2,4,2,2,1,4,3,4,4,4,2,1,1,2,4,1,1,4,4,1,4,4,1
S10,2,4,3,1,2,3,5,1,3,3,4,1,2,3,2,3,1,2,2,1,2,4,1,2,5,2,3,4,2,3
```

OMR (optical mark recognition) scanners for processing student answers on "bubble sheets" commonly output csv files. These files are Excel friendly; Excel will readily open a csv file, no questions asked.

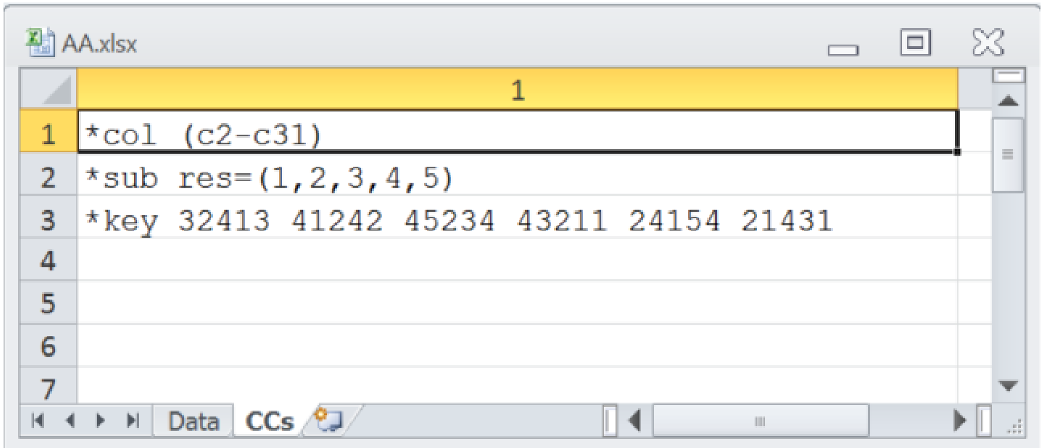
Importing csv files so that they're ready for Lertap is a process described in this topic. The following two snapshots display the Data and CCs worksheets set up for use in the format expected when working with Lertap 5.



AA.xlsx

	1	2	3	4	5	6	7	8	9	10	11
1	Lrtp 101 test of 8 April 2016.										
2	**ID**	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**
3	S1	3	2	3	1	2	4	5	2	4	2
4	S2	3	2	4	1	3	3	2	1	4	4
5	S3	3	2	2	1	2	4	5	1	4	4
6	S4	3	2	2	1	3	4	3	2	4	3
7	S5	3	2	2	1	3	4	2	5	4	4
8	S6	3	2	1	1	2	3	1	1	2	2
9	S7	5	2	4	1	3	4	2	1	2	1
10	S8	3	2	4	1	3	4	2	2	2	1
11	S9	3	2	2	1	4	4	4	2	4	2
12	S10	2	4	3	1	2	3	5	1	3	3

Data CCs



The *col line indicates that item responses start in "c2" of the Data worksheet and end in "c31" -- short for column 2 and column 31. The *sub line's res= statement tells Lertap that valid item responses for this test's items were 1,2,3,4,5, and the *key line gives the correct answer for each of the 30 items; the correct answer for the first item was 3, for the second item it was 2, and so on. More about CCs lines is here.

Ready for results? Play along if you'd like: download a copy of the dataset and run it with Lertap 5 (see this topic for how to get Lertap 5).

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

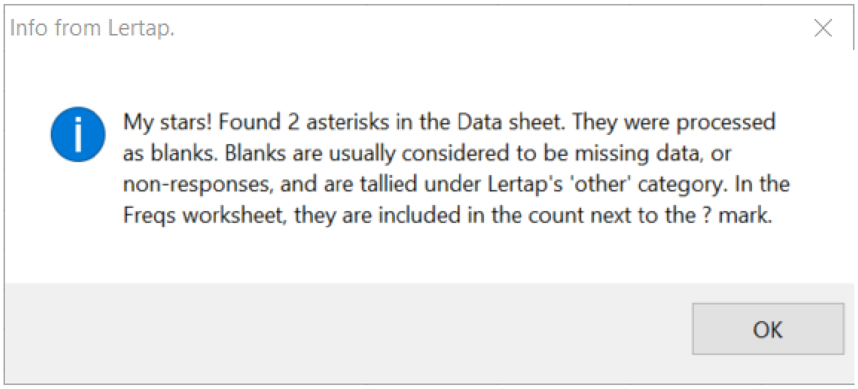
3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

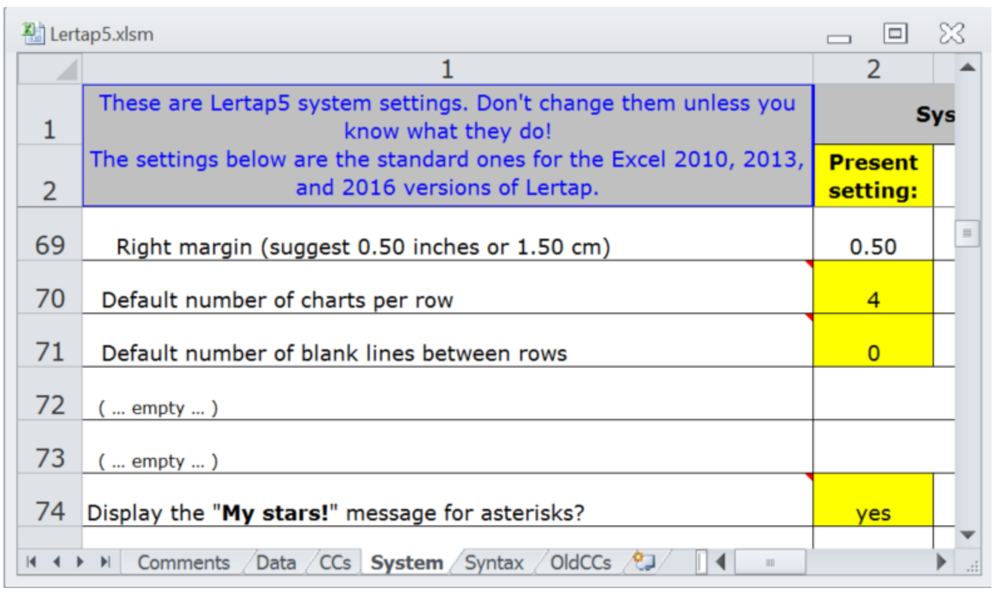
Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

As the Interpret option went about its work, it flashed up this message:



OMR scanners will sometimes record an asterisk as a student response when a student has shaded in more than one bubble on his/her bubble sheet. It's possible to ask Lertap to ignore asterisks by using a setting in line 74 of the System worksheet:



In the case of this little test, Interpret's Freqs report was pretty clean:

AA.xlsx

I1 (c2)

Option	n	/127
1	10	7.9%
2	7	5.5%
3	94	74.0%
4	8	6.3%
5	8	6.3%

I2 (c3)

Option	n	/127
Z	1	0.8%
1	12	9.4%
2	92	72.4%
3	7	5.5%
4	15	11.8%

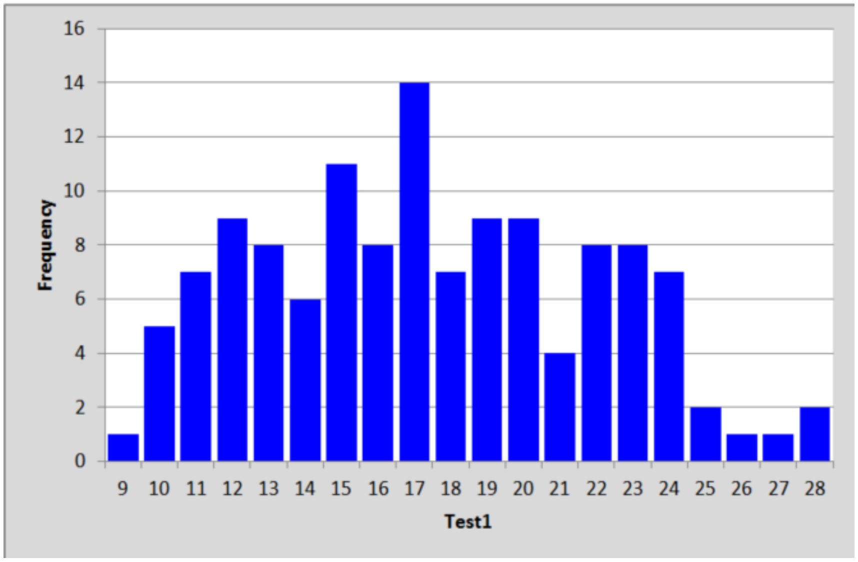
Data CCs Freqs

The number of Zs was very low, and there were very few items with a ? entry; the seventh item was one of just two items with a ? mark tally, indicating that one student apparently did not answer the item:

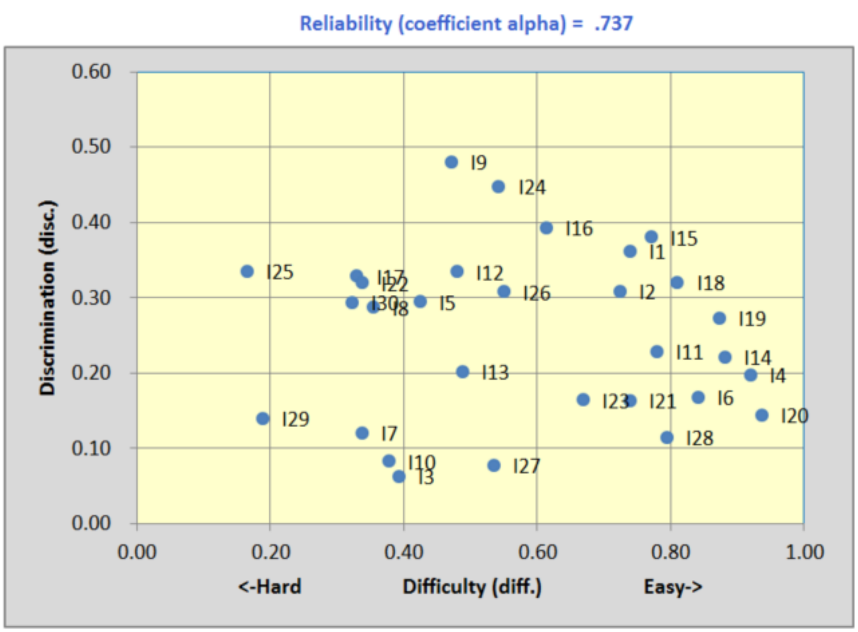
I7 (c8)

Option	n	/127
1	43	33.9%
2	34	26.8%
3	16	12.6%
4	13	10.2%
5	20	15.7%
?	1	0.8%

A histogram of the test scores produced by the Elmillon option showed no irregularities:



The scatterplot at the bottom of the [Stats1b] report revealed three items, I3, I10, and I27 with discrimination below 0.10:



A scan down the [Stats1f] report revealed a number of items with distractors (incorrect answers) selected by above-average students. Item 27 was one of them:

I27 (c28)

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	68	0.54	0.08	0.10	18.19	0.17
2	0.00	3	0.02	-0.12	-0.32	14.00	-0.76
3	0.00	31	0.24	-0.18	-0.25	15.97	-0.32
4	0.00	6	0.05	0.00	0.01	17.50	0.02 <-aa
5	0.00	18	0.14	0.04	0.07	17.89	0.11 <-aa
other	0.00	1	0.01	-0.15	-0.60	10.00	-1.65

The correct answer to this item was 1; 54% (n=68) of the students selected it. However, some fairly strong students, ones with above-average test scores, thought that distractors 4 and 5 were plausible answers. The 18 students who selected option 5, for example, were fairly capable students judged by their average test score of 17.89 (the mean of all test scores was 17.41, equal to 58%). There may have been some ambiguity in the item, something which might have caused some students to read the item in a manner different to that expected by the instructor.

As a result of items such as I27, coefficient alpha's value was on the low side -- we'd usually want to see a value of 0.80 or better, but multiple-choice tests made by lone instructors will often turn in relatively weak alpha values.

What to conclude thus far? Well, the data look clean, there aren't many unexpected "responses" -- a couple of Zs (for some reason), and an asterisk or two, nothing alarming. The score distribution seems reasonable, ranging from 9 to 28 with an average of just over 17, or 58% of the maximum possible score of 30.

Test reliability, as indexed by coefficient alpha, was on the low side. Some items, such as I27, may have had some ambiguity for the students, and the instructor might want to review them before the items are used again.

Let's move on now to running Lertap's RSA routine, a tool which can be used (*when conditions are suitable*) as a check on possible cheating, or "collusion".

Please page forward.

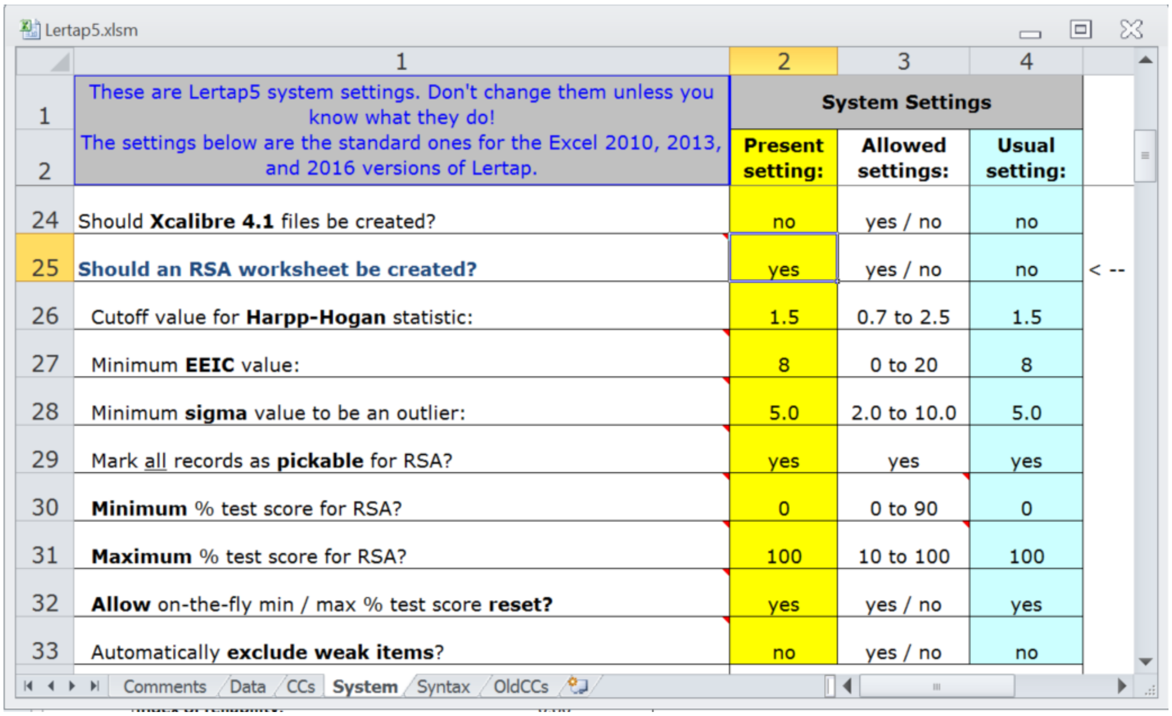
Tidbit:

Note the "*when conditions are suitable*" qualifier used in the text above. The RSA method used in Lertap works best when a test has at least 30 items and an average test score less than 80%. In personal correspondence (22 April 2016), Wesolowsky has stated that the method is "ok" when the number of questions is "around 50". He has found the method to underestimate similarity when there are "100 questions or so", and to overestimate when "there is a small number of questions".

2.8.1 Using Lertap RSA

An extensive discussion of "RSA", the response similarity analysis tool in Lertap 5, is available here.

Use of this tool requires that row 25 of Lertap's System worksheet be set to "yes":

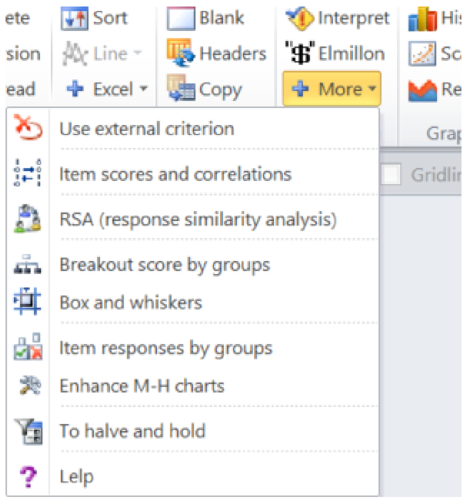

Lertap5.xlsm

	1	2	3	4	
1	These are Lertap5 system settings. Don't change them unless you know what they do!	System Settings			
2	The settings below are the standard ones for the Excel 2010, 2013, and 2016 versions of Lertap.	**Present setting:**	**Allowed settings:**	**Usual setting:**	
24	Should **Xcalibre 4.1** files be created?	no	yes / no	no	
25	**Should an RSA worksheet be created?**	yes	yes / no	no	< --
26	Cutoff value for **Harpp-Hogan** statistic:	1.5	0.7 to 2.5	1.5	
27	Minimum **EEIC** value:	8	0 to 20	8	
28	Minimum **sigma** value to be an outlier:	5.0	2.0 to 10.0	5.0	
29	Mark all records as **pickable** for RSA?	yes	yes	yes	
30	**Minimum** % test score for RSA?	0	0 to 90	0	
31	**Maximum** % test score for RSA?	100	10 to 100	100	
32	**Allow** on-the-fly min / max % test score **reset?**	yes	yes / no	yes	
33	Automatically **exclude weak items**?	no	yes / no	no	

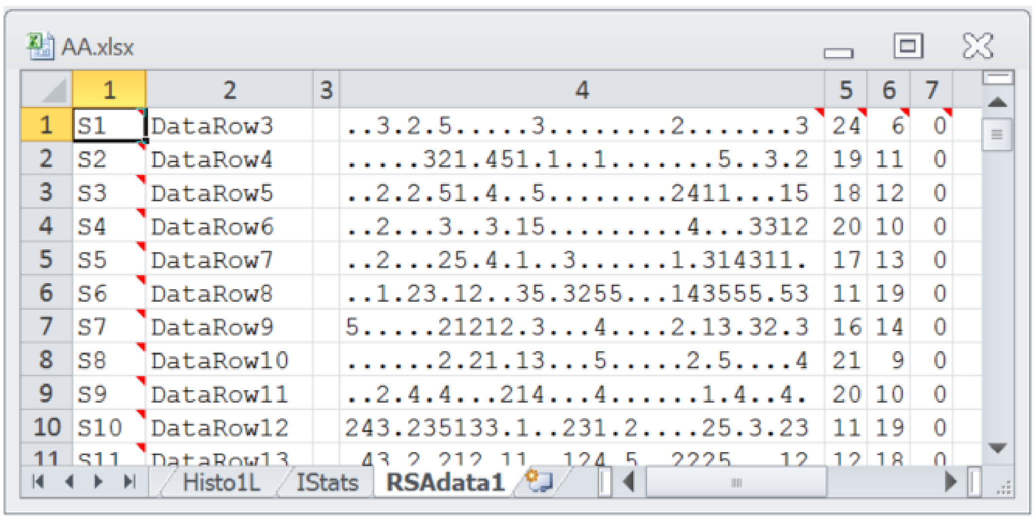
Comments / Data / CCs / System / Syntax / OldCCs

There are no less than eight rows in the System worksheet pertaining to the use of RSA, rows 26 to 33, as seen above. It is recommended that these default settings be used in almost all cases.

Let's get rolling ... the first step is to use the "Item scores and correlations" option from the Run menu:



Two new worksheets will be created, "IStats" and "RSAdata1". The snapshot below indicates the formatting of the RSAdata1 worksheet; as you look at this display, keep in mind that correct answers are indicated by "."; the other entries in each line indicate the distractors selected by students. Student S1 got the first two items right, selected option 3 (a distractor) on the third item, got the next item right, then selected option 2 on the fifth item, and so on. The 24 in column 5 is the number of items answered correctly by student S1. The 6 in the next column is this student's number of incorrect answers; the seventh column indicates the number of questions skipped (unanswered) by each student.



	1	2	3	4	5	6	7
1	S1	DataRow3		..3.2.5.....3........2.......3	24	6	0
2	S2	DataRow4		.....321.451.1..1.......5..3.2	19	11	0
3	S3	DataRow5		..2.2.51.4..5........2411...15	18	12	0
4	S4	DataRow6		..2...3..3.15.........4...3312	20	10	0
5	S5	DataRow7		..2...25.4.1..3.......1.314311.	17	13	0
6	S6	DataRow8		..1.23.12..35.3255...143555.53	11	19	0
7	S7	DataRow9		5.....21212.3...4....2.13.32.3	16	14	0
8	S8	DataRow10		......2.21.13...5.....2.5....4	21	9	0
9	S9	DataRow11		..2.4.4...214...4.......1.4..4.	20	10	0
10	S10	DataRow12		243.235133.1..231.2....25.3.23	11	19	0

At this point, some experienced users might get a fresh cup of coffee, turn off their phones, clean their glasses, and slowly scroll down the RSAdata1 worksheet. There are times when a bit of "hanky-panky" can be spotted by doing just that -- suddenly there will at times be blocks in column 4 which stand out, with response patterns almost suspicious in and of themselves.

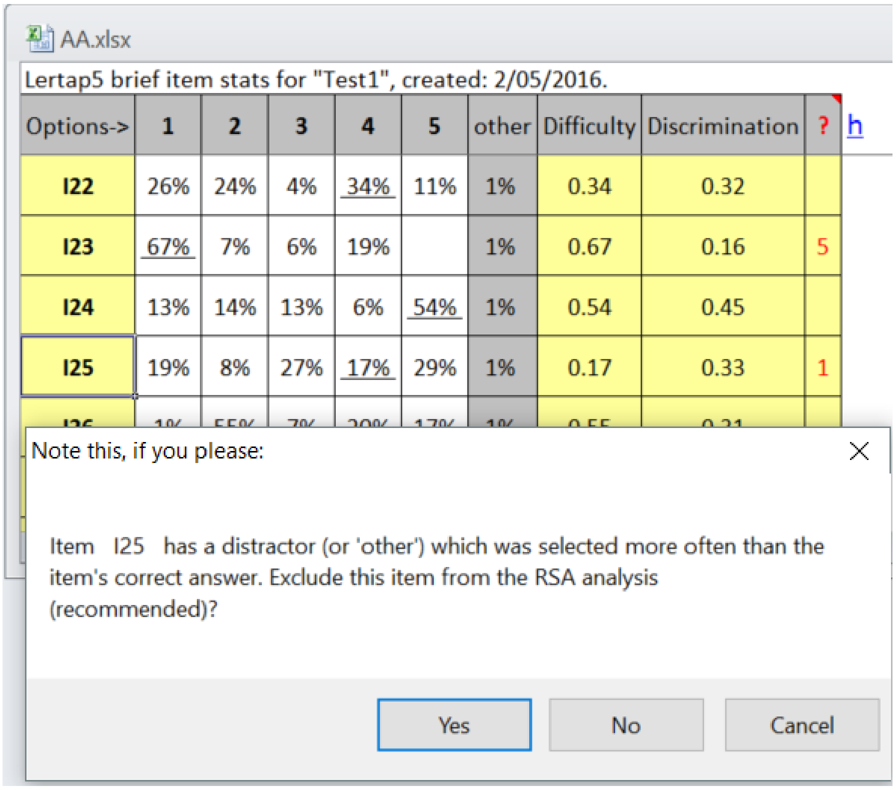
Look, for example, at the screen snapshot below, an excerpt from a 44-item test taken by some 2,500 students:

2336	S2336	............b...b.......d......a........-.-b
2337	S2337	........a.b.b.b.-c...........b.a......b.....
2338	S2338	..b.d.a...--..bc-b..dddbcb..bc.c.----abbb...
2339	S2339	d.bd..da...-bbbdbb..ddd.c...bd.c.b..baa.-...
2340	S2340	...............d....b.......ab..b..b........
2341	S2341	..b............d....b.......ab..b..b........
2342	S2342	..b............d....b.......ab..b..b........
2343	S2343	...............d....b.......ab..b..b........
2344	S2344	..b............d...bb.......ab..b..b........
2345	S2345	..b............d....b.......ab..b..b........
2346	S2346	..b............d....b.......ab..b..b........
2347	S2347	..b............d...bb.......ab..b..b........
2348	S2348	..b............d....b.......ab..b..b........
2349	S2349	...............d...bb.......ab..b..b........
2350	S2350	...............d....b.......ab..b..b........
2351	S2351	..b............d...bb.......ab..b..b........
2352	S2352	..b...cac......dbbbbdbb..b..ad.abb..b.......
2353	S2353	d.b.d.a.a.....bd-...b.b.a..b.ca.b.......-...
2354	S2354	da...d.ac..b.bbd-abbb..bdbbbdcbc..bb.c..b...

There's a suspicious block above, starting with S2340. We might not require any sort of sophisticated analysis here -- we would be unable to categorically state that the students in this block colluded, but to some readers (probably many), hanky-panky would indeed seem to be present.

But let's come back to the results from "Uni Class A". We've taken the "Item scores and correlations" option and an RSAdata1 worksheet is on hand.

We can now take the "RSA (response similarity analysis)" option from the Run menu. It produced the following message as it got underway:



The recommended application of Lertap's RSA routine involves weeding out those items where one or more of the distractors is found to be more popular than the correct answer. This has happened here with I25: 17% of the students selected option 4, the correct answer, but more students opted for almost all of the other options. We take "Yes" to exclude this item from the analysis. Ditto for I29.

The RSA routine will ask for the range of scores to use in its analysis. This will usually be from the minimum possible test score to the maximum possible test score (but in some cases it might be desirable to work within a smaller range of scores).

It took our laptop computer with an i5 processor 29 seconds to compete the RSA analysis.

It found one pair of students to have item responses judged to be "very suspect". This was reported in the "RSAcases1" worksheet:

AA.xlsx

Lertap5 RSA cases list with EEIC min = 8, produced on: 2/05/2016.

ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
S58	DataRow60	1...43.524.....14...52.3x.3.x3	15	12	2	6.00	-22.28	7.69
S96	DataRow98	1...41.524.....143..52.3x.3.x3	14					

Total number of cases displayed above: 1.
Number in the pink (very suspect cases): 1.

RSAsig1 RSAtable1 RSAcases1

An "x" in the Responses string indicates an item excluded from the analysis.

Here we see that the two students, S58 and S96, had 12 exact errors in common (EEIC) and just two response differences (D) over all the items, resulting in an Index, the "Harpp-Hogan statistic", of 6.00 and Sigma of 7.69. These are extreme values for these statistics and the reason their results have been flagged as "very suspect".

The RSAsig1 report looked like this:

Lertap5 RSAsig probabilities list with

S1 ID	S1 Data row	S2 ID	S2 Data row
S55	DataRow57	S69	DataRow71
S61	DataRow63	S126	DataRow128
S72	DataRow74	S73	DataRow75

Pairings	
Suspect:	1
Not suspect:	8,000
Total:	8,001
Inclusions	
Number of items:	28
Number of students:	127
Run control	
EEIC minimum:	8
H-H index minimum:	1.5
H-H sigma minimum:	5
Items excluded:	2
Minimum score setting:	0
Maximum score setting:	30

RSAsig1 RSAtable1

Out of a possible 8,001 pairs of students, only one pair was found to have their item responses flagged as "suspect".

More about the interpretation of RSA output is here.

Results from Wesolowsky's SCheck program are found in the next topic.

2.8.2 SCheck Results Class A

As a crosscheck on Lertap RSA, the results below were found by running Weslowsky's SCheck program. In this case, the RSA findings from Lertap are corroborated by SCheck. Both programs found that students 58 and 96 had responses which seemed to be suspicious.

AA.out - Notepad

File Edit Format View Help

```
 Summary of Significances of Identified Pairs
 --------------------------------------------

  pair          Z     A Priori    Bonferroni     Approx. Prog.
                      Signif.     Signif.        Selected Signif.
-------------------------------------------------------------
 58, 96        6.188  3.0E-10     2.4E-6         3.0E-10*
-------------------------------------------------------------



         SUMMARY
        ==========
mean of Z's = -0.0082 stdev=  0.8100

The number of pairs checked is  8001
The specified significance for program-selected pairs is   0.001

The approx. cutoff Z based on the specified significance is  4.177
The cutoff Z based on the Bonferroni bound is 5.16
start = (04-12-2016  16:02:56)  finish = (04-12-2016  16:02:59)
Number of observations above 4.177   is 1


_________________________________________________________________
 Do not copy or distribute. For use by licensee ONLY.
 This program is owned by G.O. Wesolowsky (wesolows@mcmaster.ca)
 Licensed to Demo Nelson  , Expires 09/30/2016
_________________________________________________________________
  The methodology of this program is described in:
  George O. Wesolowsky, 'Detecting Excessive Similarity in Answers
  on Multiple Choice Exams', Journal of Applied Statistics,
  Vol. 27, No. 7, 2000, pp. 909-921.
```

2.8.3 Download Uni Class A

Use these links to download the dataset as an Excel (xlsx) Lertap 5 workbook, or as a csv file with all item response data from the Data worksheet.

The item keys, the correct answers for all 30 items, are:

324134124245234432112415421431

2.9 Uni Class B

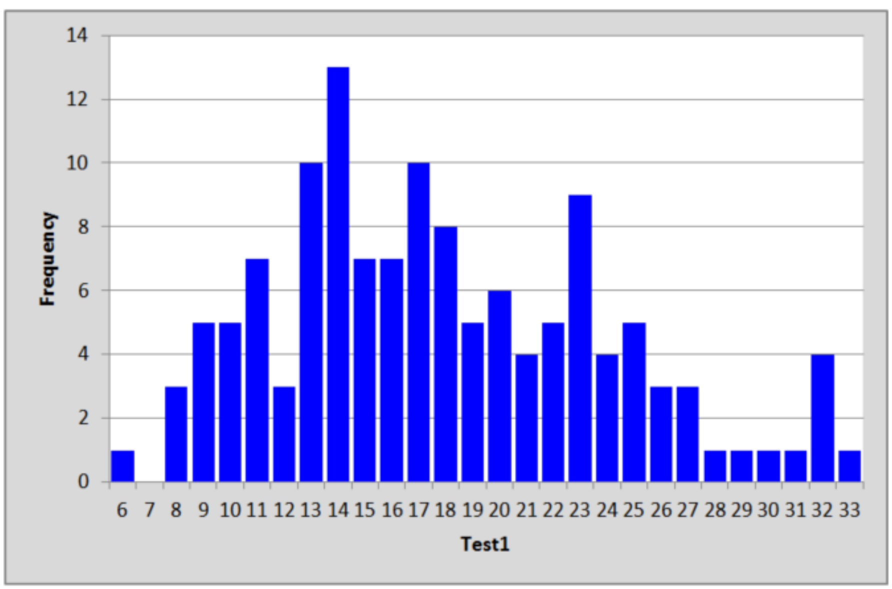
This dataset is similar to that from "Uni Class A" in that it is from a North American university and involves an instructor-created multiple-choice test. There were 34 test items responded to by 132 students.

We'll replicate the approach applied in the analysis of "Uni Class A", using Lertap's RSA routine to look for an indication of possible cheating. You will see that there were more suspects in this class, with results in some disagreement with those from the SCheck program.

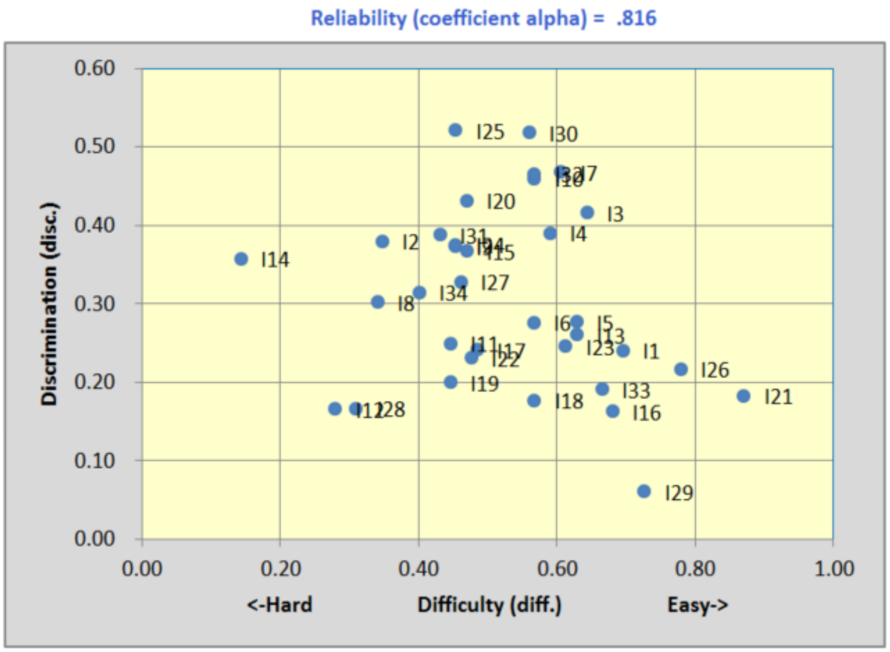
A copy of the dataset, as an Excel workbook, may be downloaded here.

The Freqs report was as clean as that in Uni Class A. In fact, it was somewhat cleaner, with no asterisks, and even fewer unanswered questions.

The distribution of test scores is shown below:



The distribution is fairly similar to that from Uni Class A. The average score was 52.5% (17.85 out of a maximum possible score of 34), standard deviation 6.1 (compared to 4.5 in Uni Class A results).



This test's coefficient alpha value was higher; there was only one item, I29, with discrimination below 0.10, and only one really easy item, I21.

Page forward for RSA results.

2.9.1 Using Lertap RSA

The results of Lertap's RSA run for this test are captured below. Four items were excluded from the analysis as their correct answers were selected by fewer students than found for one or more of the distractors (items I2, I12, I14, and I20).

It took 34 seconds to process the data.

Lertap5 RSA cases list with EEIC min = 8, produced on: 24/7/19.

ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
S36	DataRow38	.x..5315514x1x5.431x..344.52.255.5	10	15	7	2.14	-26.09	7.34
S72	DataRow74	.x1.5315514x1x5.431x...33.52.2...3	11					
S69	DataRow71	.x1.5315514x1x5.431x..344..2.2.43	11	15	7	2.14	-27.48	7.88
S72	DataRow74	.x1.5315514x1x5.431x...33.52.2...3	11					
S36	DataRow38	.x..5315514x1x5.431x..344.52.255.5	10	17	6	2.83	-30.12	8.92
S69	DataRow71	.x1.5315514x1x5.431x..344..2.2.43	11					
S69	DataRow71	.x1.5315514x1x5.431x..344..2.2.43	11	15	5	3.00	-28.48	8.28
S70	DataRow72	.x1.5315534x1x5.431x...33..2.2..3	13					
S70	DataRow72	.x1.5315534x1x5.431x...33..2.2..3	13	16	4	4.00	-30.11	8.92
S72	DataRow74	.x1.5315514x1x5.431x...33.52.2...3	11					
S78	DataRow80	.x.1233455.x.x2542.x...41..2..51.5	14	15	2	7.50	-29.20	8.56
S97	DataRow99	.x.14334.5.x.x2542.x...41..2..51.5	14					
S2	DataRow4	.x.12334.5.x.x2542.x...41..2..51.5	15	15	1	15.00	-29.89	8.83
S97	DataRow99	.x.14334.5.x.x2542.x...41..2..51.5	14					
S2	DataRow4	.x.12334.5.x.x2542.x...41..2..51.5	15	16	1	16.00	-31.15	9.32
S78	DataRow80	.x.1233455.x.x2542.x...41..2..51.5	14					

Total number of cases displayed above: 9.
Number in the pink (very suspect cases): 9.

Pairings	
Suspect:	9
Not suspect:	8,637
Total:	8,646
Inclusions	
Number of items:	30
Number of students:	132
Run control	
EEIC minimum:	8
H-H index minimum:	1.5
H-H sigma minimum:	5
Items excluded:	4
Minimum score setting:	0
Maximum score setting:	34

Nine student pairs were found by Lertap RSA to have "very suspect" responses.

The next page has results obtained from the SCheck program.

2.9.2 SCheck Results Class B

The results below were produced by Wesolowsky's SCheck program.

Two of the 11 pairs below, {36, 70} and {95, 128}, were not flagged by Lertap's RSA. The two programs agreed on the other 9 pairs. (But be sure to read the text below the table -- the two programs agree exactly if Lertap's options are changed.)

```
BB.out - Notepad
File  Edit  Format  View  Help

 Summary of Significances of Identified Pairs
 -------------------------------------------

  pair          Z     A Priori    Bonferroni     Approx. Prog.
                      Signif.     Signif.        Selected Signif.
-------------------------------------------------------------
 2, 78        7.891  1.5E-15     1.3E-11        1.5E-15*
 2, 97        7.428  5.5E-14     4.8E-10        5.5E-14*
 36, 69       6.253  2.0E-10     1.7E-6         2.9E-10
 36, 70       4.755  9.9E-7      8.5E-3         5.3E-5
 36, 72       5.514  1.8E-8      1.5E-4         1.7E-7
 60, 119      4.931  4.1E-7      3.5E-3         1.5E-5
 69, 70       6.527  3.3E-11     2.9E-7         3.3E-11*
 69, 72       5.474  2.2E-8      1.9E-4         2.4E-7
 70, 72       6.527  3.3E-11     2.9E-7         3.3E-11*
 78, 97       7.067  7.9E-13     6.9E-9         7.9E-13*
 95, 128      4.560  2.6E-6      2.2E-2         2.0E-4
-------------------------------------------------------------


          SUMMARY
         ==========
mean of Z's = -0.0381 stdev=  0.8347

The number of pairs checked is  8646
The specified significance for program-selected pairs is   0.001

The approx. cutoff Z based on the specified significance is  4.317
The cutoff Z based on the Bonferroni bound is 5.17
start = (04-12-2016  15:45:02)  finish = (04-12-2016  15:45:04)
Number of observations above 4.317   is 11
```

When Lertap's RSA analysis was initially run, producing the results seen on the previous page, four items were omitted due to "overly-popular distractors", that is, distractors (incorrect answers) that were selected more frequently than the items' correct answers. Excluding such items is recommended when Lertap RSA runs. However, if the correct answer is selected almost as often as the "popular distractor", and if the item's discrimination is at least 0.20, then leaving the item in the analysis may well be appropriate.

In this case, if all four initially excluded items, I2, I12, I14, and I20, are included, then a Lertap RSA analysis will include 10 of the pairs identified by SCheck, missing out only the {95, 128} pair. Here is the RSAcases report after the four items have been included:

Lertap5 RSA cases list with EEIC min = 8, produced on: 24/7/19.

ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
S36	DataRow38	....53155143135.4312..344.52.255.5	10	16	10	1.60	-26.03	6.37
S70	DataRow72	..1.53155343135.4312...33..2.2..3.	13					
S60	DataRow62	.4.....2.....444441..33....5......	23	11	5	2.20	-27.21	6.79
S119	DataRow121	44.....2.43.5444441..33....5..3...	18					
S36	DataRow38	....53155143135.4312..344.52.255.5	10	18	8	2.25	-28.94	7.42
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S69	DataRow71	..1.53155143135.4312..344..2.2.43.	11	18	8	2.25	-30.33	7.92
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S36	DataRow38	....53155143135.4312..344.52.255.5	10	20	6	3.33	-33.89	9.21
S69	DataRow71	..1.53155143135.4312..344..2.2.43.	11					
S69	DataRow71	..1.53155143135.4312..344..2.2.43.	11	18	5	3.60	-32.24	8.62
S70	DataRow72	..1.53155343135.4312...33..2.2..3.	13					
S70	DataRow72	..1.53155343135.4312...33..2.2..3.	13	19	5	3.80	-32.96	8.88
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S78	DataRow80	.4.1233455.5.12542.....41..2..51.5	14	18	3	6.00	-33.91	9.22
S97	DataRow99	.4.14334.5.5.12542.2...41..2..51.5	14					
S2	DataRow4	.4.12334.5.5.12542.....41..2..51.5	15	18	2	9.00	-34.59	9.47
S97	DataRow99	.4.14334.5.5.12542.2...41..2..51.5	14					
S2	DataRow4	.4.12334.5.5.12542.....41..2..51.5	15	19	1	19.00	-36.51	10.16
S78	DataRow80	.4.1233455.5.12542.....41..2..51.5	14					

Total number of cases displayed above: 10.
Number in the pink (very suspect cases): 10.

Now, a look at the bottom of the corresponding RSAsig report:

Lertap5 RSAsig probabilities list with EEIC min = 8, created on: 24/7/19.

S1 ID	S1 Data row	S2 ID	S2 Data row	S1 Correct	S2 Correct	EEIC	D	H-H index	Log(PROB)	H-H sigma
S46	DataRow48	S70	DataRow72	16	13	11	13	0.85	-19.18	-3.88
S46	DataRow48	S72	DataRow74	16	11	11	13	0.85	-18.65	-3.69
S35	DataRow37	S101	DataRow103	12	13	12	14	0.86	-21.66	-4.78
S60	DataRow62	S131	DataRow133	23	15	9	10	0.90	-21.95	-4.89
S49	DataRow51	S114	DataRow116	33	32	1	1	1.00	-20.55	-4.38
S96	DataRow98	S98	DataRow100	19	17	10	10	1.00	-19.33	-3.94
S119	DataRow121	S131	DataRow133	18	15	11	10	1.10	-24.24	-5.72
S60	DataRow62	S95	DataRow97	23	26	6	5	1.20	-20.34	-4.31
S60	DataRow62	S128	DataRow130	23	27	6	5	1.20	-20.34	-4.31
S95	DataRow97	S128	DataRow130	26	27	7	1	7.00	-22.67	-5.15
Pairings								n	8,636	8,636

The {S95, S128} pair of students has an H-H Index of 7.00, and a sigma greater than 5 (in absolute value, that is, ignoring the negative sign on -5.15).

If the EEIC cutoff is now changed to 6 (instead of 8), then the {S95, S128} pair will also be added to Lertap's RSAcases list, resulting in total agreement with SCheck:

Lertap5 RSA cases list with EEIC min = 6, produced on: 24/7/19.

ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
S36	DataRow38	....53155143135.4312..344.52.255.5	10	16	10	1.60	-26.03	6.38
S70	DataRow72	..1.53155343135.4312...33..2.2..3	13					
S60	DataRow62	.4.....2.....444441..33....5.....	23	11	5	2.20	-27.21	6.80
S119	DataRow121	44.....2.43.5444441..33....5..3..	18					
S36	DataRow38	....53155143135.4312..344.52.255.5	10	18	8	2.25	-28.94	7.43
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S69	DataRow71	..1.53155143135.4312..344..2.2.43	11	18	8	2.25	-30.33	7.93
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S36	DataRow38	....53155143135.4312..344.52.255.5	10	20	6	3.33	-33.89	9.22
S69	DataRow71	..1.53155143135.4312..344..2.2.43	11					
S69	DataRow71	..1.53155143135.4312..344..2.2.43	11	18	5	3.60	-32.24	8.63
S70	DataRow72	..1.53155343135.4312...33..2.2..3	13					
S70	DataRow72	..1.53155343135.4312...33..2.2..3	13	19	5	3.80	-32.96	8.89
S72	DataRow74	.11.53155143135.4312...33.52.2...3	11					
S78	DataRow80	.4.1233455.5.12542.....41..2..51.5	14	18	3	6.00	-33.91	9.23
S97	DataRow99	.4.14334.5.5.12542.2...41..2..51.5	14					
S95	DataRow97	.1.....2.....5..4.1..33....5.....	26	7	1	7.00	-22.67	5.16
S128	DataRow130	.......2.....5..4.1..33....5.....	27					
S2	DataRow4	.4.12334.5.5.12542.....41..2..51.5	15	18	2	9.00	-34.59	9.48
S97	DataRow99	.4.14334.5.5.12542.2...41..2..51.5	14					
S2	DataRow4	.4.12334.5.5.12542.....41..2..51.5	15	19	1	19.00	-36.51	10.18
S78	DataRow80	.4.1233455.5.12542.....41..2..51.5	14					

Total number of cases displayed above: 11.
Number in the pink (very suspect cases): 11.

Harpp & Hogan (1993) suggested an EEIC=6 setting but modified it to EEIC=8 in Harpp, Hogan, & Jennings (1996). (Reference list is here).

The EEIC=8 setting can be viewed as conservative in that it makes it more difficult for a pair's results to be flagged as being possibly indicative of collusion.

This example has shown that looking at RSAsig results might suggest what could happen if the EEIC cutoff setting is changed: initially we had 10 student pairs in the RSAcases report when EEIC=8, but a look at the bottom of the corresponding RSAsig report signaled that we might want to set EEIC=6. When we did, the number of student pairs in RSAcases went up to 11.

It would bear repeating, perhaps, to again say that *none of these results represent conclusive evidence of cheating*.

Tidbit: the Windows version of Lertap 5 allows for some RSA settings to be changed on-the-fly, making it easy to see the effects of changing them.

2.9.3 Download Uni Class B

Use these links to download the dataset as an Excel (xlsx) Lertap 5 workbook, or as a csv file with all item response data from the Data worksheet.

The item keys, the correct answers for all 34 items, are:

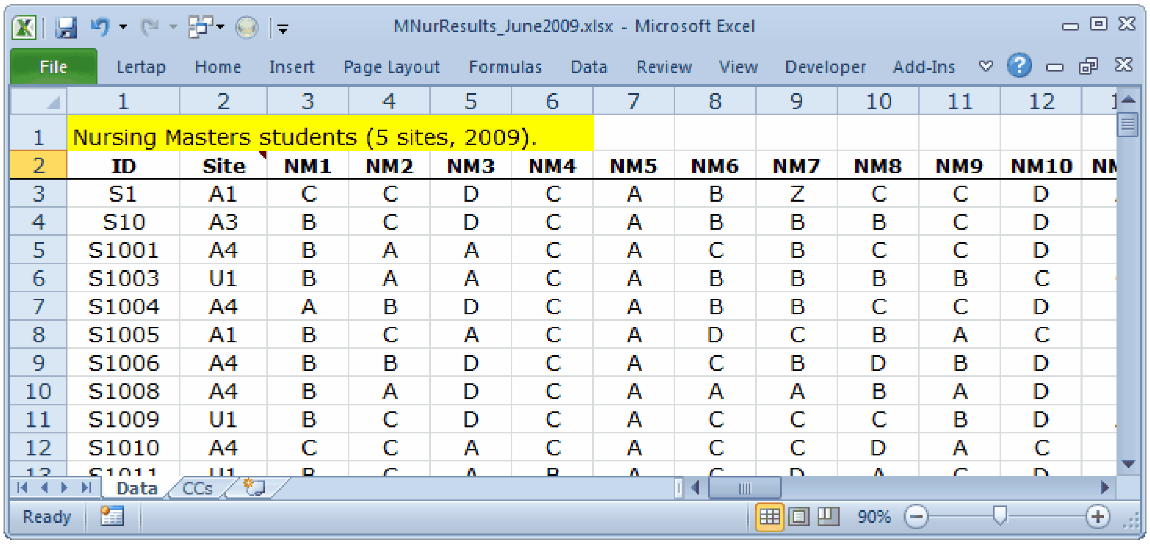
5233312332212233514115252244154344

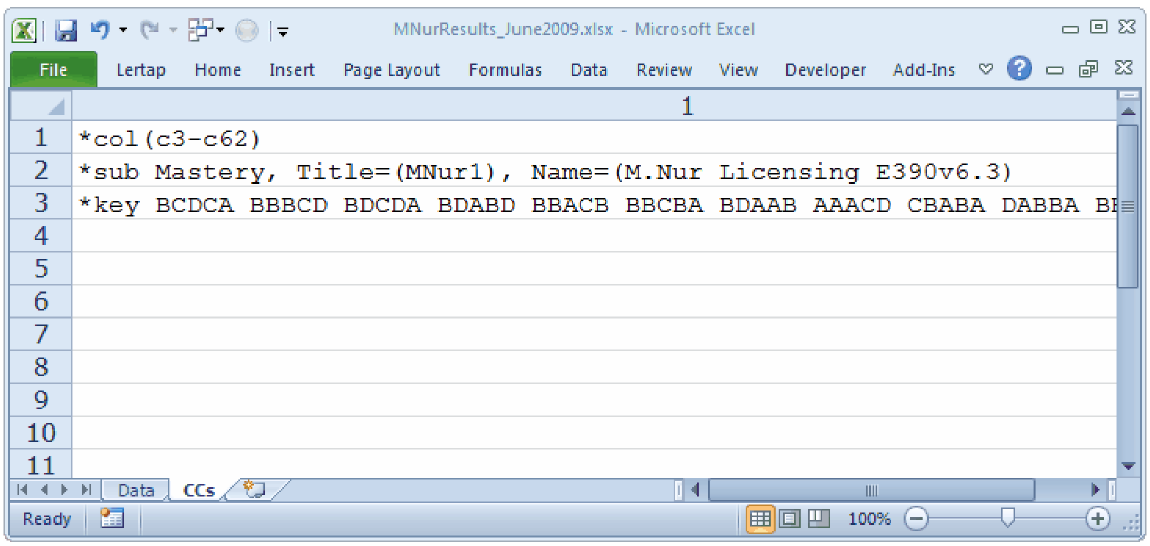
2.10 M.Nursing

A professional nursing organization contracted for the development of a master's degree course in nursing, a course subsequently licensed to universities and medical colleges for delivery. In 2009, the course was available in more than two dozen countries. All instruction is given in English (no matter the country), and all examinations are in English.

The results presented in this topic are from the administration of one of the exams, a 60-item test covering aspects of health care management. Data are from five testing venues, three in an English-speaking country, one in Asia, and one in South America. In all cases students (n=1,769) marked their answers on mark-sense scoring sheets. These were locally scanned to produce Excel-compatible files with the students' responses.

The exam has a cutoff point, or "**mastery score**", of 70%. Any student scoring less than 70% is placed in a conditional category, and can be required to repeat the course content covered by the exam (assignment work and class participation help to decide on whether or not students have to repeat).





Testing venue is coded as "Site" in column 2 of the Data worksheet. Sites with a code beginning with the letter A are from the English-speaking country; H is for the Asian test venue; U for South America. The scanner recorded a Z as an item "response" whenever the student did not answer an item, or shaded in more than one bubble on the mark-sense sheet.

Note the presence of "**Mastery**" on the *sub line. This word is required in order to produce the mastery test statistics supported by Lertap. The mastery percentage level is set to 70% in the System worksheet, a value which may be changed. It is also possible to specify another level on the *sub line itself. For example, Mastery=50% would set the level at 50%. (Mastery=50, on the other hand, would set the cutoff to a test score of 50. Read more about this here.)

Ready for results? (You can play along if you've downloaded this dataset.)

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

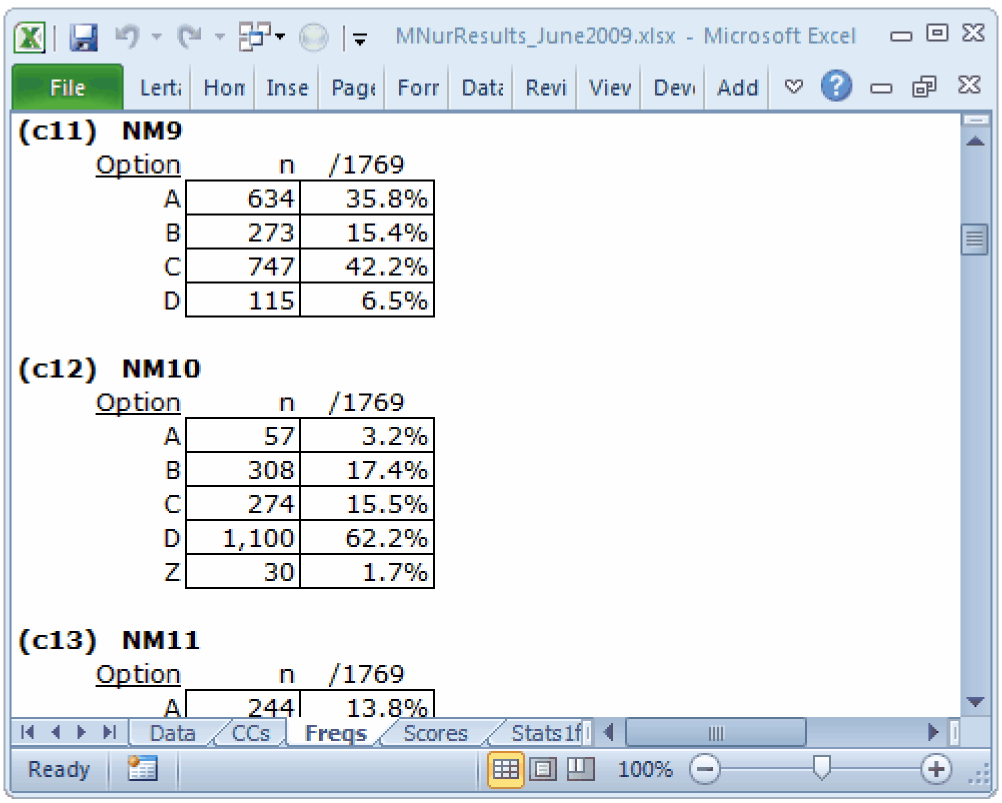
2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

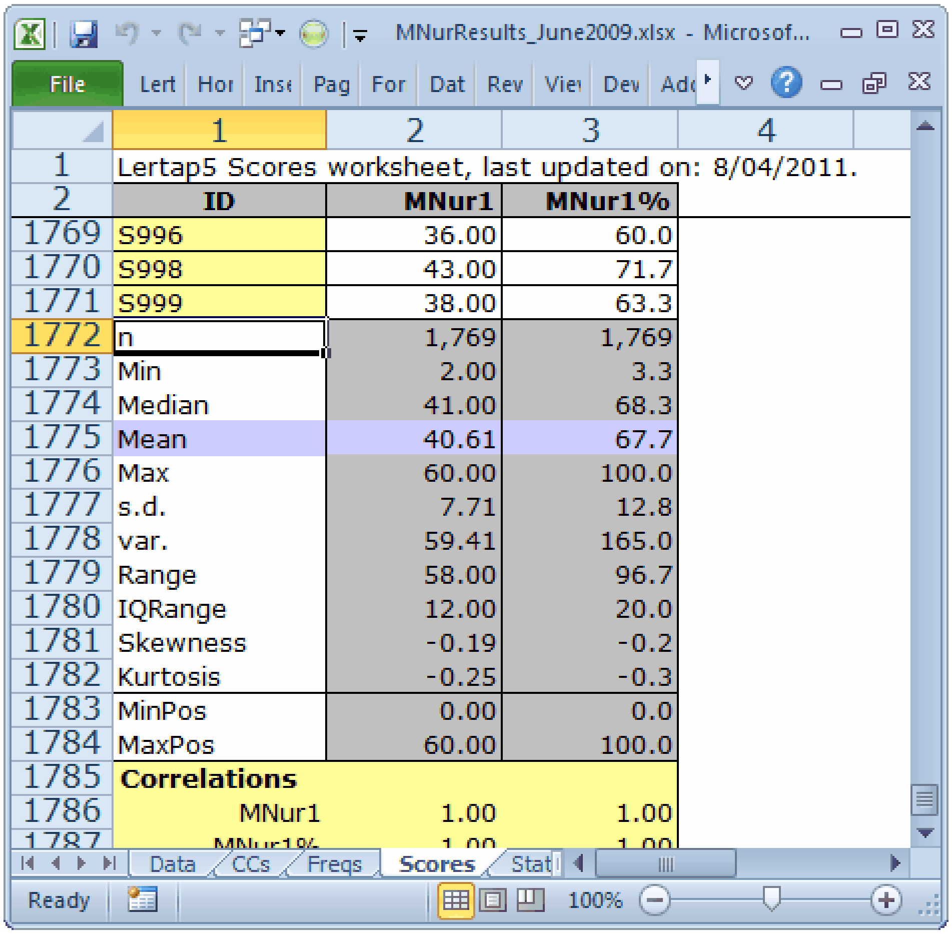
Make it one step?

Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.



A scan of the Freqs report revealed the largest number of Zs for NM10, the item whose responses are recorded in c12 of the Data worksheet. Most items had less than 1% in their Z tally. Many had no Zs, such as NM9.

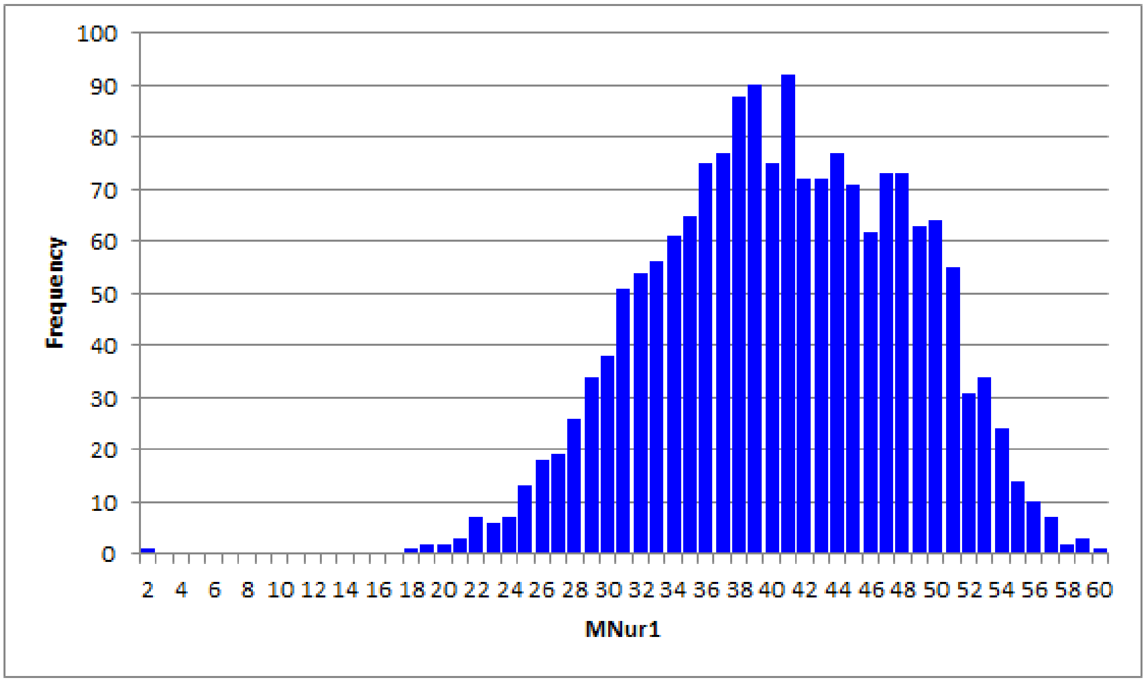


Lertap5 Scores worksheet, last updated on: 8/04/2011.

	ID	MNur1	MNur1%
1769	S996	36.00	60.0
1770	S998	43.00	71.7
1771	S999	38.00	63.3
1772	n	1,769	1,769
1773	Min	2.00	3.3
1774	Median	41.00	68.3
1775	Mean	40.61	67.7
1776	Max	60.00	100.0
1777	s.d.	7.71	12.8
1778	var.	59.41	165.0
1779	Range	58.00	96.7
1780	IQRange	12.00	20.0
1781	Skewness	-0.19	-0.2
1782	Kurtosis	-0.25	-0.3
1783	MinPos	0.00	0.0
1784	MaxPos	60.00	100.0
1785	**Correlations**		
1786	MNur1	1.00	1.00

Whenever mastery mode is on in Lertap, the Scores report includes both the actual test score and its percentage equivalent.

In this case the lowest score, "Min", is just 2, or 3.3%. A histogram reveals that this is an outlier:

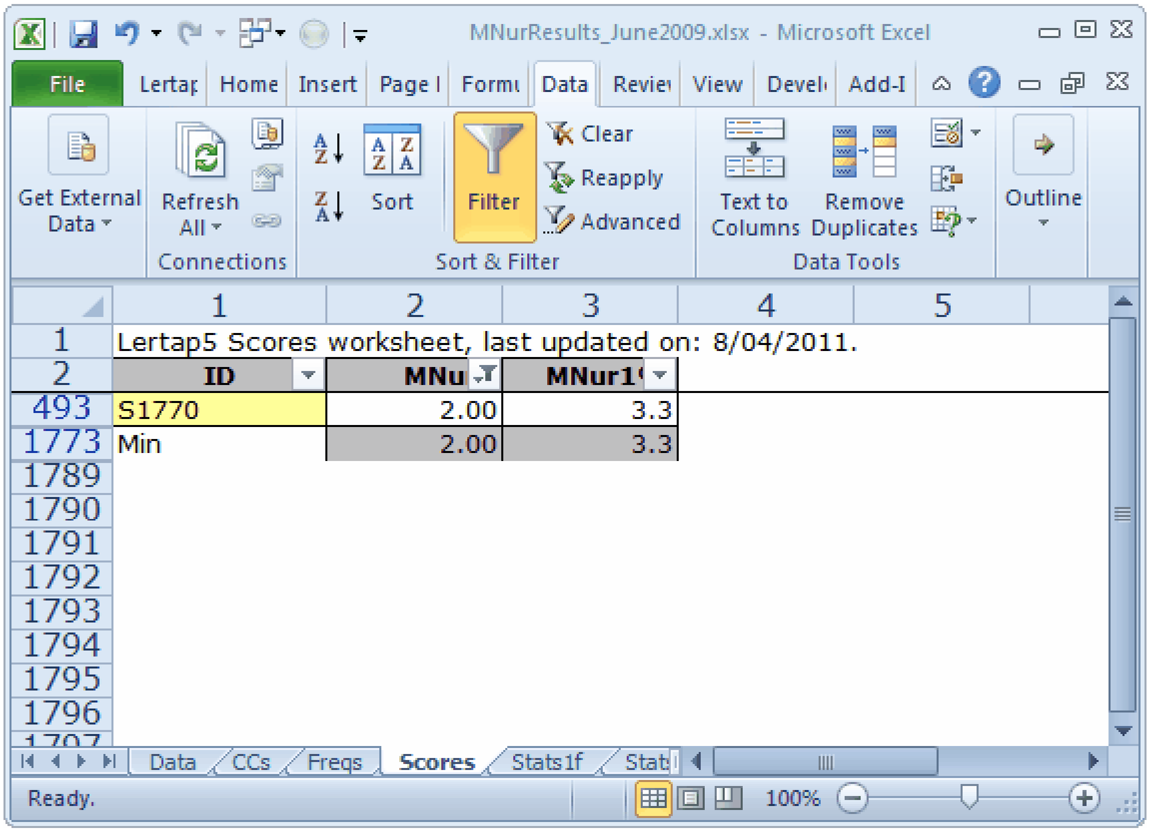


The next-lowest score, after 2, is 18. There is clearly something odd about the score of 2.

How to find out which student got a score of 2?

A quick way is to go to the Scores worksheet, then use Excel's Data tab, then click on Filter.

Click on the little arrowhead in the second column. Unclick Select All. Click 2.00.



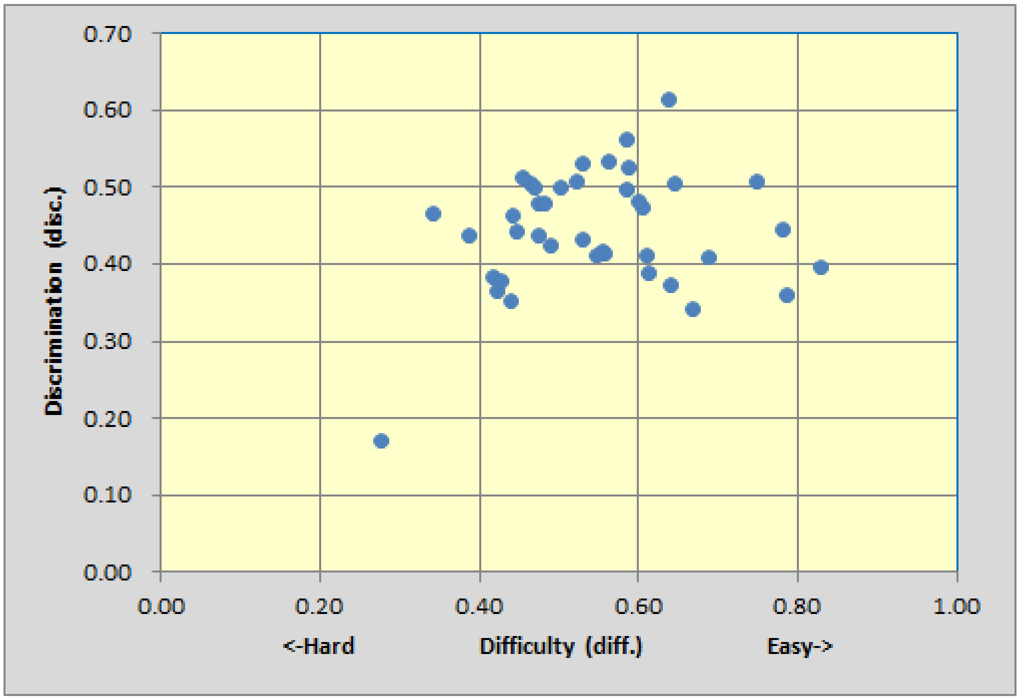
Filter will show all rows in the Scores worksheet which have a value of 2.00 in column 2. The ID of the student with this score is S1770.

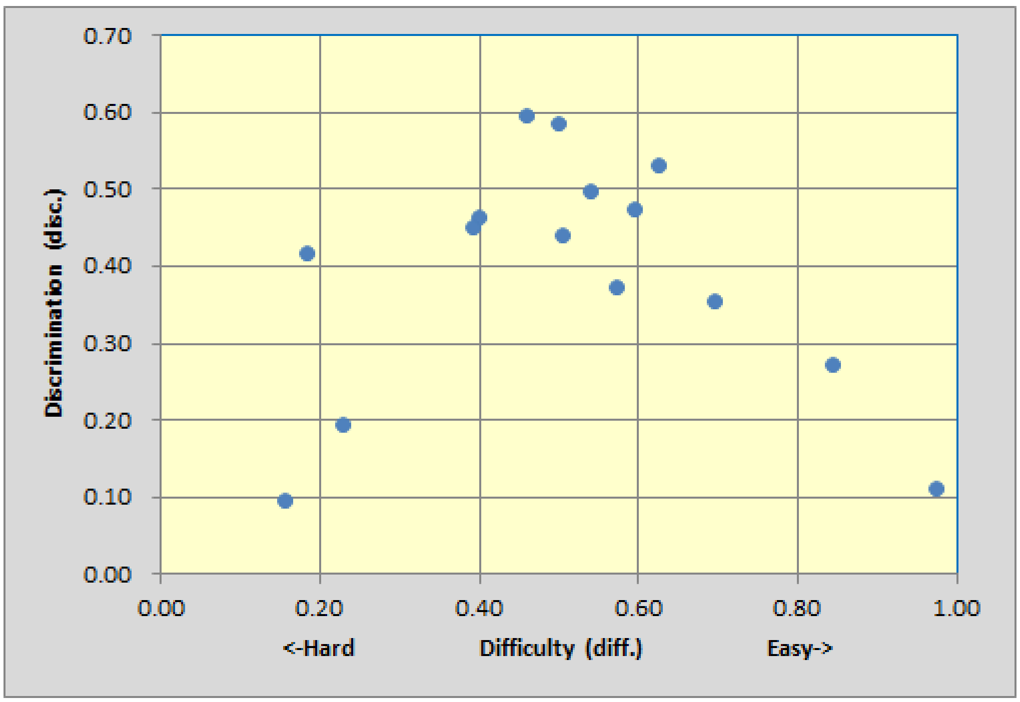
To see item responses for S1770, go to the Data worksheet, and search for S1770 (use "Find", an option on the Home tab). You'll see that this student was from the South American testing site (U1), and that, after item NM12, his or her item "responses" were all Zs. This student's results are (or certainly seem to be) an anomaly. There would be much to say for deleting the student's record from the Data worksheet, and then re-running Lertap's Interpret and Elmillon options. The results seen in the topics immediately following are based on a Data worksheet with this student's results taken out.

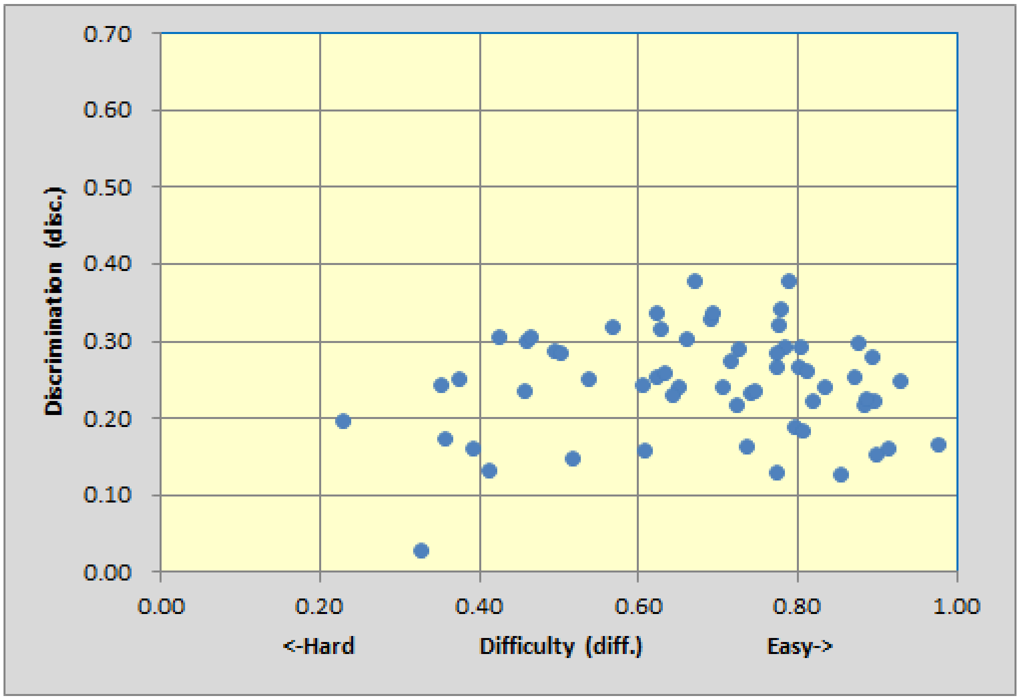
Keep moving forward!

2.10.1 Sample Results

Before getting into some of the results for this test, consider the following three scatterplots from selected Stats1b reports (the last one of these three pertains to this test):


0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Discrimination (disc.)
0.00
0.20
0.40
0.60
0.80
1.00
<-Hard
Difficulty (diff.)
Easy->


0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Discrimination (disc.)
0.00
0.20
0.40
0.60
0.80
1.00
<-Hard
Difficulty (diff.)
Easy->



The top scatter is from a 40-item test with reliability (coefficient alpha) of 0.92.

The middle scatter is from our 15-item MathsQuiz with reliability 0.80 (with I11's key corrected).

The bottom scatter is from the 60-item nursing test with reliability of 0.82.

Is is well known that, all things being equal, a longer test will have better reliability than a shorter one. One way to increase the reliability of a test is to add more items to it.

The **Spearman-Brown prophecy formula** is used to calculate what the reliability of a test would be if it were lengthened by adding more items. We will pretend to add items to the top and middle tests to bring them up to 60 items. This will enable some interesting comparisons to be made.

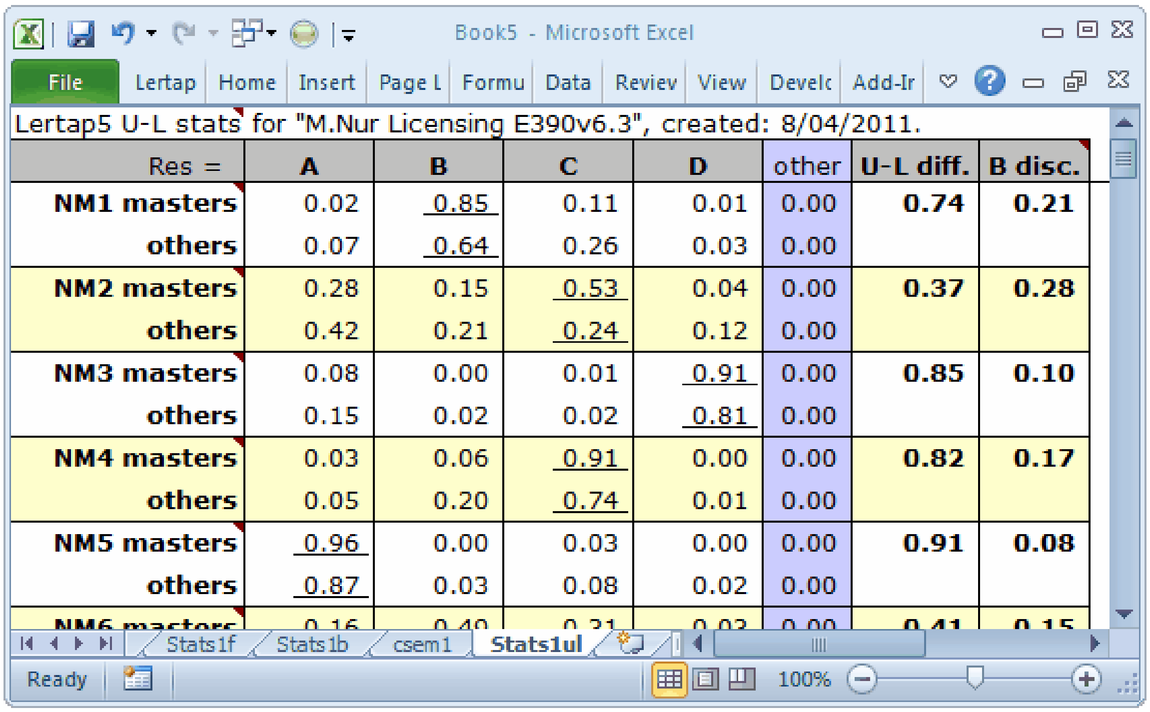
For the 40-item test, making it a 60-item test by adding another 20 items of similar quality, and then applying Spearman-Brown, would bring the reliability to 0.945. Adding 45 items to the 15-item test would raise its reliability to 0.941. We would then have three tests, each with 60 items. Two have reliabilities bordering on 0.94, while one, the nursing exam, has a lower reliability figure, just 0.82.

Why is it that our 60-item nursing exam's reliability is considerably lower than that found for the other two 60-item tests?

Is there something about the top and middle scatterplots which distinguishes them from the bottom one? Yes, for sure. The big difference is that the bottom test has no items with disc. figures above 0.40. Tests will have better reliability as their items come to be more discriminating. In turn, item discrimination is related to item difficulty -- very difficult items, where, say, only 10% get the item right, and very easy items, where only 10% get the item wrong, are obviously not discriminating.

Well and good? Well, maybe not. The nursing exam is a mastery test. Such tests are often intentionally designed to have a greater proportion of items which tend to be relatively easy. This will limit item discrimination figures which, in turn, will serve to bring down reliability, at least as measured by coefficient alpha.

When a *sub line includes the word mastery, Lertap significantly alters one of its reports: Stats1ul. (The other two reports, Stats1b and Stats1f, are not affected by using the word mastery on *sub.)



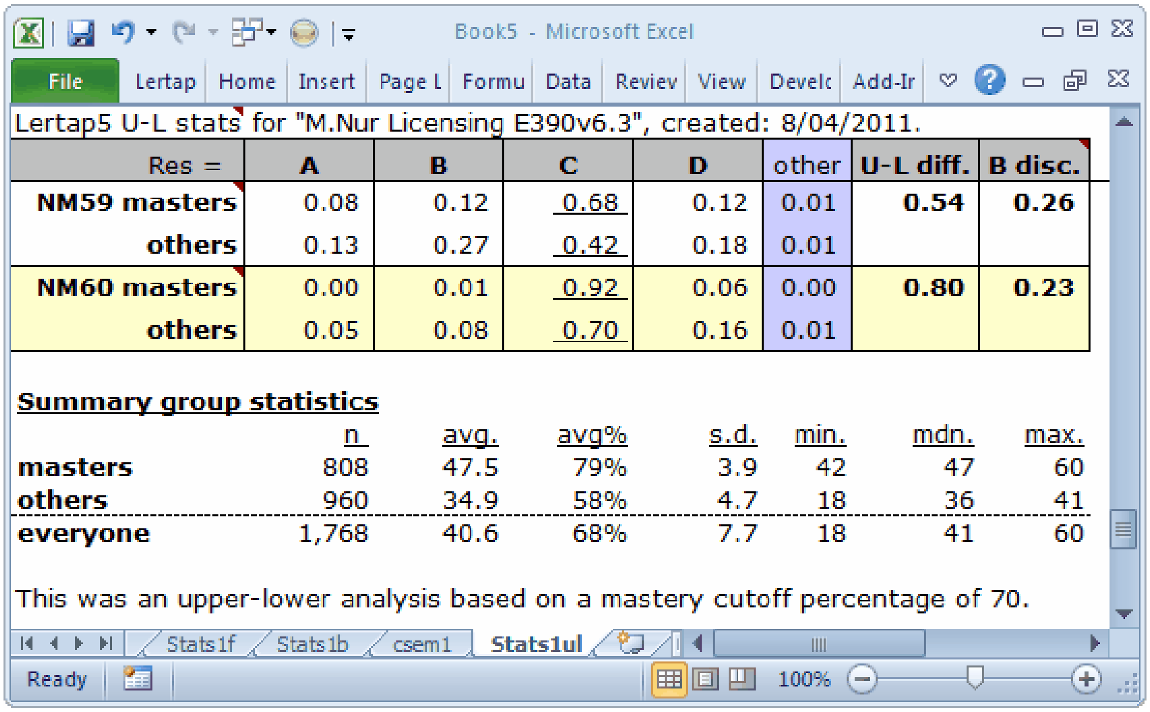
Lertap5 U-L stats for "M.Nur Licensing E390v6.3", created: 8/04/2011.

Res =	A	B	C	D	other	U-L diff.	B disc.
NM1 masters	0.02	0.85	0.11	0.01	0.00	**0.74**	**0.21**
others	0.07	0.64	0.26	0.03	0.00		
NM2 masters	0.28	0.15	0.53	0.04	0.00	**0.37**	**0.28**
others	0.42	0.21	0.24	0.12	0.00		
NM3 masters	0.08	0.00	0.01	0.91	0.00	**0.85**	**0.10**
others	0.15	0.02	0.02	0.81	0.00		
NM4 masters	0.03	0.06	0.91	0.00	0.00	**0.82**	**0.17**
others	0.05	0.20	0.74	0.01	0.00		
NM5 masters	0.96	0.00	0.03	0.00	0.00	**0.91**	**0.08**
others	0.87	0.03	0.08	0.02	0.00		
NM6 masters	0.16	0.40	0.31	0.03	0.00	**0.41**	**0.15**

Test takers have now been classified as "masters" if their test scores are 70% or greater. Otherwise, they fall into what Lertap calls the "others" group.

Stats1ul item summaries indicate the proportion in each group selecting an item option, with the proportions corresponding to the keyed-correct answer underlined.

The "**B disc.**" figure is simply the difference between the proportion correct for the two groups (Brennan, 1972).



Lertap5 U-L stats for "M.Nur Licensing E390v6.3", created: 8/04/2011.

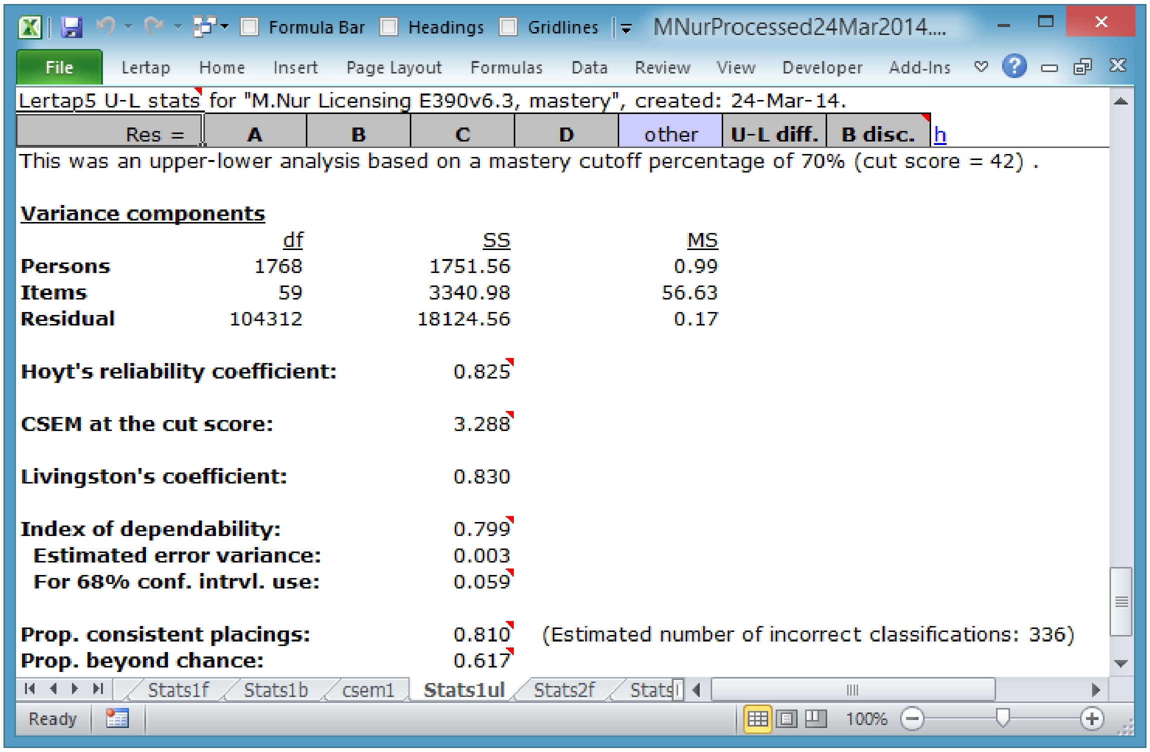
Res =	A	B	C	D	other	U-L diff.	B disc.
NM59 masters	0.08	0.12	0.68	0.12	0.01	**0.54**	**0.26**
others	0.13	0.27	0.42	0.18	0.01		
NM60 masters	0.00	0.01	0.92	0.06	0.00	**0.80**	**0.23**
others	0.05	0.08	0.70	0.16	0.01		

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
masters	808	47.5	79%	3.9	42	47	60
others	960	34.9	58%	4.7	18	36	41
everyone	1,768	40.6	68%	7.7	18	41	60

This was an upper-lower analysis based on a mastery cutoff percentage of 70.

The summary group statistics indicate that 960 students failed to reach the mastery score on this test. This is a substantial number, corresponding to 54% of the students.



The bottom of the Stats1ul report for the mastery test analysis displays statistics discussed in Chapter 7 of the manual, in a technical paper on the use of cut scores in Lertap, and, most completely, in "Lelp", Lertap's online help system.

Of these statistics, the "**Prop. consistent placings:**", at 0.810, suggests that about 81% of the test takers have been correctly classified as being either in the "masters" group, or the "others" group. If we could test the same students again, the great majority, about 81%, would end up with the same classification.

This is what is meant by "**classification consistency**"--we've used the test once as a means of classifying the students, finding some to be "masters" and others not masters. Now, let's imagine that it's possible to use the test again, just as a check: how many of the students will end up with the same classification?

Hopefully it'll be all of them, but in reality this will seldom be the case. In this M.Nursing example we have a classification consistency estimate of 81%. The implication is that *19% of the students might be expected to change if we could test them again.* Some students will go from "masters" to "others"; some students will go from "others" to "masters". This is *approximately a fifth of the students*, some 336 of them (*something to think about, isn't it?*).

For more about classification consistency, read that technical paper. And see the "La Florida" sample too — one of the two tests used in La Florida had a classification consistency, as measured by "Prop. consistent placings", of over 96%.

Next topic?

Tidbits:

It is possible for a test to have items with good discrimination, and yet end up with relatively low reliability. This may happen (for example) when the test is not "pure", that is, has items which are really testing different, and relatively unrelated, things, or "factors". A **principal components** analysis may be used to see if this might be happening; read about it in one of the Lertap "erudite epistles". (It's the "*About eigenvalues, scree tests, and coefficient alpha*" paper; Download it if you'd like. Read about Lertap and principal components here.)

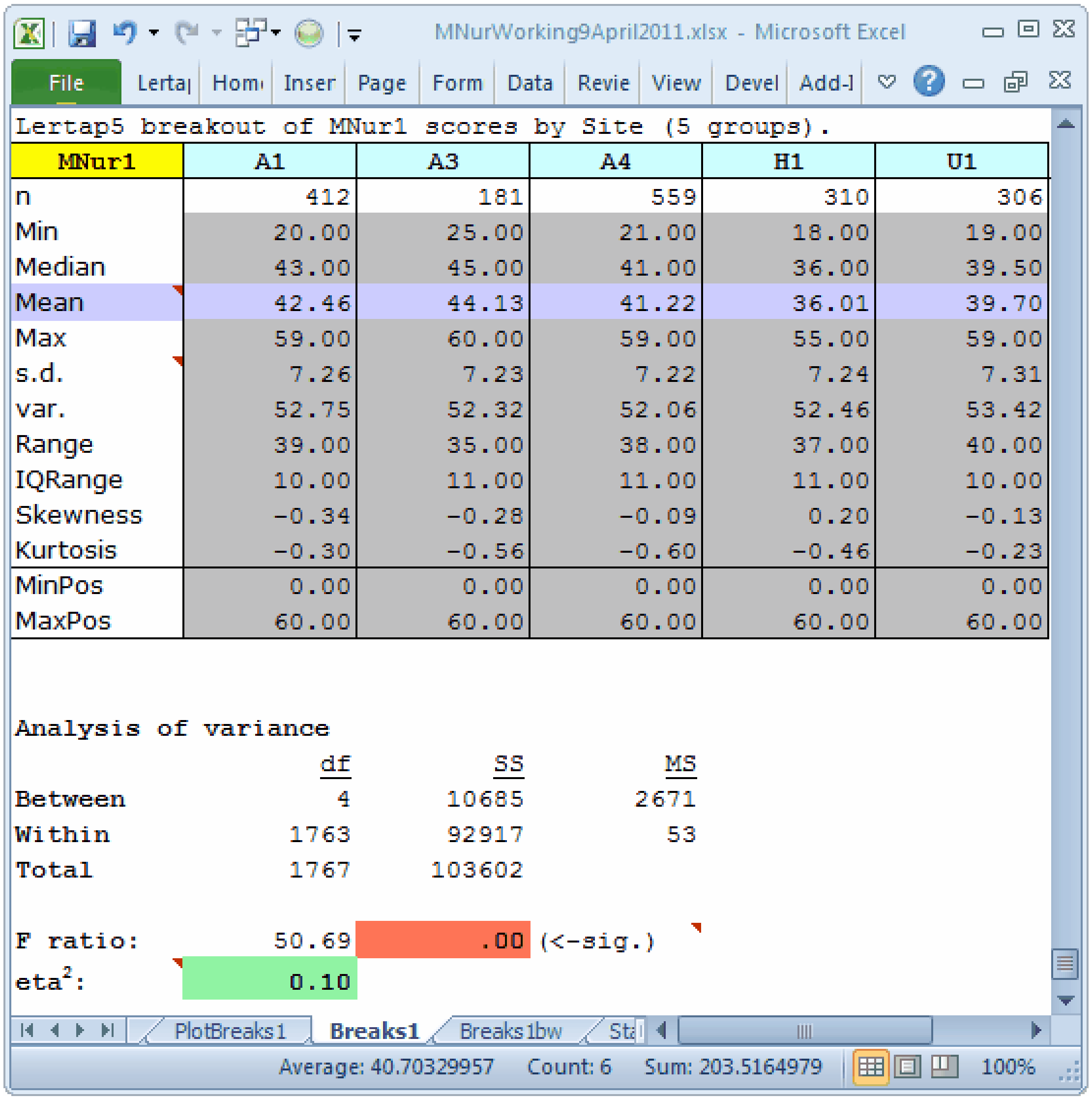
The technical paper referred to just above about the use of cut scores points out that a test may have low reliability, and yet have a high "Prop. consistent placings". Here's another chance to download it.

2.10.2 Group differences

Lertap may be used to look at how the five groups in this example may have differed in their scores, and in their item responses.

The best reference for how to use Lertap to look for differences among scores is here.

In this example, the groups are test sites; their codes are found in column 2 of the Data worksheet. Here are excerpts from two of Lertap's group breakouts reports:



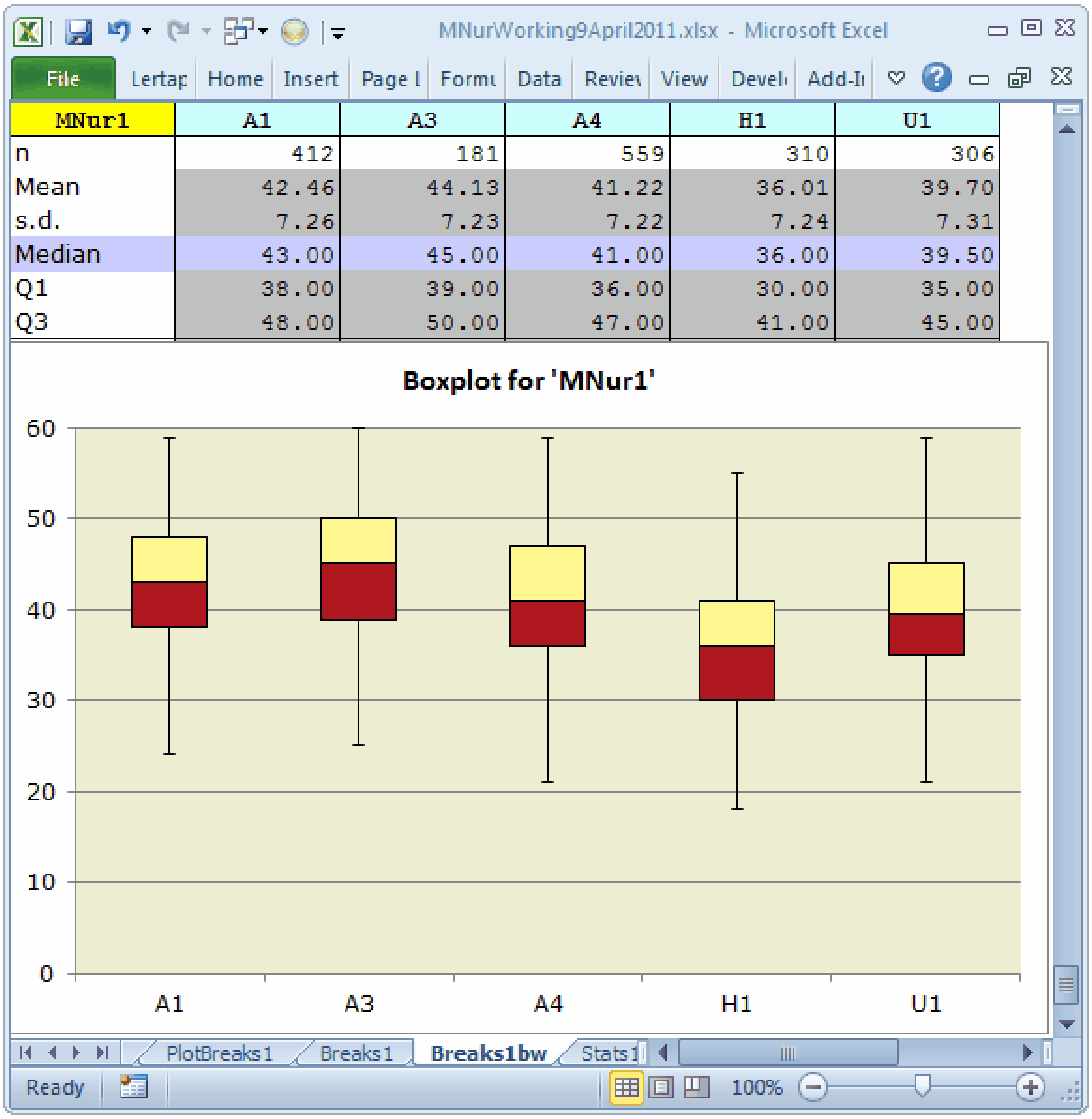
Lertap5 breakout of MNur1 scores by Site (5 groups).

MNur1	A1	A3	A4	H1	U1
n	412	181	559	310	306
Min	20.00	25.00	21.00	18.00	19.00
Median	43.00	45.00	41.00	36.00	39.50
Mean	42.46	44.13	41.22	36.01	39.70
Max	59.00	60.00	59.00	55.00	59.00
s.d.	7.26	7.23	7.22	7.24	7.31
var.	52.75	52.32	52.06	52.46	53.42
Range	39.00	35.00	38.00	37.00	40.00
IQRange	10.00	11.00	11.00	11.00	10.00
Skewness	-0.34	-0.28	-0.09	0.20	-0.13
Kurtosis	-0.30	-0.56	-0.60	-0.46	-0.23
MinPos	0.00	0.00	0.00	0.00	0.00
MaxPos	60.00	60.00	60.00	60.00	60.00

Analysis of variance

	df	SS	MS
Between	4	10685	2671
Within	1763	92917	53
Total	1767	103602	

F ratio:	50.69	.00	(<-sig.)
eta^2:	0.10		



MNur1	A1	A3	A4	H1	U1
n	412	181	559	310	306
Mean	42.46	44.13	41.22	36.01	39.70
s.d.	7.26	7.23	7.22	7.24	7.31
Median	43.00	45.00	41.00	36.00	39.50
Q1	38.00	39.00	36.00	30.00	35.00
Q3	48.00	50.00	47.00	41.00	45.00

The **Breaks1** report, top screen snap above, suggests that there are statistically-significant differences among the five groups (F=50.69, p<.00). When sample sizes are large, care is often required when interpreting the significance of an F ratio as it may be an artefact of sample size, and potentially not meaningful. In such cases we'd be sure to take in the eta^2 figure. It corresponds to the **correlation ratio**, a not-uncommon index of practical significance, an "**effect-size**" estimator, having a range of 0.00 (no effect) to 1.00 (big effect). Here, eta^2 = 0.10, meaning that 10% of the total-score variance observed in these five samples is related to group membership.

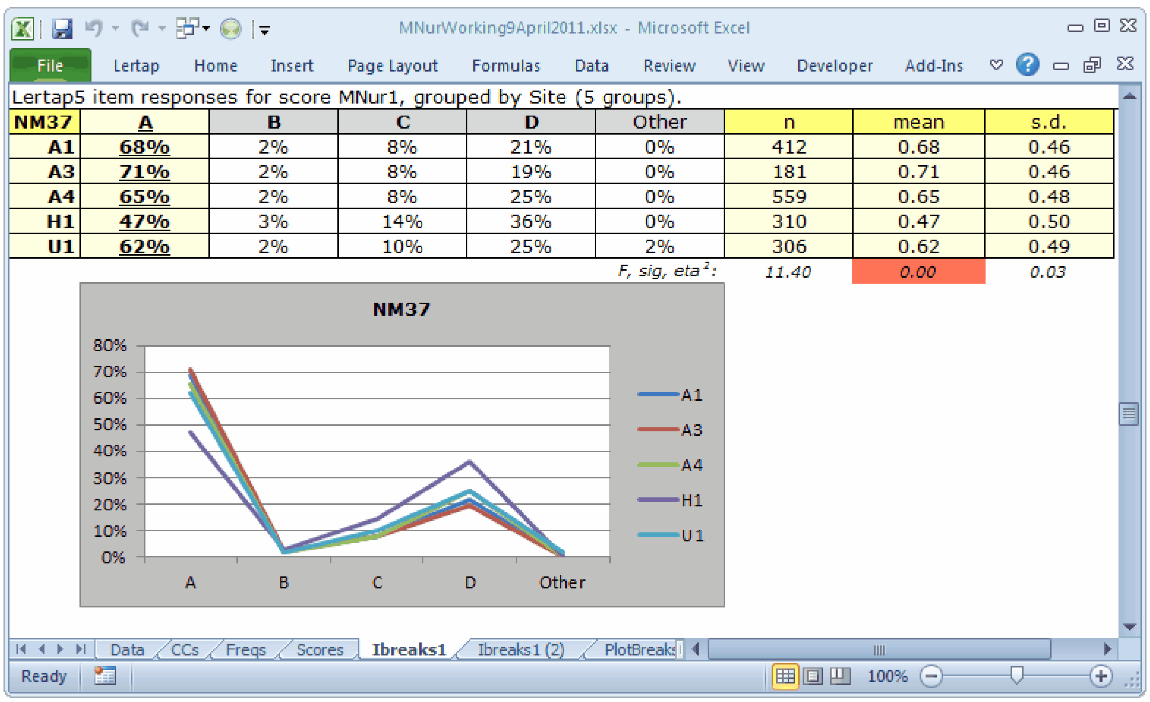
The "significant" F and the 0.10 eta^2 are both suggesting that there are total-score differences among the groups.

But, fortunately perhaps, we don't have to rely on these statistics for what we want to do here. If there are meaningful group differences, we should be able to see them in a graph.

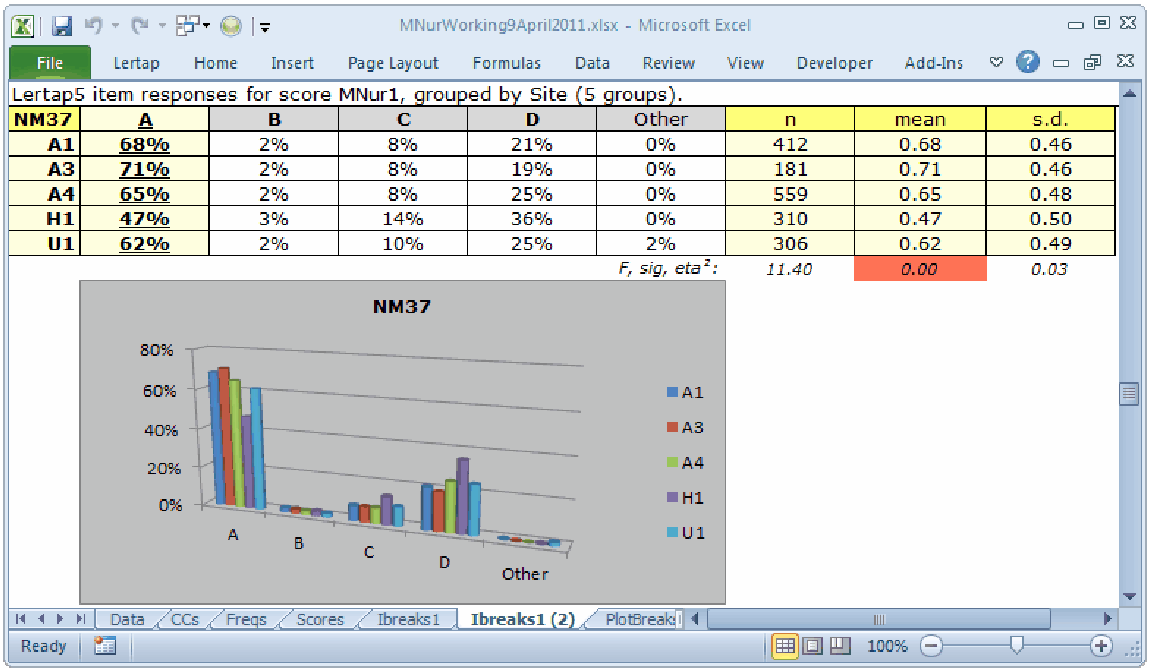
And we do. Look at the **Boxplot** seen in the **Breaks1bw** report above.

There are indeed differences in the group test scores, the "total scores". The best results were turned in by the three sites whose codes begin with A. As mentioned earlier, these sites were all in the same country, one where English is by far the predominant language. The H1 and U1 sites were situated in countries where English is not the dominant language. This may have been a reason for the lower scores found in U1 and, especially, in H1. (There could be other reasons, of course, but we know that the English-language "problem" was of special interest in this case.)

We could try to find out more. We could ask Lertap to prepare some new reports which summarize results at the item level. This level of analysis is referred to as "Ibreaks". The best reference for Ibreaks is currently here.


Lertap5 item responses for score MNur1, grouped by Site (5 groups).

NM37	A	B	C	D	Other	n	mean	s.d.
A1	68%	2%	8%	21%	0%	412	0.68	0.46
A3	71%	2%	8%	19%	0%	181	0.71	0.46
A4	65%	2%	8%	25%	0%	559	0.65	0.48
H1	47%	3%	14%	36%	0%	310	0.47	0.50
U1	62%	2%	10%	25%	2%	306	0.62	0.49
					F, sig, eta²:	11.40	0.00	0.03



NM37	A	B	C	D	Other	n	mean	s.d.
A1	68%	2%	8%	21%	0%	412	0.68	0.46
A3	71%	2%	8%	19%	0%	181	0.71	0.46
A4	65%	2%	8%	25%	0%	559	0.65	0.48
H1	47%	3%	14%	36%	0%	310	0.47	0.50
U1	62%	2%	10%	25%	2%	306	0.62	0.49
					F, sig, eta²:	11.40	0.00	0.03

The two tables and corresponding charts above are for a single item, NM37. Only 47% of the H1 students got this item right while over 60% got it right at the other four sites. For some reason, distractor D was more popular with the H1 students.

The chart seen in the first Ibreaks1 table for item NM37 is called a "Line" chart in Excel. These charts can be a bit noisy when there are many groups, that is, difficult to readily interpret.

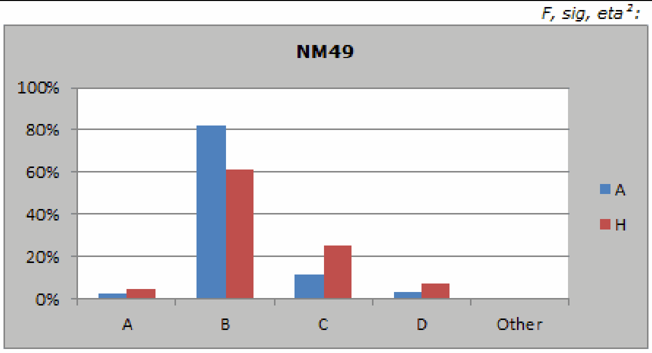
The second chart is a "Clustered Cylinder" Column chart. Is it easier to interpret? Perhaps yes, perhaps not, depending on your point of view.

What we might like to do is compare H1 with the three A sites. This might enable us to more easily detect where the H1 students seemed to go astray, at least when compared to the A sites.

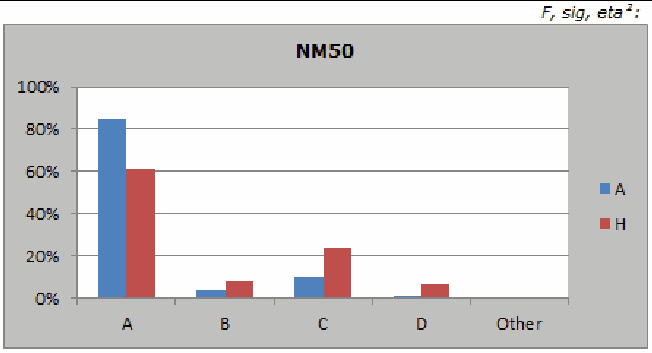
We'll use Lertap's recoder to (1) combine the three A sites into one group, with a code of A; (2) change the H1 code to H; and (3), exclude the U1 site altogether. Use will also be made of the "chart changer" macro, a tool which turns Excel Line charts into something else, such as (in the case below), Column charts.

Lertap5 item responses for score MNur1, grouped by Site recoded (2 groups).

NM49	A	**B**	C	D	Other	n	mean	s.d.
A	3%	**82%**	12%	3%	0%	1152	0.82	0.38
H	5%	**62%**	25%	7%	1%	310	0.62	0.49
					F, sig, eta²:	*61.16*	*0.00*	*0.04*



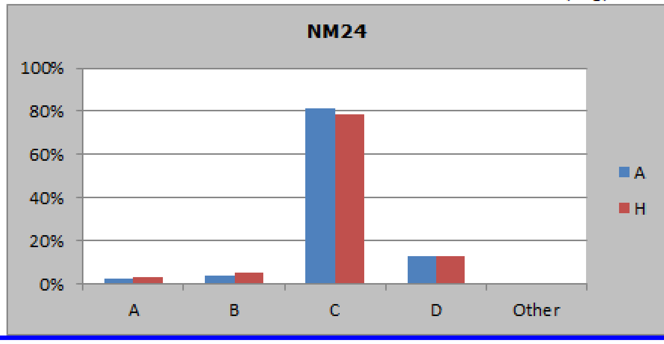
NM50	**A**	B	C	D	Other	n	mean	s.d.
A	**85%**	4%	10%	1%	0%	1152	0.85	0.36
H	**61%**	8%	24%	7%	0%	310	0.61	0.49
					F, sig, eta²:	*89.35*	*0.00*	***0.06***



Here we can see two items, NM49 and NM50, where the H students had noticeably poorer results.

Should you muster your forces, and repeat this work, you'll find that there were numerous other items where the A / H differences were also quite marked. There were also some, such as NM24, with almost no difference:

NM24	A	B	C	D	Other	n	mean	s.d.
A	2%	4%	**81%**	13%	0%	1152	0.81	0.39
H	3%	5%	**79%**	13%	0%	310	0.79	0.41
					F, sig, eta²:	*1.01*	*0.31*	*0.00*



The large number of items with weak results for the H site led to a mini-investigation of possible problems. Would you like to know what was the outcome? So would we. Language differences were thought to be a likely problem, but then the U1 site, in another country where English is not the dominant language at all, returned results which were not as weak. This test had been administered before, but apparently was not subjected to the sort of scrutiny we have exemplified here. At the end of the day, the results of the mini-investigation have remained confidential.

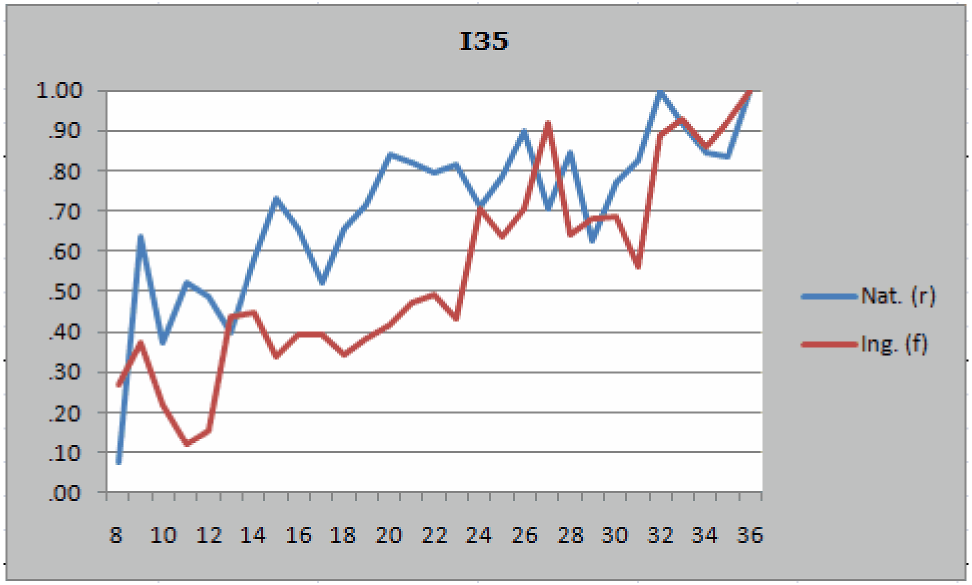
Adelante!

Tidbits:

The histogram option likes to work with a Breaks1 report. Can be useful.

Lertap has support for **DIF**, differential item functioning analysis. You might have an interest in DIF when you have two groups known to be equal in ability (or proficiency), and want to see if there were differences in how they did on individual test items. What's that you say? Haven't we just seen a job with two groups, A and H? Yes, good point -- but DIF digs much deeper, making even more informative charts, with even more statistics. Not to be missed. Check it out with on-line help.

Here's an example of a DIF chart from Lertap / Excel, indicating that on an item called I35, a greater proportion of the blue group got I35 right over most of the score range selected for the plot, 8 to 36:



2.10.3 Download

The M.Nursing workbook may be downloaded by clicking here.

Some worthy questions: how many items get mentioned in the Stats1b **? column**?

What do the quintile plots look like? (They're no longer really quintiles as, with the mastery option on, only two groups are involved.)

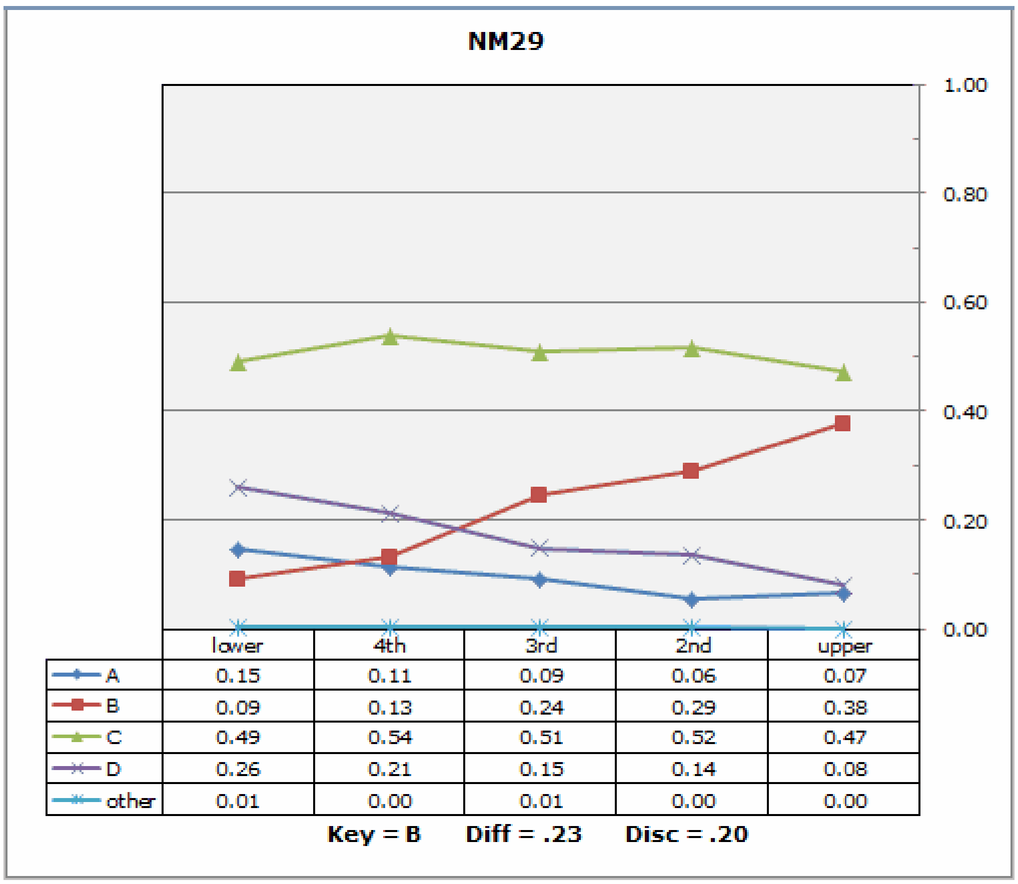
To get a "conventional" analysis in addition to the mastery analysis, use these *CCs lines:

```
*col(c3-c62)
*sub Mastery, Title=(MNur1), Name=(M.Nur Licensing E390v6.3)
*key BCDCA BBBCD BDCDA BDABD BBACB BBCBA BDAAB AAACD CBABA
DABBA BBDDC BBCCC
*col(c3-c62)
*sub Title=(NM Nur), Name=(M.Nur Licensing E390v6.3)
*key BCDCA BBBCD BDCDA BDABD BBACB BBCBA BDAAB AAACD CBABA
DABBA BBDDC BBCCC
```

These lines define two subtests. They're really identical as they employ the same items, but the second *sub line does not include the word Mastery. The Stats**1**ul report will have just two groups, "masters" and "others", but the Stats**2**ul report will have the quintiles, the five groups, ranging from "lower" to "upper".

(Note: there's no need at all to use different colors for CCs lines, as we have done here -- but a splash of color can make the lines corresponding to each subtest easier to contemplate.)

Question NM29 is interesting. It has a very popular distractor, yet does not get flagged in the **? column.** Its quintile plot is shown here:



Note (above) the popularity of distractor C, even in the upper group. Why isn't this distractor flagged in the corresponding Statsb report? (Refer to the corresponding Statsf report for the answer -- but first you have to download the workbook, and make changes to the CCs worksheet.)

What would happen to "Prop. consistent placings" if Mastery=40%? (You might be surprised.)

*col(c3-c62)
*sub Mastery, Title=(MNur-70), Name=(M.Nur Licensing E390v6.3)
*key BCDCA BBBCD BDCDA BDABD BBACB BBCBA BDAAB AAACD CBABA DABBA BBDDC BBCCC
*col(c3-c62)
*sub Title=(MNur-NM), Name=(M.Nur Licensing E390v6.3)
*key BCDCA BBBCD BDCDA BDABD BBACB BBCBA BDAAB AAACD CBABA DABBA BBDDC BBCCC
*col(c3-c62)
*sub Mastery=40%, Title=(MNur-40), Name=(M.Nur Licensing E390v6.3)
*key BCDCA BBBCD BDCDA BDABD BBACB BBCBA BDAAB AAACD CBABA DABBA BBDDC BBCCC

Now, with the CCs lines above, we'll have three subtests. The Stats**3**ul report will reveal the effect of Mastery=40 on "Prop. consistent placings". (The result might very well knock your socks off, even if you're not wearing any.)

Tidbit:

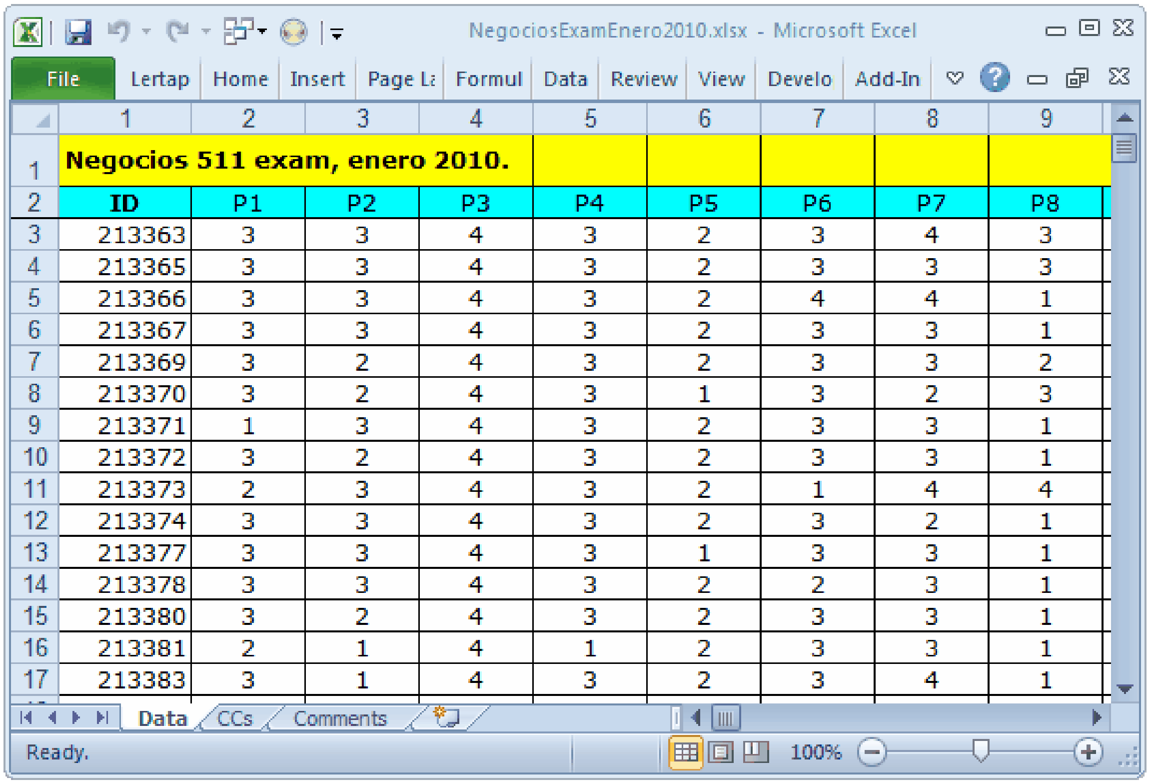
You may encounter a **bug** in Lertap if you try this example. If you're told that a Statsul report cannot be made as "everyone reached mastery", be sure to specify the mastery level each time the Mastery word is used on *sub. Here, for example, put Mastery=70% in the first *sub line, should you run into this problem.

2.11 Negocios

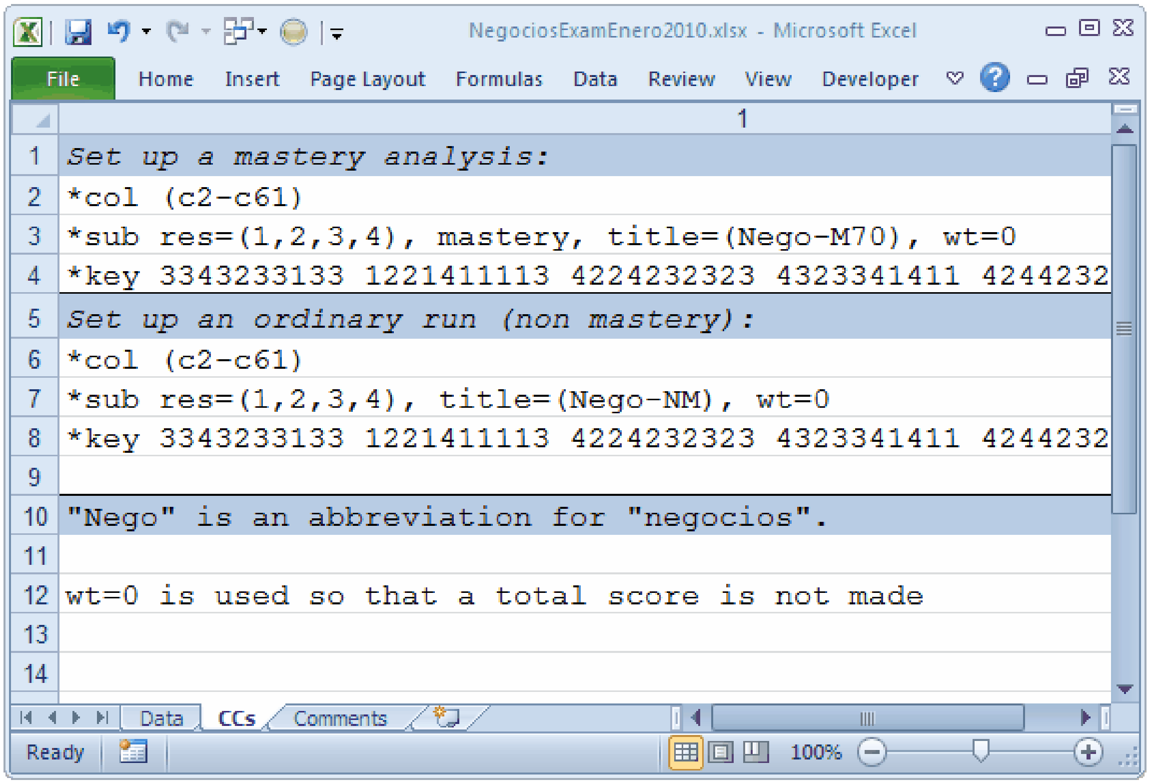
This data set involves exam responses from a postgraduate class in business studies (n=500).

Cheating was suspected by one of the exam proctors, and Lertap's **RSA analysis** was used to see if there may have been some unexpectedly-similar item responses.

And there were. When confronted with evidence, two students confessed that they had cheated by sharing answers during the administration of the test.



	1	2	3	4	5	6	7	8	9
1	Negocios 511 exam, enero 2010.								
2	ID	P1	P2	P3	P4	P5	P6	P7	P8
3	213363	3	3	4	3	2	3	4	3
4	213365	3	3	4	3	2	3	3	3
5	213366	3	3	4	3	2	4	4	1
6	213367	3	3	4	3	2	3	3	1
7	213369	3	2	4	3	2	3	3	2
8	213370	3	2	4	3	1	3	2	3
9	213371	1	3	4	3	2	3	3	1
10	213372	3	2	4	3	2	3	3	1
11	213373	2	3	4	3	2	1	4	4
12	213374	3	3	4	3	2	3	2	1
13	213377	3	3	4	3	1	3	3	1
14	213378	3	3	4	3	2	2	3	1
15	213380	3	2	4	3	2	3	3	1
16	213381	2	1	4	1	2	3	3	1
17	213383	3	1	4	3	2	3	4	1



The Data and CCs worksheets for this workbook are (partially) displayed above. Student responses were processed by a scanner. Unanswered questions were coded as blanks by the scanner; those questions on which a student shaded in more than one bubble were coded as an asterisk (*).

Two subtests are defined by respective *col, *sub, and *key lines. Since the *col line points to the same items, and the *key lines are identical, the only differences to be found in Lertap's reports will be in Stats1ul (mastery) and Stats2ul (non-mastery). The main reason for having the second subtest is so that the Stats2ul report can be used to get quintile plots of item functioning (the mastery U-L report, Stats1ul, will have only two groups, "masters" and "others"; the item response plots from a mastery Statsul report are informative, but having five groups often gives a better idea of the performance of an item's options).

Download this dataset if you'd like, then, as per usual, get results:

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

Next?

Tidbits:

The CCs worksheet seen in this example has several lines which do not begin with an asterisk. These are comments only--because they do not begin with an *, they are ignored by Lertap. There is no limit to the number of comment lines which may be used. Also note: Lertap stops reading CCs lines whenever it encounters a blank line. In this example, line 9 is blank, meaning that Lertap will not bother to read what follows (in this case that's of no consequence anyway as all following lines are comments).

Some tips which are handy at times:

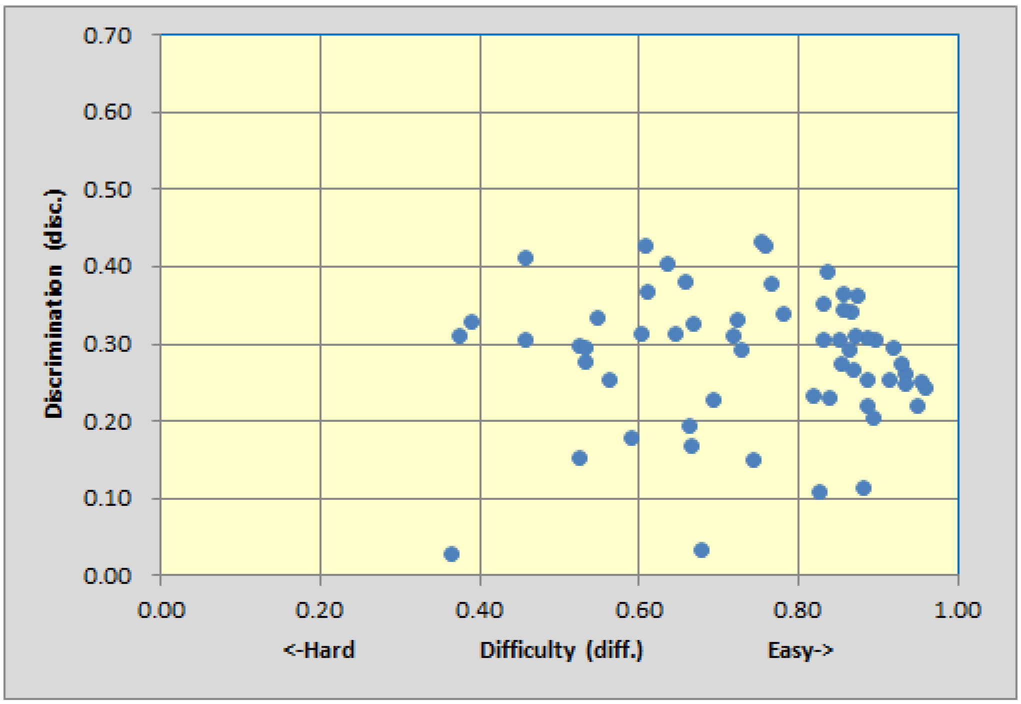
1	Leaving a blank line after the first CCs *col line is okay, and may be used to make sure that Lertap produces only one report: Freqs. The Freqs report is then examined just to determine how the item responses look, checking to see that there are no unexpected "responses".
2	Lertap stops reading Data records after it has encountered a blank line (that is, a blank row) in the Data worksheet. This tip may be used just to check the CCs lines for errors; it saves time when there may be hundreds of results in the Data worksheet -- insert a blank line (row) after, say, ten rows, and you'll quickly see if the CCs lines are doing the job you hope they will.
3	In a job with lots of subtests (remember: each subtest begins with a *col line), any subtest may be taken "off line" by making its lines into comments: just insert a character before each line's asterisk (for example x*col..., x*sub..., x*key... and so on).

2.11.1 Sample results

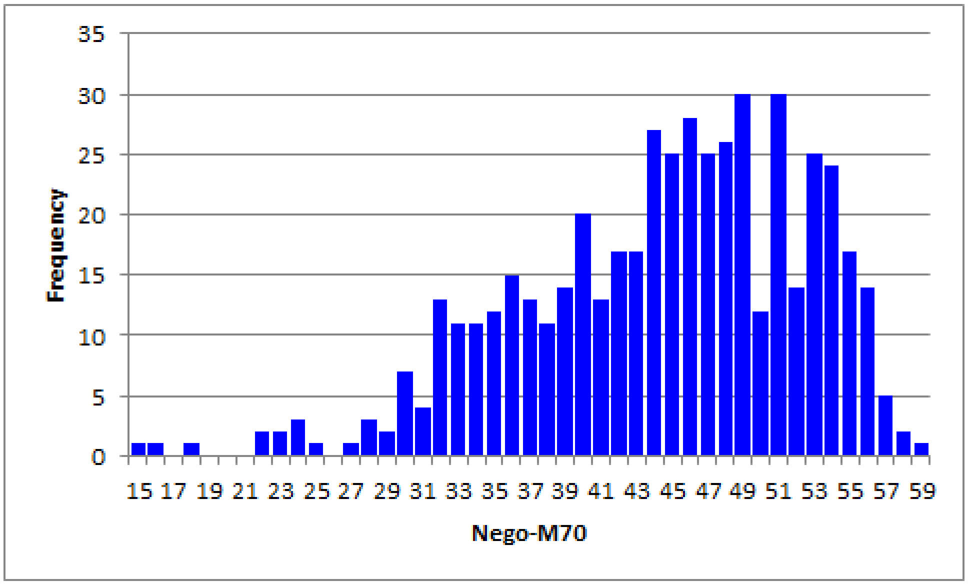
Before applying RSA, be sure to read all of the references first. Start here.

Not all tests will work well with Harpp-Hogan / Lertap RSA. The test should have at least 30 items, and an average score not greater than 80% of the maximum possible score. The test items should use at least four options (true-false questions do not work well with Harpp-Hogan). Bad items, questions with negative discrimination, should be excluded from RSA. Items which have positive discrimination, but have one or more distractors with response proportions much greater than those for the correct answer, should be omitted.

The Stats1f figures (same as Stats2f in this case) show a reliability of 0.86 for this 60-item test.



The test has a large number of easy items (diff. values above 0.80), which is often a characteristic of "mastery" tests. Nonetheless, there are five items with discriminations of 0.40 or greater, and also a fair smattering of items with discrimination in the 0.30 to 0.40 band. Not bad. (The two items whose blips are at the bottom of the scatterplot are P57 on the left, and P17 on the right. Both have disc. values below 0.05.)



The score histogram has three stragglers on the left. We leave it to you to find out if the corresponding data records reveal a preponderance of strange answers -- blanks (unanswered questions), or asterisks (more than one answer selected by a student). (The M.Nursing sample shows how this can be done.)

When we ran Lertap's RSA option, Lertap paused twice, on P15 and P34. We told it to omit P15 but keep P34. Of these two items, P15 had a distractor selected by 45% of the students, quite a bit above the 37% for the correct answer. On P34, one distractor had 48%, with the correct answer at 46%, not much difference (these decisions to omit or keep items are a bit arbitrary -- both P15 and P34 had good discrimination; with 60 items and good reliability, the RSA results would probably have been about the same had P15 been retained for the analysis).

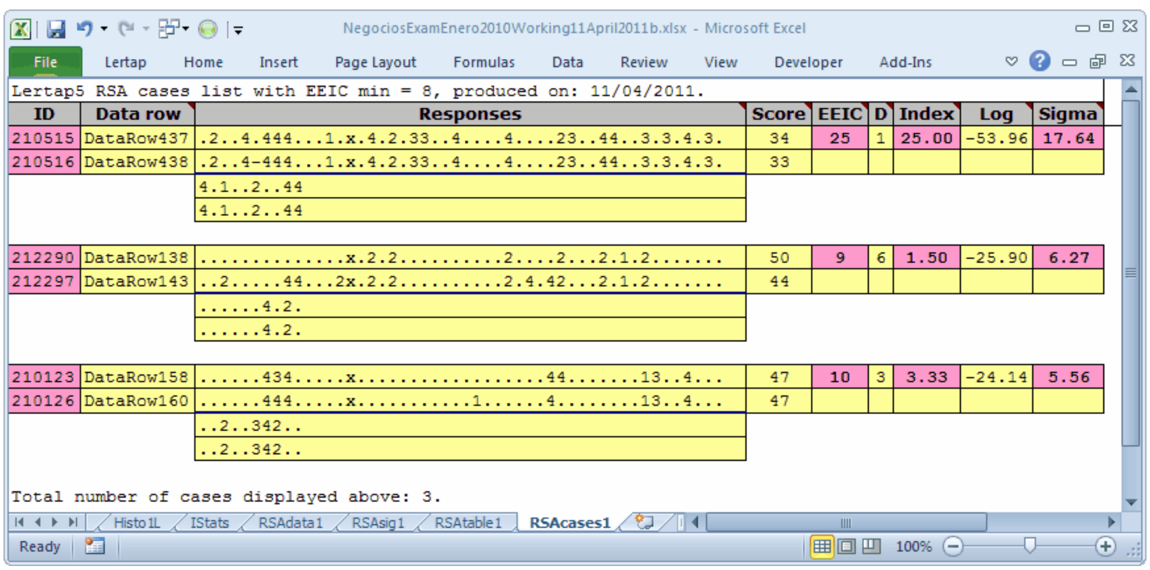
It doesn't make much sense to have Lertap run its RSA analysis with all of the scores. At the low end, say below 32 in this example, we might not be too concerned with cheating as, if it existed, it didn't do the students much good. At the high end, say above 52, there can't be many exact errors in common (EEIC), so students with these scores may as well be omitted. Running RSA with a low score of 32 and high score of 52 is what we decided on for this example.

Pairings	
Suspect:	3
Not suspect:	73,533
Total:	73,536
Inclusions	
Number of items:	59
Number of students:	384
Run control	
EEIC minimum:	8
H-H index minimum:	1.5
H-H sigma minimum:	5
Items excluded:	1
Minimum score setting:	32
Maximum score setting:	52

n	73,533	73,533
minimum	-34.22	-9.64
median	-10.28	0.06
mean	-10.43	0.00
maximum	-2.40	3.25
s.d.	2.47	1.00
variance	6.09	1.00
range	31.81	12.89
IQrange	3.22	1.30
skewness	-0.50	-0.50
kurtosis	0.86	0.86
	expect	**found**
within 1 sigma	68.30%	69.47
1 to 2 sigma	27.20%	26.18
2 to 3 sigma	4.28%	3.72
3 to 4 sigma	0.26%	0.52
4 to 5 sigma	0.01%	0.09
over 5 sigma	0.00%	0.02

The RSAsig1 sheet shows that just 3 pairs were suspect, not many at all. (This is indicated above left, under the "Pairings" section.)

The "expect" and "found" results to the lower right are typical; ideally the values in these two columns will be very similar, more than what has resulted here -- the values on the left, "expect", are those found under a normal, or Gaussian, distribution; those on the right represent the values found in our sample.



Lertap5 RSA cases list with EEIC min = 8, produced on: 11/04/2011.

ID	Data row	Responses	Score	EEIC	D	Index	Log	Sigma
210515	DataRow437	.2..4.444...1.x.4.2.33..4....4....23..44..3.3.4.3.	34	25	1	25.00	-53.96	17.64
210516	DataRow438	.2..4-444...1.x.4.2.33..4....4....23..44..3.3.4.3.	33					
		4.1..2..44						
		4.1..2..44						
212290	DataRow138	.............x.2.2.........2....2...2.1.2.......	50	9	6	1.50	-25.90	6.27
212297	DataRow143	..2.....44...2x.2.2.........2.4.42...2.1.2.......	44					
		......4.2.						
		......4.2.						
210123	DataRow158	......434.....x.................44.......13..4...	47	10	3	3.33	-24.14	5.56
210126	DataRow160	......444.....x..........1......4.......13..4...	47					
		..2..342..						
		..2..342..						

Total number of cases displayed above: 3.

The RSAcases1 report reveals the identities of the 3 suspect pairs of students. Students 210515 and 210516 had only one response difference over the 60 items. On 25 of the questions, they chose exactly the same wrong answer (the same distractor).

These two students were brothers, and, when invited in to "chat with the Dean", they admitted that they had cheated. For reasons unknown to us, the other two pairs of students were not challenged.

Moving on?

Tidbit:

There's much more to read about with regard to Lertap and its RSA analysis. Don't cheat yourself out of some top-flight reading. Start here.

2.11.2 Download

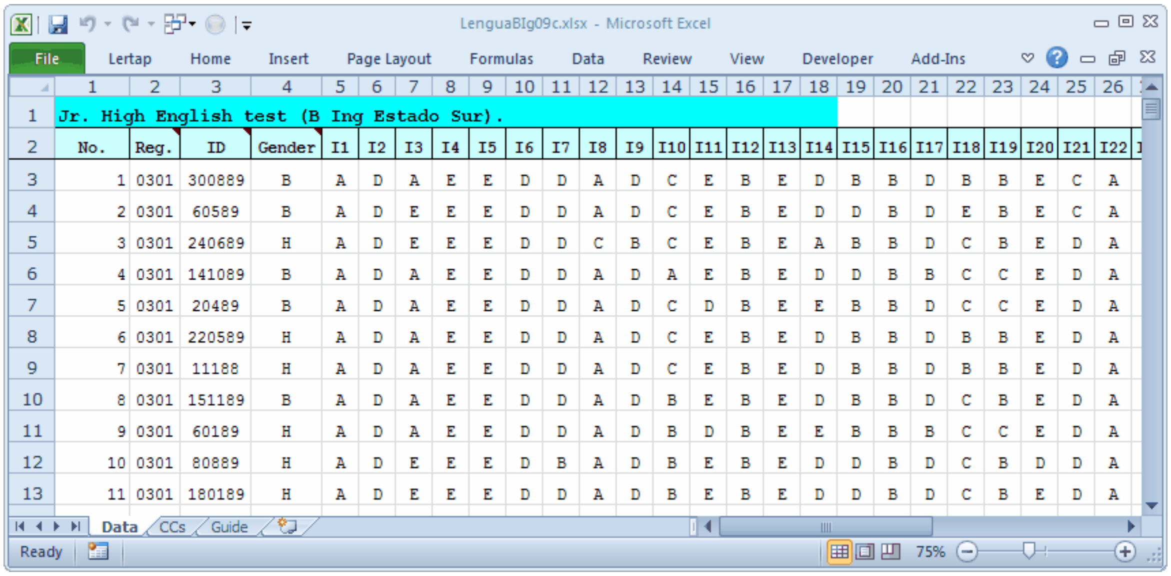
The Negocios workbook may be downloaded by clicking here.

2.12 LenguaBIg

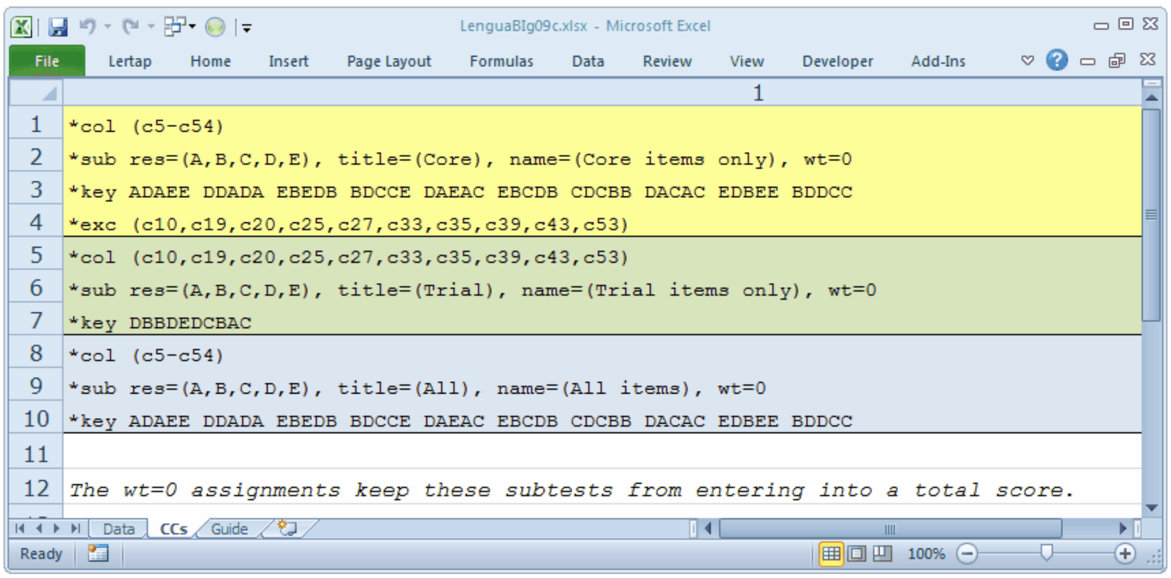
These results are from a **high-stakes** test of topics having to do with characteristics of the English language. It was given to over 5,000 junior high school students in an Asian country.

Students had to have a score of at least 50% to pass the test. However, in our discussion of this sample, the pass rate will not be our focus. (Please see the M.Nursing sample for mention of cut scores, pass rates, and mastery testing.)

Here, we'll concentrate on using multiple subtests, and an external criterion analysis, in order to see how Lertap may be used to check out the quality of trial, or "pretest", or "**pilot**" items. These are questions embedded in the exam which are not scored; these questions are themselves under test, as it were, to see if they're good enough.

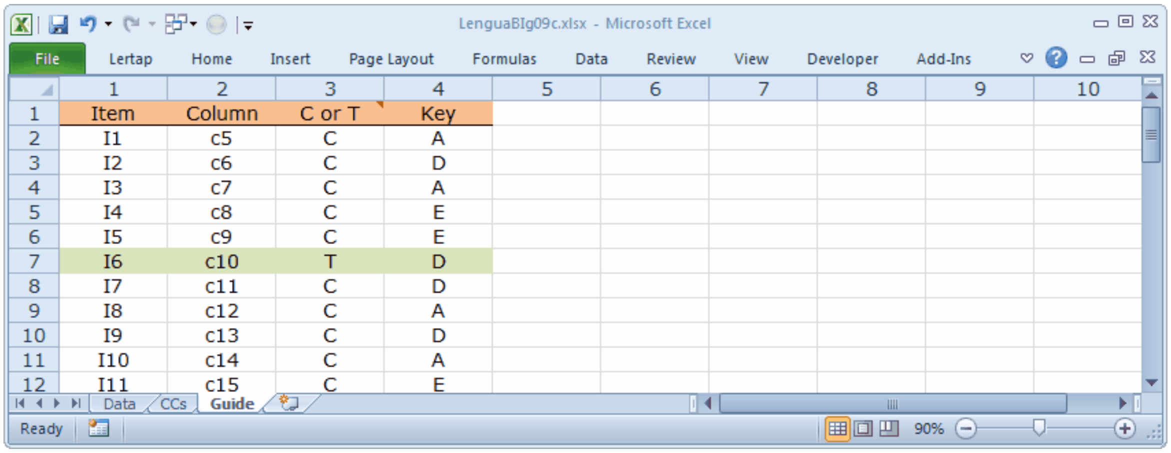


Jr. High English test (B Ing Estado Sur).																									
No.	Reg.	ID	Gender	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18	I19	I20	I21	I22
1	0301	300889	B	A	D	A	E	E	D	D	A	D	C	E	B	E	D	B	B	D	B	B	E	C	A
2	0301	60589	B	A	D	E	E	E	D	D	A	D	C	E	B	E	D	D	B	D	E	B	E	C	A
3	0301	240689	H	A	D	E	E	E	D	D	C	B	C	E	B	E	A	B	B	D	C	B	E	D	A
4	0301	141089	B	A	D	A	E	E	D	D	A	D	A	E	B	E	D	D	B	B	C	C	E	D	A
5	0301	20489	B	A	D	A	E	E	D	D	A	D	C	D	B	E	E	B	B	D	C	C	E	D	A
6	0301	220589	H	A	D	A	E	E	D	D	A	D	C	E	B	E	D	B	B	D	B	B	E	D	A
7	0301	11188	H	A	D	A	E	E	D	D	A	D	C	E	B	E	D	B	B	D	B	B	E	D	A
8	0301	151189	B	A	D	A	E	E	D	D	A	D	B	E	B	E	D	B	B	D	C	B	E	D	A
9	0301	60189	H	A	D	A	E	E	D	D	A	D	B	D	B	E	E	B	B	B	C	C	E	D	A
10	0301	80889	H	A	D	E	E	E	D	B	A	D	B	E	B	E	D	D	B	D	C	B	D	D	A
11	0301	180189	H	A	D	E	E	E	D	D	A	D	B	E	B	E	D	D	B	D	C	B	E	D	A



```
*col (c5-c54)
*sub res=(A,B,C,D,E), title=(Core), name=(Core items only), wt=0
*key ADAEE DDADA EBEDB BDCCE DAEAC EBCDB CDCBB DACAC EDBEE BDDCC
*exc (c10,c19,c20,c25,c27,c33,c35,c39,c43,c53)
*col (c10,c19,c20,c25,c27,c33,c35,c39,c43,c53)
*sub res=(A,B,C,D,E), title=(Trial), name=(Trial items only), wt=0
*key DBBDEDCBAC
*col (c5-c54)
*sub res=(A,B,C,D,E), title=(All), name=(All items), wt=0
*key ADAEE DDADA EBEDB BDCCE DAEAC EBCDB CDCBB DACAC EDBEE BDDCC
```

The wt=0 assignments keep these subtests from entering into a total score.



	Item	Column	C or T	Key
2	I1	c5	C	A
3	I2	c6	C	D
4	I3	c7	C	A
5	I4	c8	C	E
6	I5	c9	C	E
7	I6	c10	T	D
8	I7	c11	C	D
9	I8	c12	C	A
10	I9	c13	C	D
11	I10	c14	C	A
12	I11	c15	C	E

This 50-item test had 40 core items, and 10 trial, or "pretest", items.

The trial items were I6, I15, I16, I21, I23, I29, I31, I35, I39, and I49. These items were not scored. They were included to see how well they would perform; if they turned out well, they would be considered for inclusion in a future version of the test.

The first subtest, "Core", has an ***exc** line which serves to exclude the trial items. The second subtest, "Trial", has only the trial items, while "All", the last subtest, uses all 50 items.

The "Guide" worksheet was used to make it easier to write the CCs lines; what is particularly useful is the "map" which links items to columns.

The "C or T" column in this "Guide" indicates if an item was a "Core" item, or a "Trial" item.

The first trial item is I6, found in c10 of the Data worksheet. (Note that c10 is the first entry in the *exc line.) The map also makes it easier to write the *key line for the trial items. But fear not: you don't have to have a worksheet like our "Guide". It has nothing to do with what Lertap does -- it's just a helper for making CCs lines easier to put together.

Okay then, let's get some results, eh? We'll go through the 3 Steps, and meet you in the next topic.

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

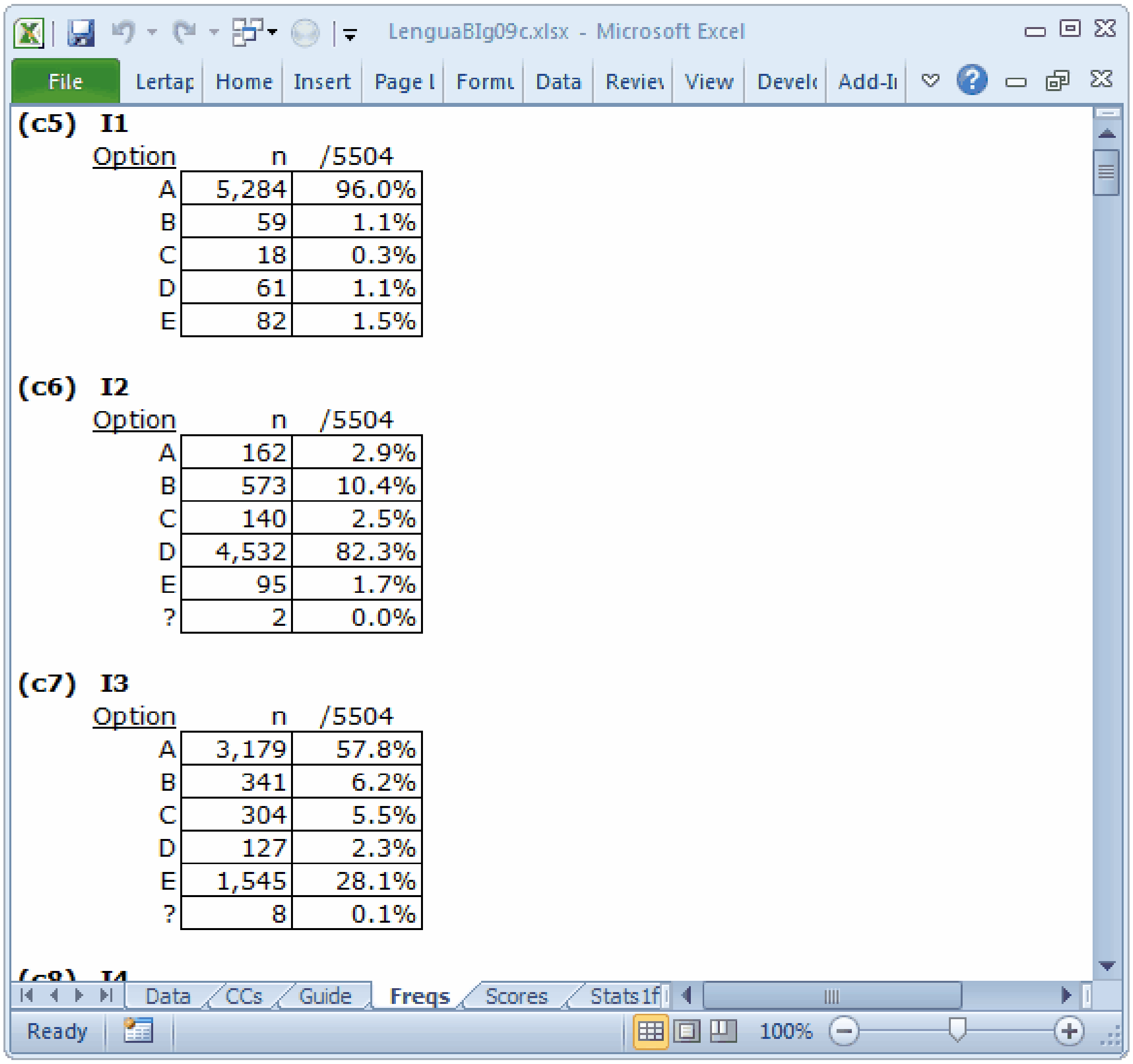
Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

2.12.1 Sample results 1

The 3 Steps resulted in a veritable swag of new "reports", or worksheets, going from Freqs to Scores to Stats1f to Stats1b, and so on, out to csem3, and Stats3ul, no less than 14 new, whopping worksheets to study. (Our computer, low on sugar, took about 4 minutes to complete the Steps.)

Were there any data processing errors? Freqs, what say you?


LenguaBIg09c.xlsx - Microsoft Excel

(c5) I1

Option	n	/5504
A	5,284	96.0%
B	59	1.1%
C	18	0.3%
D	61	1.1%
E	82	1.5%

(c6) I2

Option	n	/5504
A	162	2.9%
B	573	10.4%
C	140	2.5%
D	4,532	82.3%
E	95	1.7%
?	2	0.0%

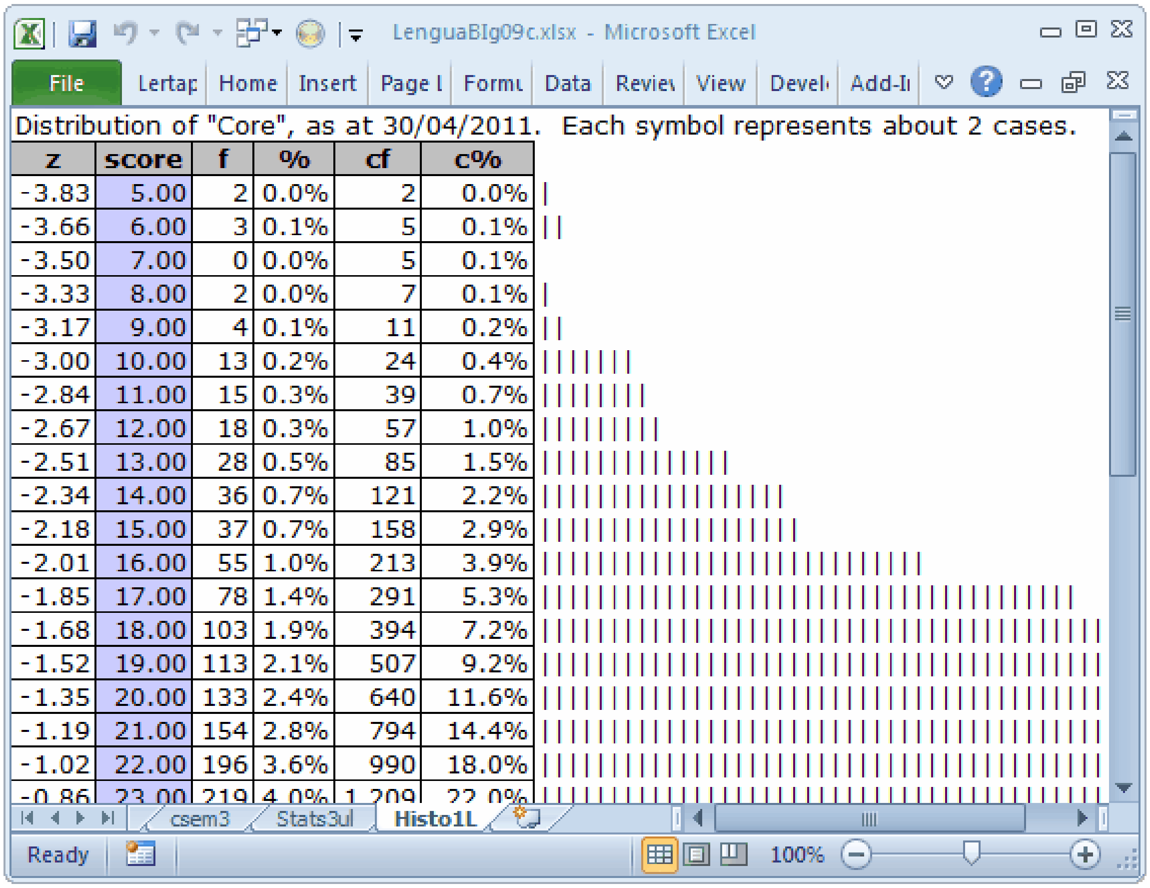
(c7) I3

Option	n	/5504
A	3,179	57.8%
B	341	6.2%
C	304	5.5%
D	127	2.3%
E	1,545	28.1%
?	8	0.1%

These results were prepared by submitting mark-sense answer sheets to a scanner. The scanner "wrote" a blank whenever a student did not answer a question, or whenever the student's answer was impossible for the scanner to decipher (this will happen when a student shades in more than one answer, for example). In the Freqs report, blanks are tallied in the ? row.

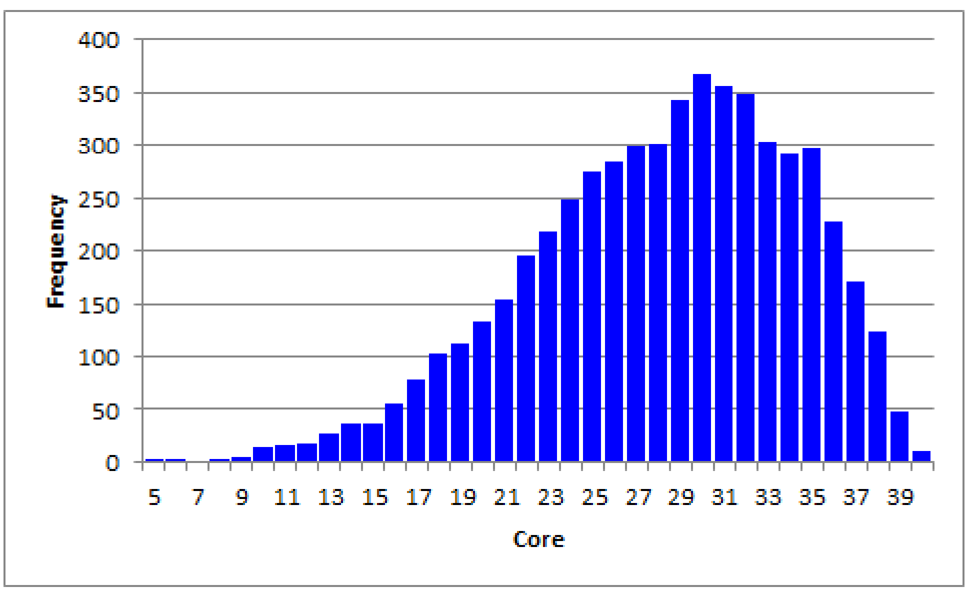
I38, not shown here, had a tally of 20 in the ? row, which is just 0.04% of the total number of students, 5504. Most items had far fewer tallies, as seen above for I1, I2, and I3. All okay.

What about the possibility of score outliers, very low scores which may indicate a bad record (row) in the Data worksheet? We'll focus on the Scores worksheet, and get a histogram for the first score, which is called "Core".



Distribution of "Core", as at 30/04/2011. Each symbol represents about 2 cases.

z	score	f	%	cf	c%	
-3.83	5.00	2	0.0%	2	0.0%	|
-3.66	6.00	3	0.1%	5	0.1%	||
-3.50	7.00	0	0.0%	5	0.1%	
-3.33	8.00	2	0.0%	7	0.1%	|
-3.17	9.00	4	0.1%	11	0.2%	||
-3.00	10.00	13	0.2%	24	0.4%	||||||
-2.84	11.00	15	0.3%	39	0.7%	|||||||
-2.67	12.00	18	0.3%	57	1.0%	|||||||||
-2.51	13.00	28	0.5%	85	1.5%	||||||||||||||
-2.34	14.00	36	0.7%	121	2.2%	||||||||||||||||||
-2.18	15.00	37	0.7%	158	2.9%	|||||||||||||||||||
-2.01	16.00	55	1.0%	213	3.9%	||||||||||||||||||||||||||||
-1.85	17.00	78	1.4%	291	5.3%	||||||||||||||||||||||||||||||||||||||||
-1.68	18.00	103	1.9%	394	7.2%	||||||||||||||||||||||||||||||||||||||||||
-1.52	19.00	113	2.1%	507	9.2%	||||||||||||||||||||||||||||||||||||||||||
-1.35	20.00	133	2.4%	640	11.6%	||||||||||||||||||||||||||||||||||||||||||
-1.19	21.00	154	2.8%	794	14.4%	||||||||||||||||||||||||||||||||||||||||||
-1.02	22.00	196	3.6%	990	18.0%	||||||||||||||||||||||||||||||||||||||||||
-0.86	23.00	219	4.0%	1,209	22.0%	||||||||||||||||||||||||||||||||||||||||||

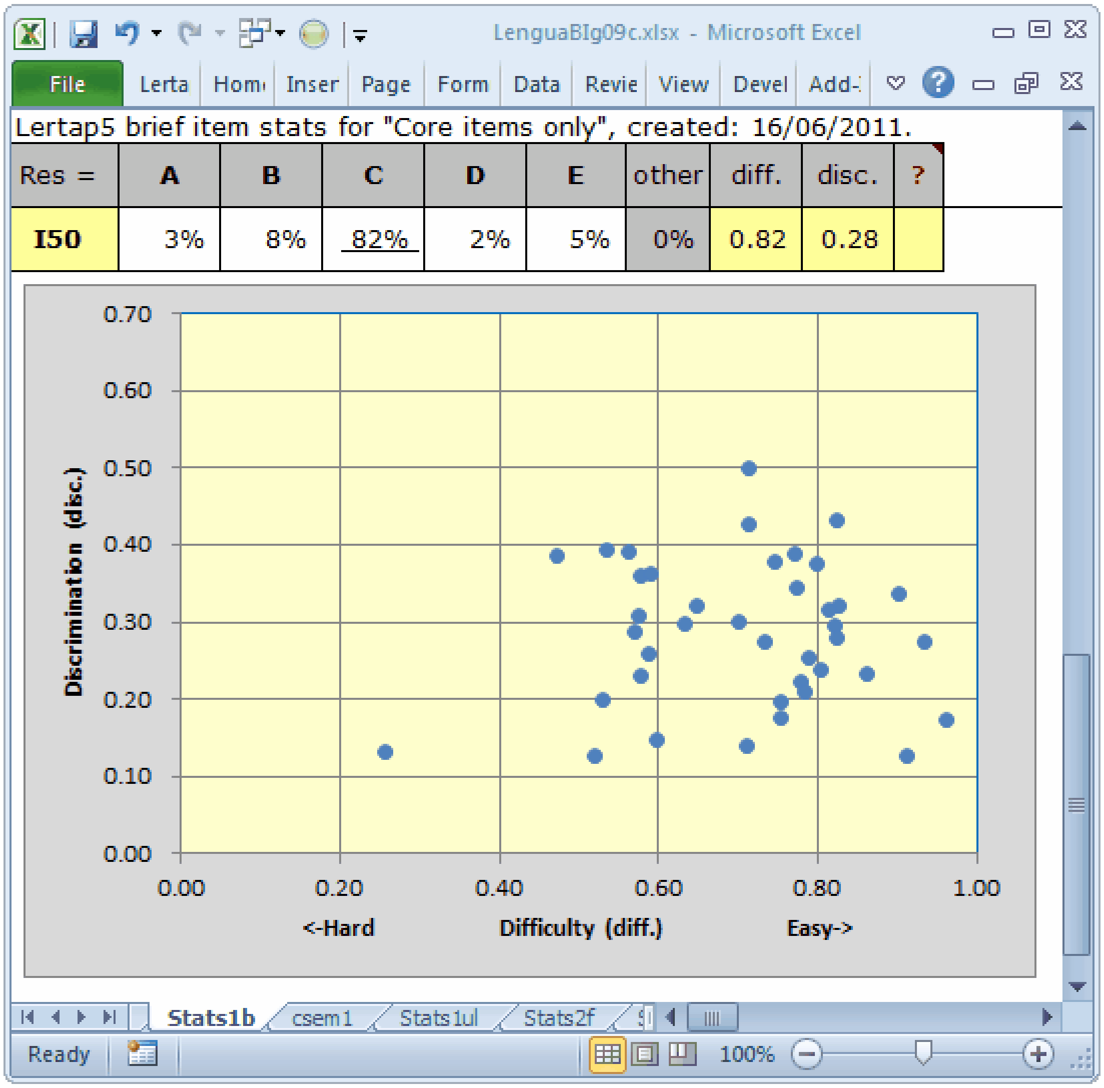


There do seem to be some curiously low scores. We've got three scores of 6.00, and two of 5.00. We should find out who they were, something which is not difficult to do. First, we made sure the **Scores** worksheet was in view, then went to the Data tab on the Excel ribbon, and selected Filter. We asked for a filter of 5.00 and 6.00 on column 2, where the "Core" scores are located. We made note of the IDs of the students who had these scores, went over to the Data worksheet, and used Find (on the Home tab) to look at their item responses. We found nothing unusual -- the students seemed to have taken a stab at each item. (In truth, we only looked at the Data records corresponding to a score of 5.00, and there is something unusual: the IDs are really birth dates, so it's possible to have two or more records with the same ID.)

Okay then, the data seem to be in order.

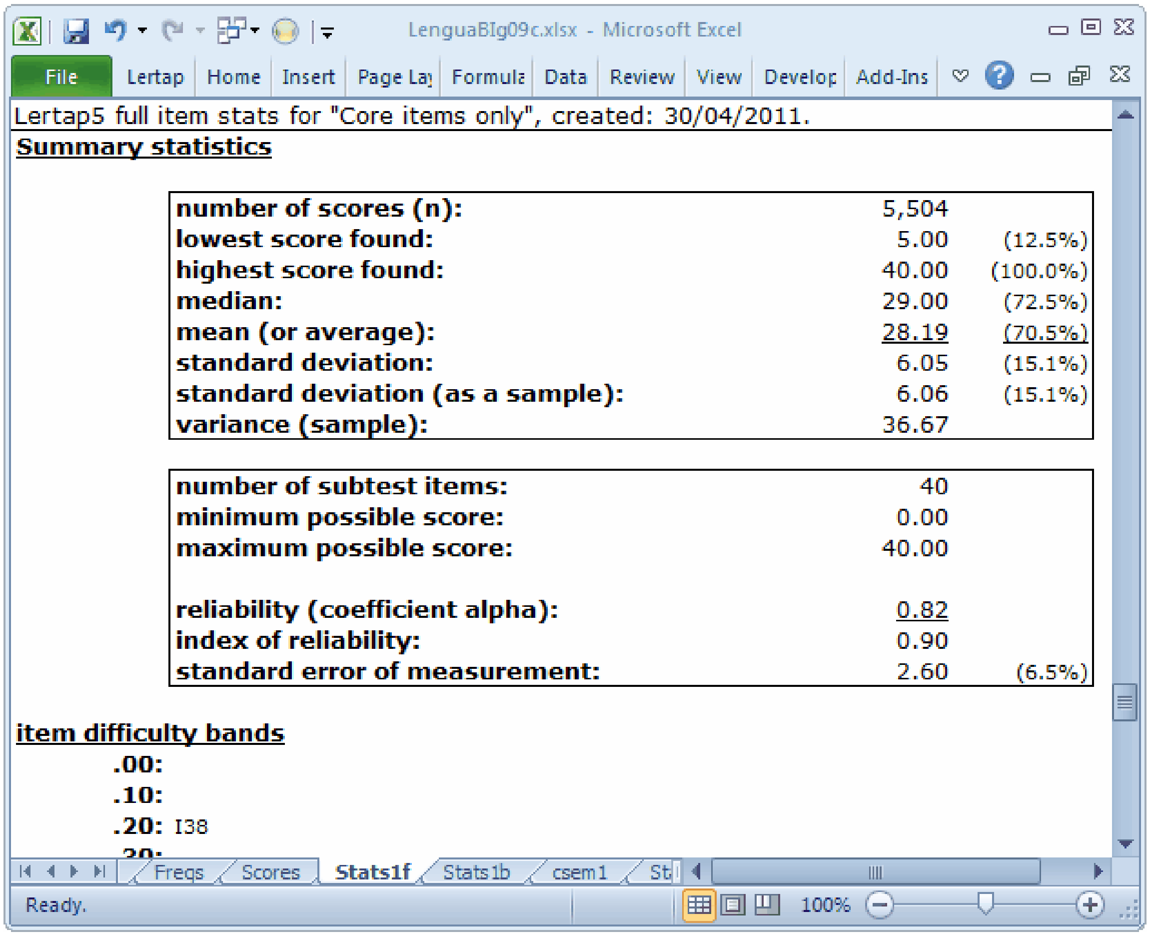
Our goal here is to look at the quality of the ten trial items. We'll do that in the next topic.

But first, let's just have a quick squiz of the results for the "Core" test, the 40 items which were scored.



There are quite a few easy-ish items, that is, questions whose diff. values are greater than 0.80 This isn't necessarily bad, especially if the test falls into a "mastery" sort of exam, which this one does: students had to have a score of at least 50% in order to pass (corresponding to a score of at least 20 on the 40 items). In this case, 4997 students passed, 91% of all students, something which is easy to determine by using the Histo1L "report".

The Stats1f report will give us some crucial data, such as coefficient alpha, the reliability figure computed by Lertap.



Lertap5 full item stats for "Core items only", created: 30/04/2011.

Summary statistics

number of scores (n):	5,504	
lowest score found:	5.00	(12.5%)
highest score found:	40.00	(100.0%)
median:	29.00	(72.5%)
mean (or average):	28.19	(70.5%)
standard deviation:	6.05	(15.1%)
standard deviation (as a sample):	6.06	(15.1%)
variance (sample):	36.67	

number of subtest items:	40	
minimum possible score:	0.00	
maximum possible score:	40.00	
reliability (coefficient alpha):	0.82	
index of reliability:	0.90	
standard error of measurement:	2.60	(6.5%)

item difficulty bands

.00:
.10:
.20: I38

A reliability of 0.82 might be acceptable; could have been better, maybe. What's missing are items with disc. values above 0.40. We've only got three (refer to the scatterplot above).

We could use the Spearman-Brown formula to see what the test's reliability would be if we added 20 items; then we could make some comparisons to the three 60-item tests discussed in the M.Nursing sample. We did this, and found we'd have a test with a reliability of 0.87 if we could add 20 items to the present test, items with similar statistical characteristics. That's better, but, this being a pass-fail test, a statistic which might be as important as reliability, if not more so, would be the "**Prop. consistent placings**" index. Finding what that might be is left as an exercise; download this dataset and see what you find.

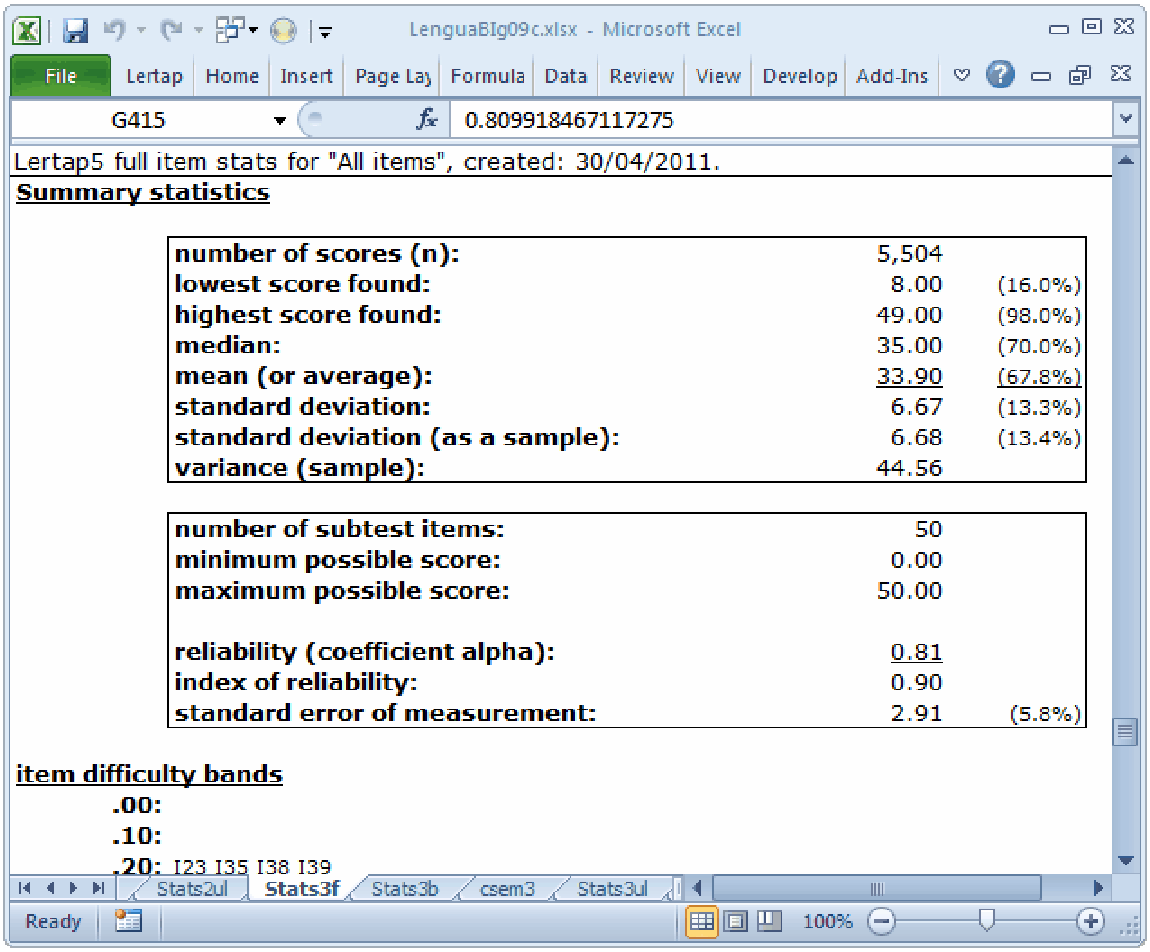
Let's move on, looking for an answer to our basic question: were the trial items any good? Advance!

2.12.2 Sample results 2

What about those ten trial items? Were they any good?

A simple way to answer this question would be to compare the reliability of the "Core" subtest and the "All" subtest. We already know the former from the previous topic: 0.82 is the reliability for the 40-item core subtest.

Here are the stats for the "All" subtest, with 50 items:



Lertap5 full item stats for "All items", created: 30/04/2011.

Summary statistics

number of scores (n):	5,504	
lowest score found:	8.00	(16.0%)
highest score found:	49.00	(98.0%)
median:	35.00	(70.0%)
mean (or average):	33.90	(67.8%)
standard deviation:	6.67	(13.3%)
standard deviation (as a sample):	6.68	(13.4%)
variance (sample):	44.56	

number of subtest items:	50	
minimum possible score:	0.00	
maximum possible score:	50.00	
reliability (coefficient alpha):	0.81	
index of reliability:	0.90	
standard error of measurement:	2.91	(5.8%)

item difficulty bands

.00:

.10:

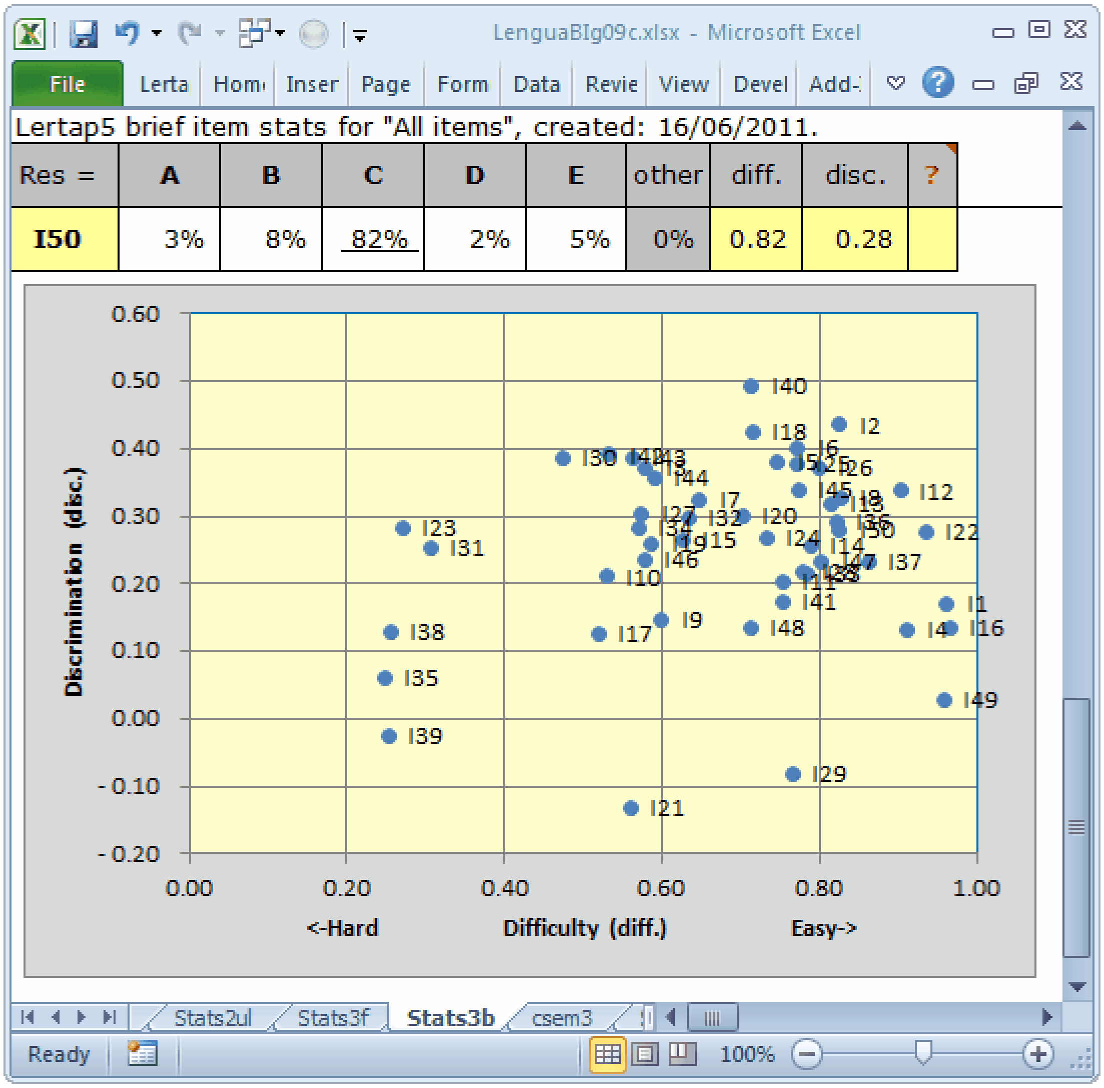
.20: I23 I35 I38 I39

Stats2ul | Stats3f | Stats3b | csem3 | Stats3ul

The reliability has gone down a bit, to 0.81. Still sort of okay, you say? Not really. Adding items to a test should increase reliability.

Here is another chance to benefit from application of the Spearman-Brown formula. Adding 10 items to our 40-item test should, by Spearman-Brown, take the reliability to 0.85; we'd truly expect this to happen if we were adding items whose quality was equal to that of the original items. Obviously, our ten new items are either of lower quality, or are just not testing the same concepts as the original items.

Here's that nifty scatterplot of item diff. by disc. for the 50-item subtest, "All":



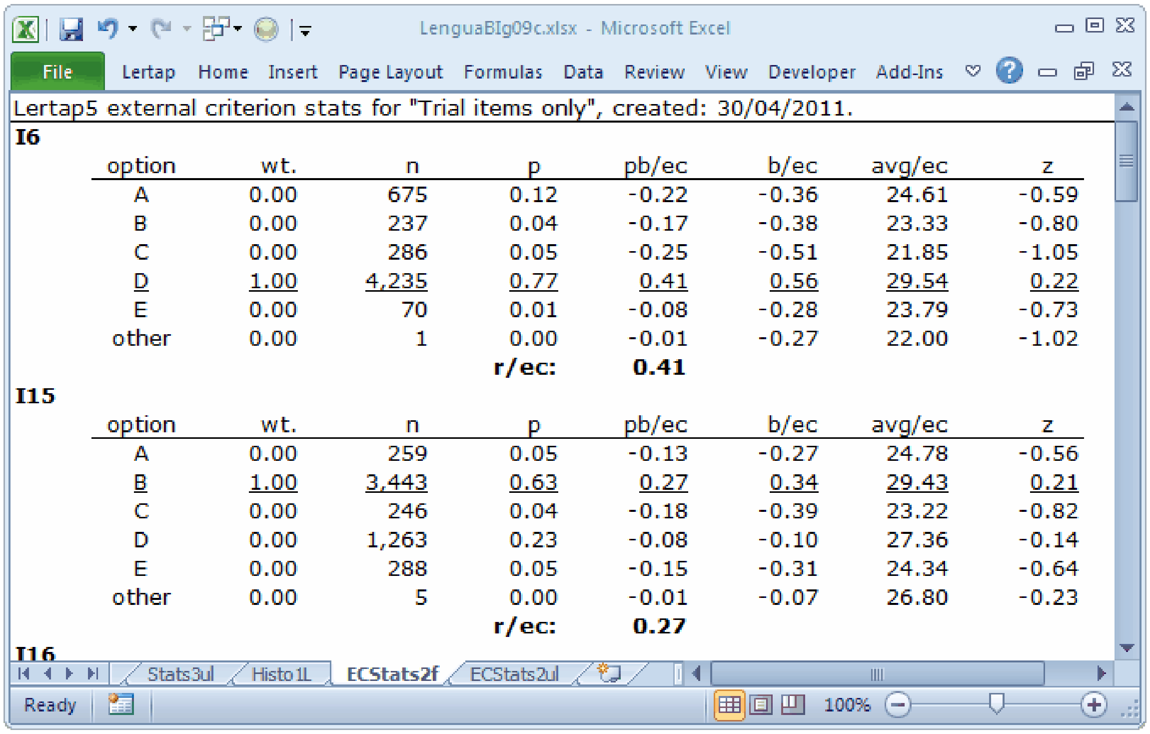
We didn't have any items with a disc. value less than 0.10 before. Now we have five. Three of these five are negative, always a bad sign. (Note: Lertap always shows item names next to their blips in the scatterplot, as shown here. To delete the names, right-click on just one of them, such as I39, and choose Delete.)

Obviously, things are not going well for our ten trial items.

We will get more detail by going through an external criterion analysis, an option available on the Run menu's More option, just under the Elmillon option.

We selected the "Core" score as the external criterion, and then the second subtest, "Trial", as the one having the items to be correlated with the external criterion.

This added another two reports, ECStats2f, and ECStats2ul.



Lertap5 external criterion stats for "Trial items only", created: 30/04/2011.

I6

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
A	0.00	675	0.12	-0.22	-0.36	24.61	-0.59
B	0.00	237	0.04	-0.17	-0.38	23.33	-0.80
C	0.00	286	0.05	-0.25	-0.51	21.85	-1.05
D	1.00	4,235	0.77	0.41	0.56	29.54	0.22
E	0.00	70	0.01	-0.08	-0.28	23.79	-0.73
other	0.00	1	0.00	-0.01	-0.27	22.00	-1.02
			r/ec:	**0.41**			

I15

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
A	0.00	259	0.05	-0.13	-0.27	24.78	-0.56
B	1.00	3,443	0.63	0.27	0.34	29.43	0.21
C	0.00	246	0.04	-0.18	-0.39	23.22	-0.82
D	0.00	1,263	0.23	-0.08	-0.10	27.36	-0.14
E	0.00	288	0.05	-0.15	-0.31	24.34	-0.64
other	0.00	5	0.00	-0.01	-0.07	26.80	-0.23
			r/ec:	**0.27**			

I16

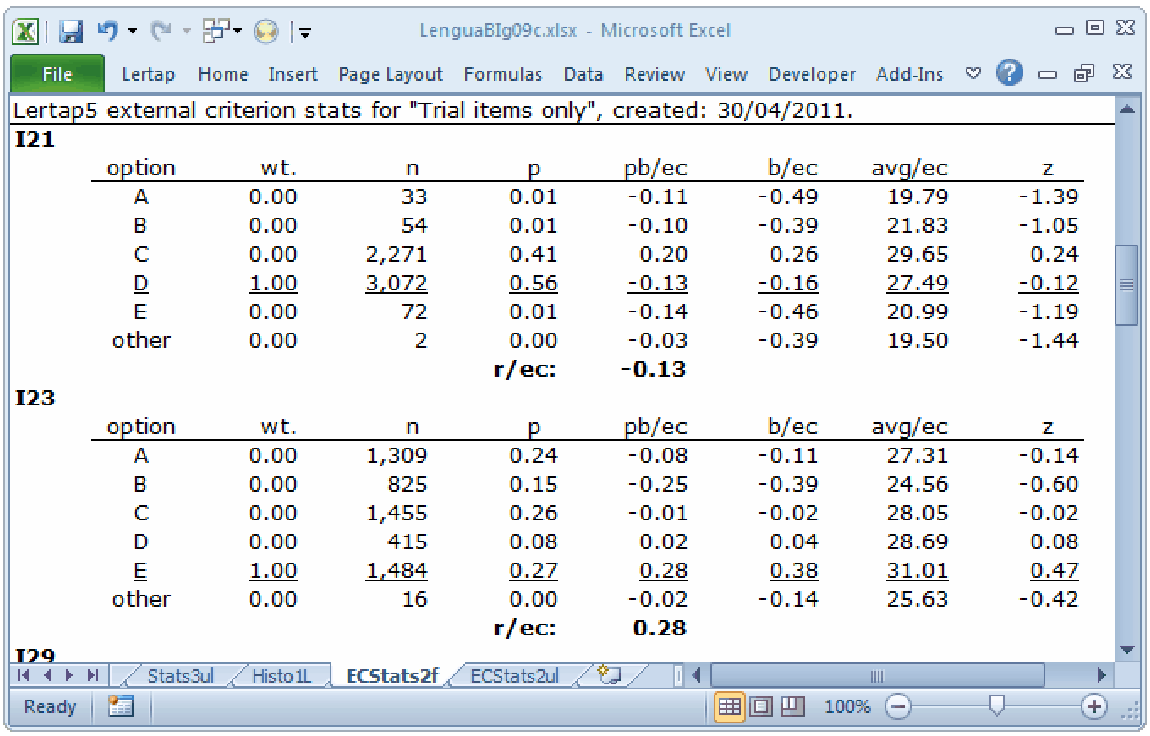
The ECStats2f report is not quite the same as an ordinary Statsf report. The criterion score used for the four statistics columns, pb/ec, b/ec, avg/ec, and z are all based on the external criterion score, which, for this analysis, is "Core", the score on the 40 core items. But what we want to see is the same sort of pattern we ordinarily look for in a discriminating item: the highest avg/ec value should be found for the keyed-correct answer.

Avg/ec is the average score on the external criterion, "Core", found for those students who selected the corresponding option. For example, on I6 the keyed-correct answer is D, underlined to remind us that it's the right answer. The average "Core" score for the 4,235 students who chose this option is 29.54. Meanwhile, the average "Core" scores on I6 for those selecting any of the four distractors are below the over-all "Core" average, which was 28.19. Note: the z column makes it easy to see if avg/ec is above or below average: above and z is positive; below and z is negative.

The **r/ec** figure found in ECStatsf reports is the correlation between the item and the external criterion. In technical terms, it's not a point-biserial (pb) correlation coefficient, nor is it a biserial (b) coeffcient. It's what's known in the business as the Pearson product-moment correlation. If an item is scored on a dichotomous basis, right-wrong, then r/ec will be the same as the pb/ec value for the item's correct answer.

What do we want to see for our ten trial items, as far as r/ec goes? Well, first of all, a positive value. Secondly, a large positive value. What's large (does size matter)? Preferably above 0.30. If hard-pressed, we might accept anything at 0.20 and above. If at or above 0.40 we'll pin a gold star on the item.

I6 and I15 performed well and well-enough, respectively. I6's correlation with the criterion, **r/ec**, is 0.41, a gold-star outcome; browsing down the **avg/ec** and **z** columns shows a good pattern for I6 (and I15 for that matter), with a positive z for the correct answer (the "key") of D (or B, in the case of I15). These two items would be "keepers"; adding them to them to the core test would improve reliability (you could try it yourself instead of just sitting there, reading; you know what a lack of exercise will do).



Lertap5 external criterion stats for "Trial items only", created: 30/04/2011.

I21

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
A	0.00	33	0.01	-0.11	-0.49	19.79	-1.39
B	0.00	54	0.01	-0.10	-0.39	21.83	-1.05
C	0.00	2,271	0.41	0.20	0.26	29.65	0.24
D	1.00	3,072	0.56	-0.13	-0.16	27.49	-0.12
E	0.00	72	0.01	-0.14	-0.46	20.99	-1.19
other	0.00	2	0.00	-0.03	-0.39	19.50	-1.44
			r/ec:	**-0.13**			

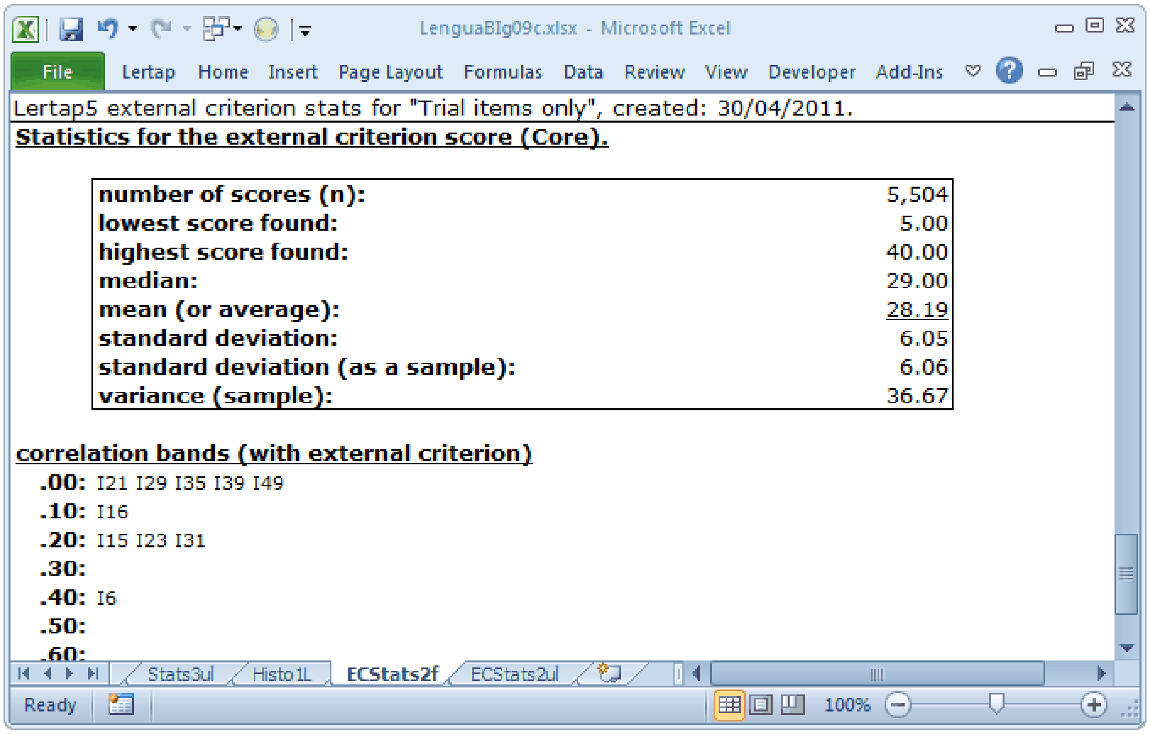
I23

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
A	0.00	1,309	0.24	-0.08	-0.11	27.31	-0.14
B	0.00	825	0.15	-0.25	-0.39	24.56	-0.60
C	0.00	1,455	0.26	-0.01	-0.02	28.05	-0.02
D	0.00	415	0.08	0.02	0.04	28.69	0.08
E	1.00	1,484	0.27	0.28	0.38	31.01	0.47
other	0.00	16	0.00	-0.02	-0.14	25.63	-0.42
			r/ec:	**0.28**			

I23 might be another keeper, but probably not I21. The z for I21's correct answer is negative. The average "Core" score for the 3,072 students who selected option D on I21, the keyed-correct answer, is below the overall "Core" average of 28.19.

The following little table indicates that only four of the ten items show promise: I6, I15, I23, and I31. This is seen in the correlation bands (with external criterion) table. This table is based on the items' r/ec values. We have five items, I21, I29, I35, I39, and I49, in the "sin bin", with r/ec values less than 0.10. In fact, of these five, items

I21, I29, and I39 have negative r/ec's. (You need a copy of the whole ECStats2f report to confirm this, and, yes you can get one -- download the dataset.)

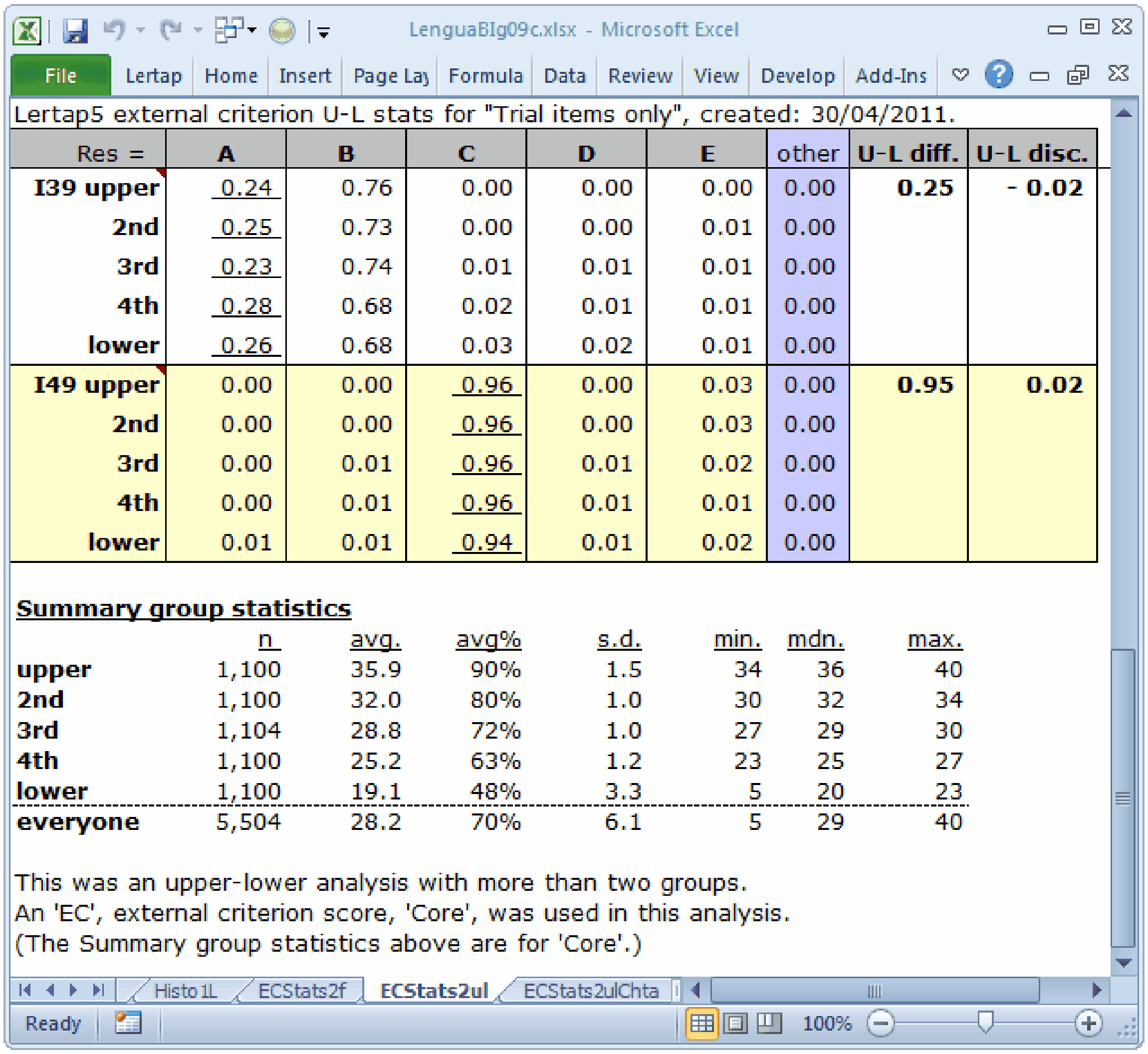

Lertap5 external criterion stats for "Trial items only", created: 30/04/2011.

Statistics for the external criterion score (Core).

number of scores (n):	5,504
lowest score found:	5.00
highest score found:	40.00
median:	29.00
mean (or average):	28.19
standard deviation:	6.05
standard deviation (as a sample):	6.06
variance (sample):	36.67

correlation bands (with external criterion)

.00: I21 I29 I35 I39 I49
.10: I16
.20: I15 I23 I31
.30:
.40: I6
.50:
.60:



LenguaBIg09c.xlsx - Microsoft Excel

File | Lertap | Home | Insert | Page Lay | Formula | Data | Review | View | Develop | Add-Ins

Lertap5 external criterion U-L stats for "Trial items only", created: 30/04/2011.

Res =	A	B	C	D	E	other	U-L diff.	U-L disc.
I39 upper	0.24	0.76	0.00	0.00	0.00	0.00	**0.25**	**- 0.02**
2nd	0.25	0.73	0.00	0.00	0.01	0.00		
3rd	0.23	0.74	0.01	0.01	0.01	0.00		
4th	0.28	0.68	0.02	0.01	0.01	0.00		
lower	0.26	0.68	0.03	0.02	0.01	0.00		
I49 upper	0.00	0.00	0.96	0.00	0.03	0.00	**0.95**	**0.02**
2nd	0.00	0.00	0.96	0.00	0.03	0.00		
3rd	0.00	0.01	0.96	0.01	0.02	0.00		
4th	0.00	0.01	0.96	0.01	0.01	0.00		
lower	0.01	0.01	0.94	0.01	0.02	0.00		

Summary group statistics

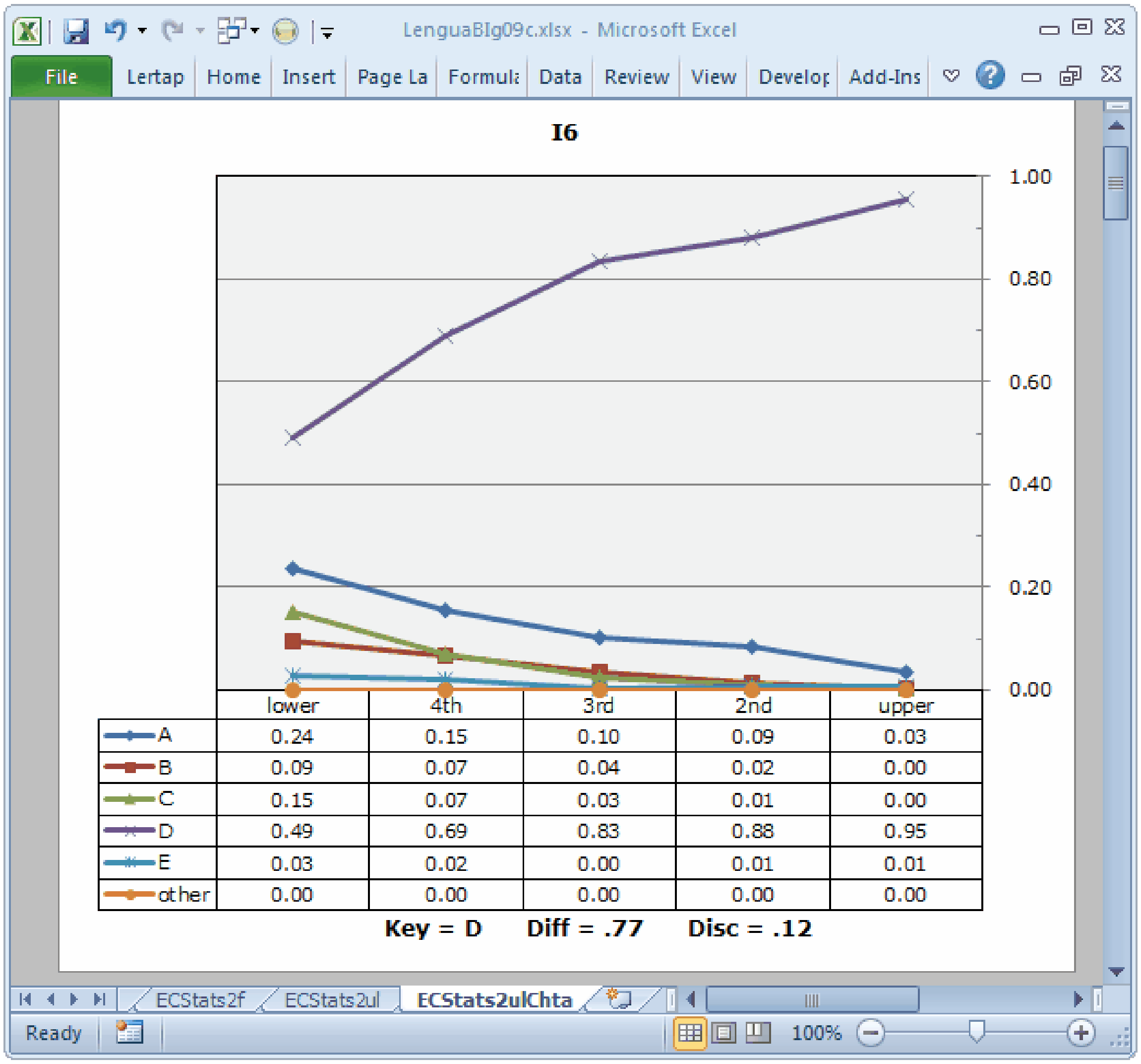
	n	avg.	avg%	s.d.	min.	mdn.	max.
upper	1,100	35.9	90%	1.5	34	36	40
2nd	1,100	32.0	80%	1.0	30	32	34
3rd	1,104	28.8	72%	1.0	27	29	30
4th	1,100	25.2	63%	1.2	23	25	27
lower	1,100	19.1	48%	3.3	5	20	23
everyone	5,504	28.2	70%	6.1	5	29	40

This was an upper-lower analysis with more than two groups.
An 'EC', external criterion score, 'Core', was used in this analysis.
(The Summary group statistics above are for 'Core'.)

Histo1L | ECStats2f | ECStats2ul | ECStats2ulChta

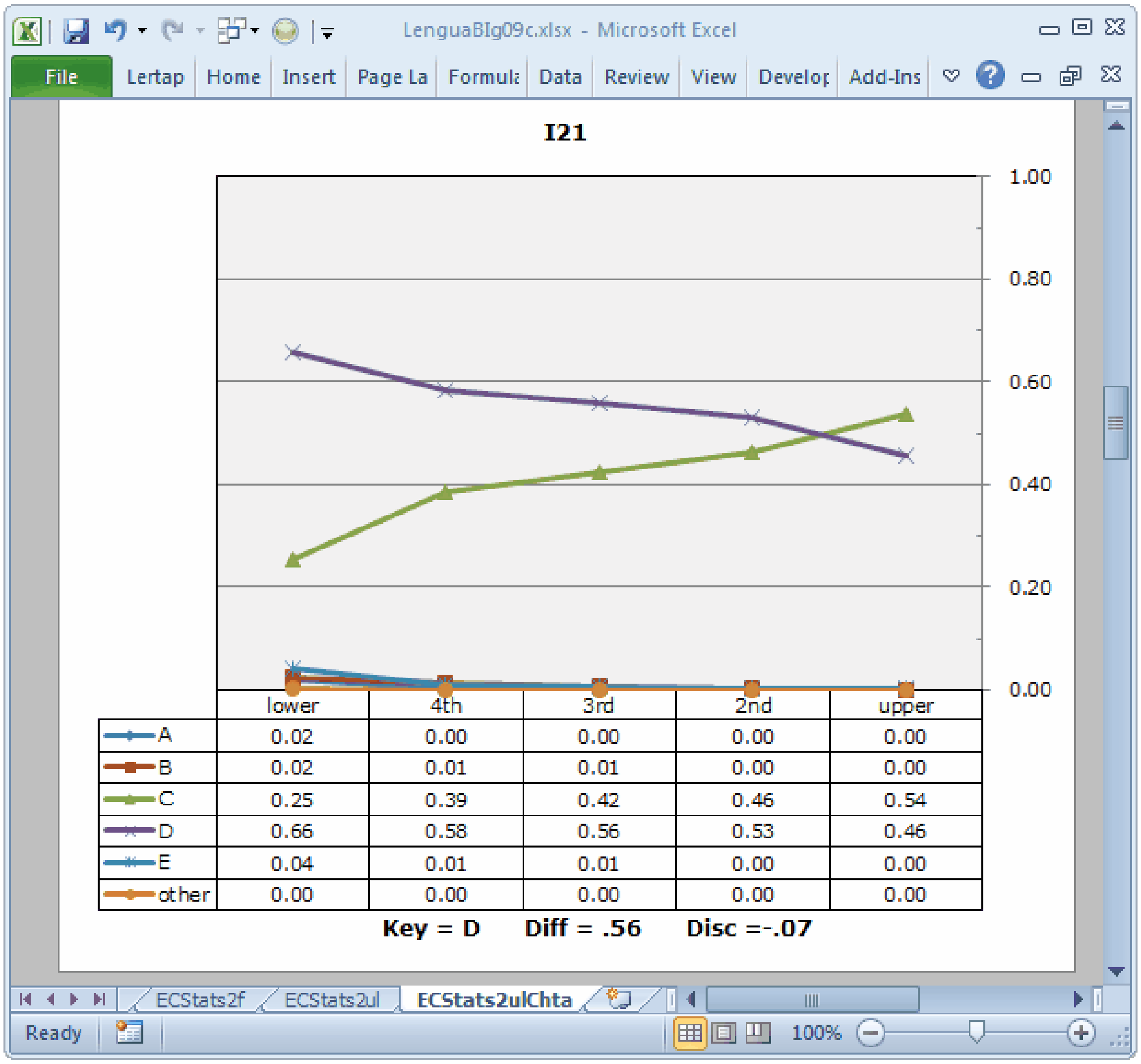
Ready 100%

The bottom of the ECStats2ul is that of an ordinary Statsul report, with an important distinction: the five quintile groups have been formed on the basis of the external criterion score. This score also serves as the foundation for the **quintile plots** which may be made from an ECStatsul report. (See how.)



	lower	4th	3rd	2nd	upper
A	0.24	0.15	0.10	0.09	0.03
B	0.09	0.07	0.04	0.02	0.00
C	0.15	0.07	0.03	0.01	0.00
D	0.49	0.69	0.83	0.88	0.95
E	0.03	0.02	0.00	0.01	0.01
other	0.00	0.00	0.00	0.00	0.00

Key = D Diff = .77 Disc = .12



Note: the Diff and Disc values in these plots are taken from the subtest's Statsb report. In this case, the subtest is the 10-item "Trial" subtest, and the corresponding report is Stats2b. These values are not very meaningful here; we have no real interest in that subtest on its own. We can ignore the Diff and Disc values, just taking in the "picture".

The plot for I6 is not bad. The trace of the correct option's response frequencies rises from left to right. This is the desired pattern for an item meant to discriminate: the strongest students, those in the "upper" group, are more likely to get the item right. At the same time, the popularity of the four distractors is greater on the left, falling off as we move to the right so that, by the time we get to the "upper" group almost no-one has been distracted by a distractor.

Not so for I21. Only two options, C and D, caught the attention of the students. Of these, D is the keyed-correct answer, but its trace line resembles that expected of a distractor, dropping from left to right when we'd expect it to do the opposite. Option C, a distractor, gains in popularity, or "endorsement", as we move from left to right, ending up as the most-selected option in the "upper" group. This is an undesired pattern for a distractor.

It could be that I21 has been mis-keyed, as was the case for I11 in our MathsQuiz sample. However, the signs are still not too promising; even should this have been the case, there would be too many people in the "upper" group selecting a distractor (which is what option D would become if we changed the keyed-correct answer to C).

Well, it seems that this palavering has stunned today's audience. People are texting on their phones, and some of them are good at it, too, looking ahead as though they're paying attention while, out of vision, their thumbs are busy. Some are taking a nap. No-one has asked a question. No-one even *looks* like they're about to as a question.

Ah, but wait! That principal from Harvard senior high school, Harriet, is alert. She says: "*Larry, this is all simply spectacular. I've had my laptop going all this time. I've downloaded this dataset, and will shortly work through it on my own. My goodness, Lertap is a marvel! Wait until I tell Ozzie about it tonight before he falls asleep! But, really, is this external-criterion analysis really required? Couldn't I just get into the various reports for the third subtest,* 'All'*, and wouldn't I be able to see how the ten items are working?*"

A good point. Indeed, scrolling this topic up to once again take in the "nifty scatterplot of item diff. by disc. for the 50-item subtest" shows our sin-bin quintet, I21, I29, I35, I39, and I49, at the bottom.

To follow up on Harriet's point, what we should do is make a little table. It'll have ten rows, one for each of the trial items, and two columns. The first column will have the r/ec values from the ECStats2f report; the second will have the disc. values from the Stats3b report, which, incidentally, are the same as the pb(r) values found in the Stats3f report. It's an interesting little exercise. You can do it yourself if you download this dataset.

You'll find that we should pin a gold star on Harriet; a comparison of the two columns will reveal only three differences, and they're very slight differences.

The ratio of core items to trial items in this sample is 40 to 10, 4-to-1. Should this ratio be lower, say 3-to-1, we might well expect our little table to show more differences. An external-criterion analysis would be the most advisable way to go about this business of checking the performance of trial items. However, in this case, it didn't make all that much difference; we would have concluded the same thing had we just run with two subtests, the 40-item core items, and "All", the 50-item subtest.

What conclusion? Four of the ten items, I6, I15, I23, and I31 are worthy of keeping, and might be included in a future version of the test.

More?

2.12.3 Download

The LenguaBIg09c.xlsx workbook may be downloaded by clicking here.

As to possible exercises to put to you, there being no baseball or cricket on TV today to watch, well, here ya'go:

On the 40-item "Core" subtest, how many items ended up with an entry in Stats1b's ? column? You'll find that there weren't many at all. Should any of the items have been double-keyed? If so, how would test reliability have been affected? What do the quintile plots look like? (Most of them are actually fairly good.)

We found four of the trial items to be worth a bit of salt, I6, I15, I23, and I31. What would the reliability of a subtest formed from the 40 core items, plus these four new ones, be? (According to Spearman-Brown, it could be around 0.83.)

You could put mastery=50 on the first *sub line to check on the "**Prop. consistent placings**" for the core items. You could also put it on the third *sub line, and then compare results: does adding the ten trial items have any impact on "**Prop. consistent placings**"? What about just adding our four promising ones, I6, I15, I23, and I31?

There are two columns in the Data worksheet, "Reg." (district, or region), and "Gender", which can be used to get score and item **breakouts** of the sort exemplified in the M.Nursing sample. However, were they to be used in breakouts, the Reg. codes would have to be recoded first. This is because they are numeric; see the caveat at the top of this topic for help and further comments -- it's easy to fix this limitation.

Being aware of the exact source of this dataset (a secret), and the importance of the test to the students involved, we can suggest that a check for **cheating** would likely uncover some cases of inappropriate behavior. If this interests you, look at the Negocios sample to get an idea of what's what, that is, how to go about checking for cheating. An RSA analysis (response-similarity analysis, Lertap's method for cheat checking) should involve only students who may have had a chance to cheat; ordinarily, this will mean that the student records involved in an RSA analysis will be from a single test venue. However, we don't have the test venue codes for this dataset --the best you could do would be to limit the RSA analysis to a single school district, using the codes found in the Data worksheet's "Reg." column. For comments on how to make a dataset which includes only selected records, read here.

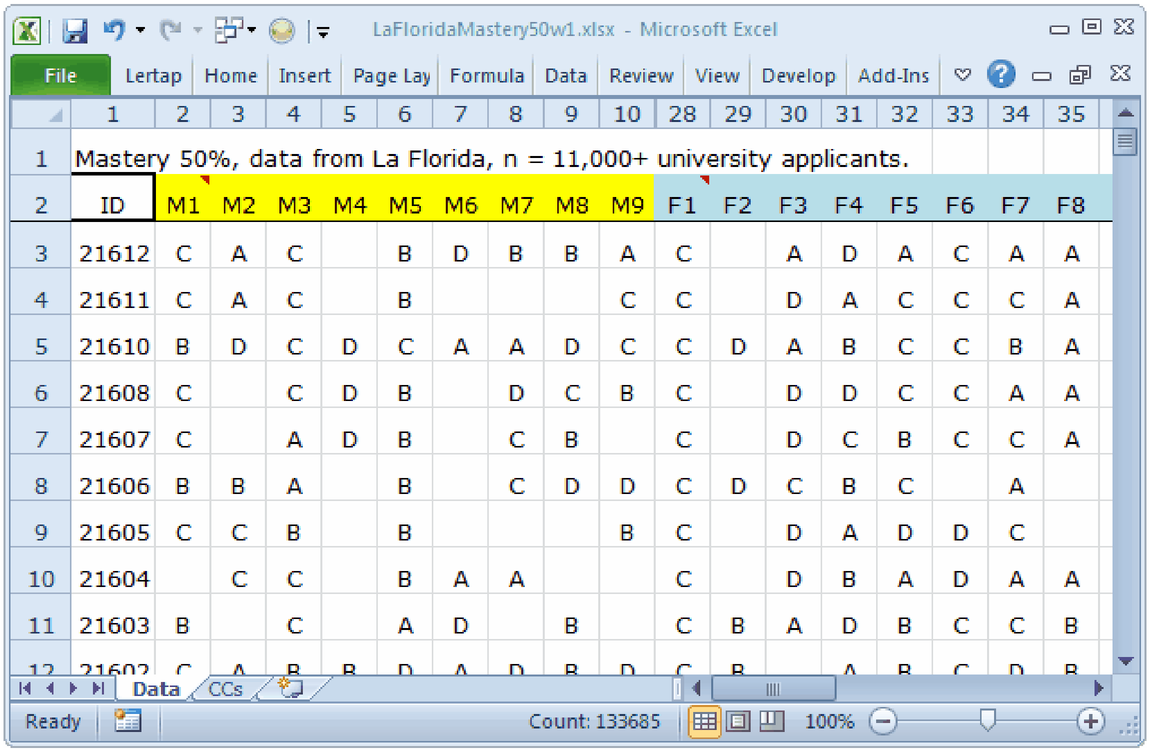
Or, you might just put your walking shoes on, and let them take you out for a stroll in the park (*watch out for Lertap hawkers*).

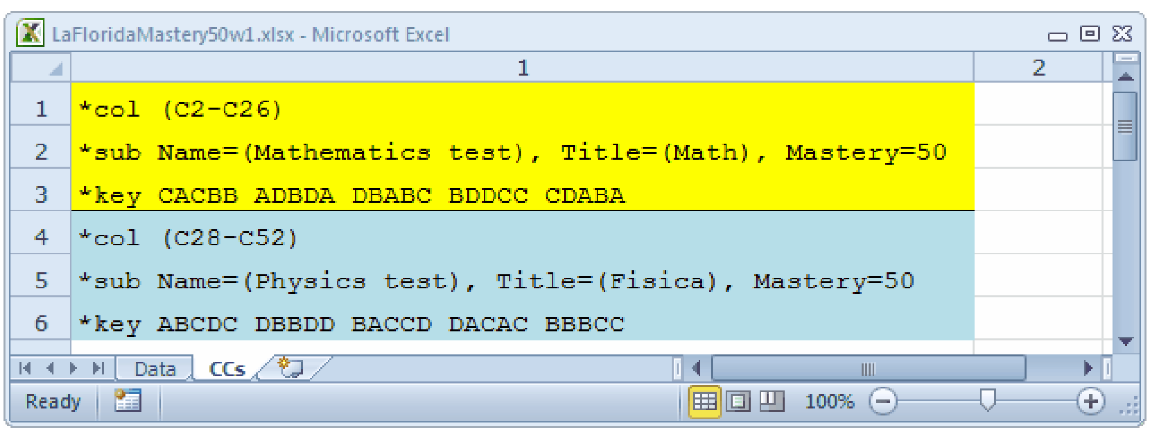
2.13 LaFlorida

Here's a 50-item screening test, used by a large South American engineering faculty to assist with the task of picking the strongest applicants from an over-abundance of erstwhile, would-be freshman students.

More than 11,000 applications were received by a faculty with space for 500 first-year students.

As part of the screening process, all applicants were required to sit a test intentionally designed to be difficult, even perhaps a bit intimidating. It started with 25 mathematics items, followed by 25 physics items ("fisica"). Students had to have a score of at least 50% to pass the test.





Note that columns 11 through 27 of the Data worksheet have been hidden in the screen snapshot above. This was to make the start of the physics items clearer: they begin in the 28th column.

Each *col line in the CCs sheet marks the definition of a subtest. This CCs worksheet defines two subtests. The first has item responses starting in column 2 (C2) of the Data worksheet, going out through C26. The item responses for the second subtest, "Fisica", begin in C28, extending out to C52.

How did we get results? You mean you don't know by now! We took a few steps.

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

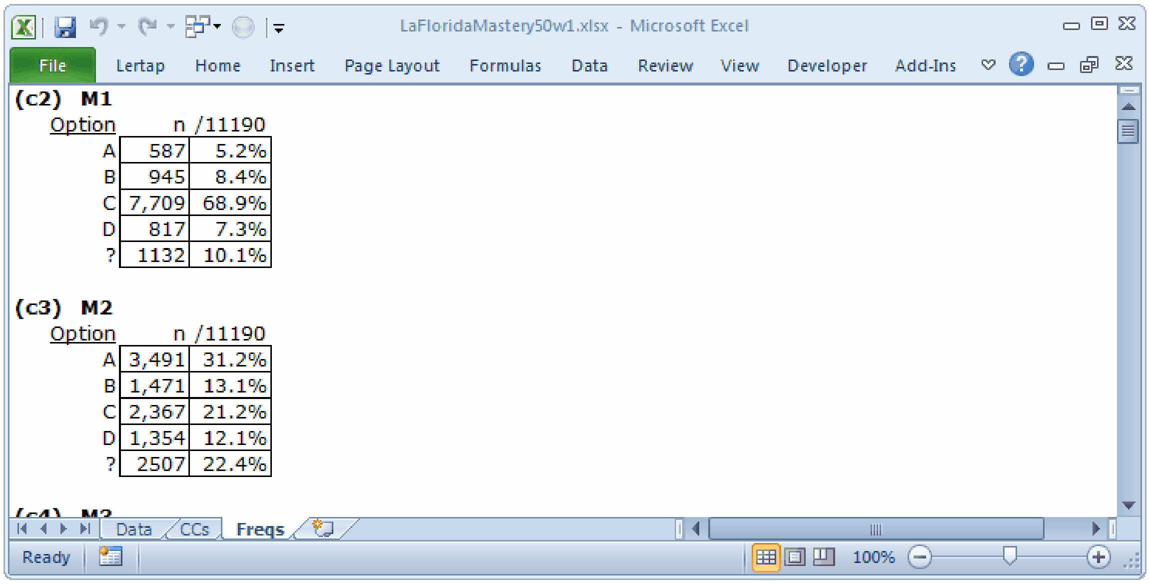
3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

It took our MacBook Pro running Windows 7 and Excel 2010 four and a half minutes to work through Lertap's **Interpret** stage. When there's more than one subtest (here we have two), Interpret may pause after it has run through the items belonging to a subtest if it finds "strange" characters for item responses. It did it with La Florida. A scanner was used to process the item responses from mark-sense sheets; unanswered questions were "recorded" as blanks in the scanner's data file, while uninterpretable, "strange", answers were recorded as asterisks. (A typical uninterpretable response occurs when students shade in more than one bubble for an item on their answer sheets, effectively giving more than one answer to the item.)



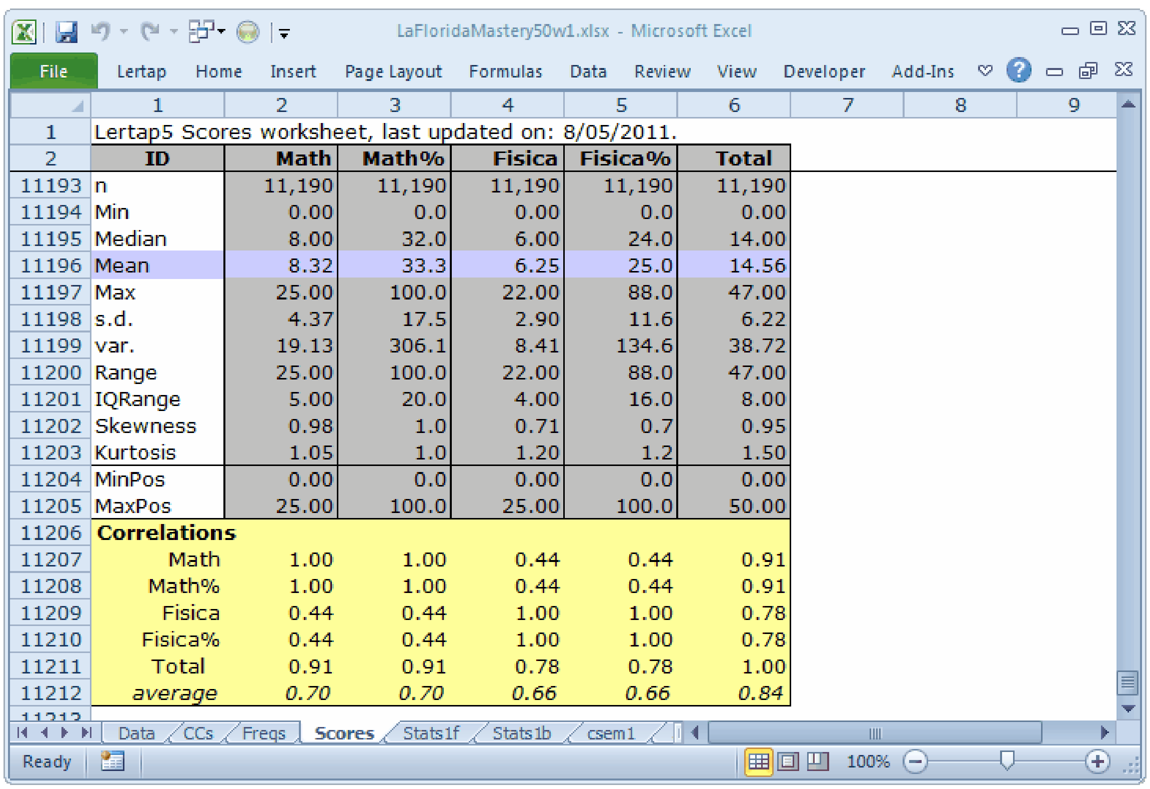
(c2) M1

Option	n	/11190
A	587	5.2%
B	945	8.4%
C	7,709	68.9%
D	817	7.3%
?	1132	10.1%

(c3) M2

Option	n	/11190
A	3,491	31.2%
B	1,471	13.1%
C	2,367	21.2%
D	1,354	12.1%
?	2507	22.4%

The ? row in a Freqs table includes both "strange" responses and blanks. All of the items in both subtests had large entries in the ? row. See for yourself: download this workbook, and have a go.

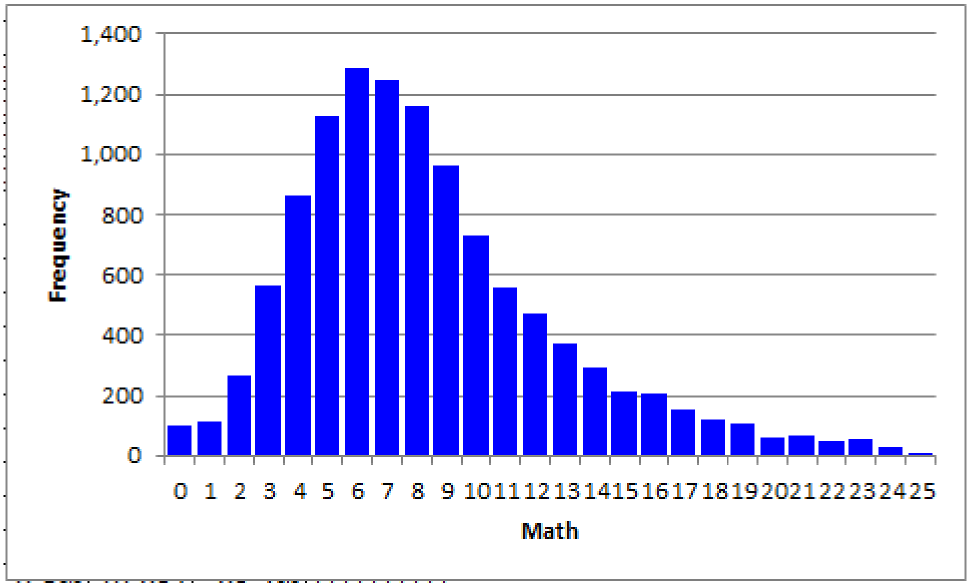


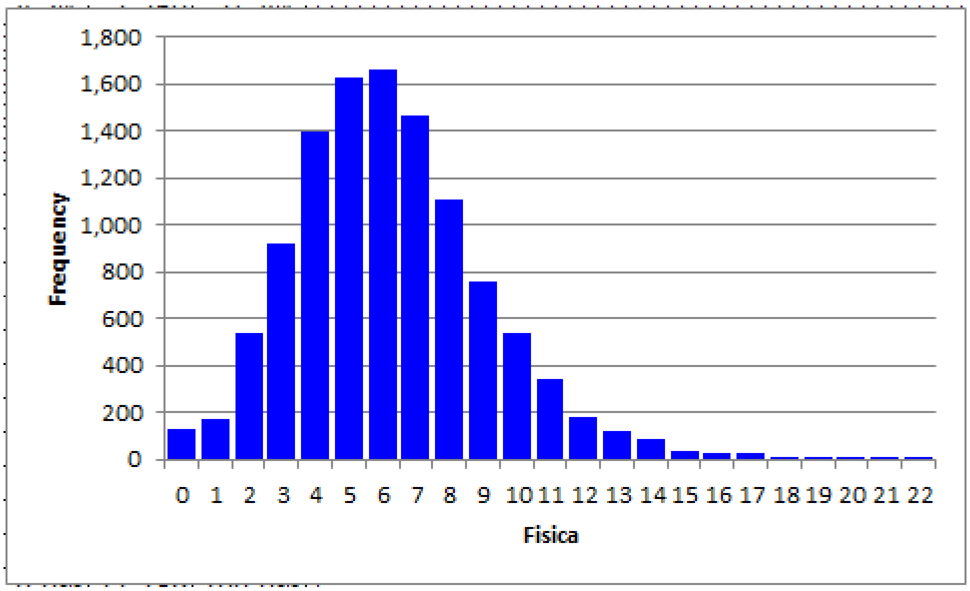
Lertap5 Scores worksheet, last updated on: 8/05/2011.

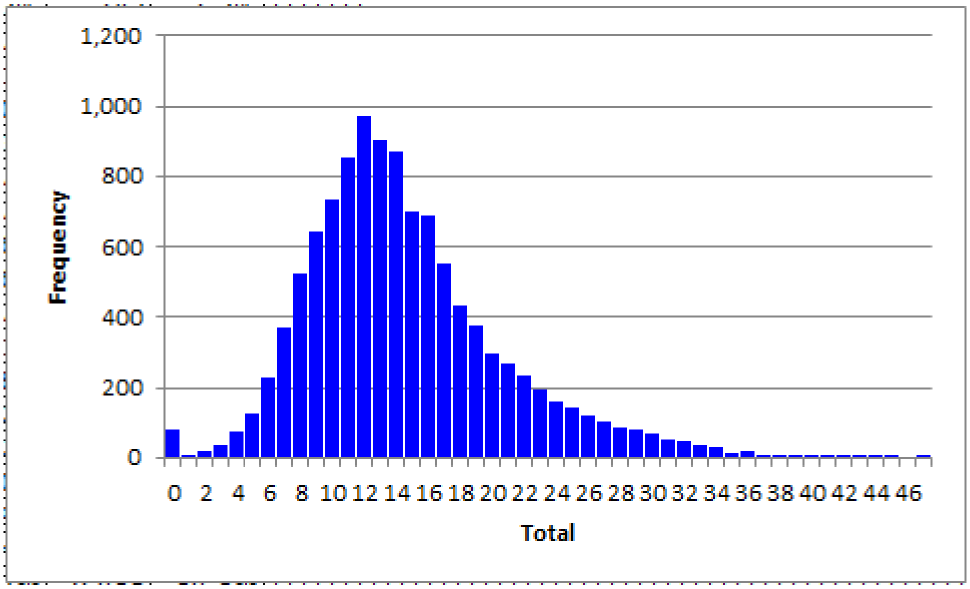
	ID	Math	Math%	Fisica	Fisica%	Total
11193	n	11,190	11,190	11,190	11,190	11,190
11194	Min	0.00	0.0	0.00	0.0	0.00
11195	Median	8.00	32.0	6.00	24.0	14.00
11196	Mean	8.32	33.3	6.25	25.0	14.56
11197	Max	25.00	100.0	22.00	88.0	47.00
11198	s.d.	4.37	17.5	2.90	11.6	6.22
11199	var.	19.13	306.1	8.41	134.6	38.72
11200	Range	25.00	100.0	22.00	88.0	47.00
11201	IQRange	5.00	20.0	4.00	16.0	8.00
11202	Skewness	0.98	1.0	0.71	0.7	0.95
11203	Kurtosis	1.05	1.0	1.20	1.2	1.50
11204	MinPos	0.00	0.0	0.00	0.0	0.00
11205	MaxPos	25.00	100.0	25.00	100.0	50.00
11206	**Correlations**					
11207	Math	1.00	1.00	0.44	0.44	0.91
11208	Math%	1.00	1.00	0.44	0.44	0.91
11209	Fisica	0.44	0.44	1.00	1.00	0.78
11210	Fisica%	0.44	0.44	1.00	1.00	0.78
11211	Total	0.91	0.91	0.78	0.78	1.00
11212	*average*	*0.70*	*0.70*	*0.66*	*0.66*	*0.84*

We told you these were difficult tests. The median score for Math was just 8.00, or 32.0%. It was even less for Fisica.

The median total score, Math + Fisica, was 14.00, just 28% of the maximum possible of 50.00.





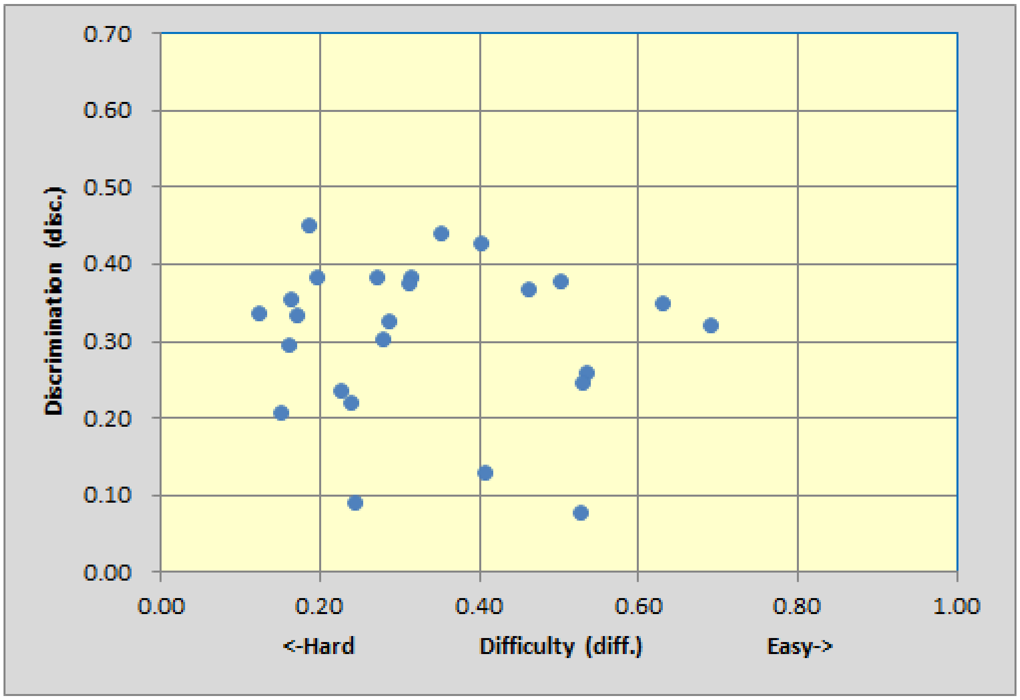


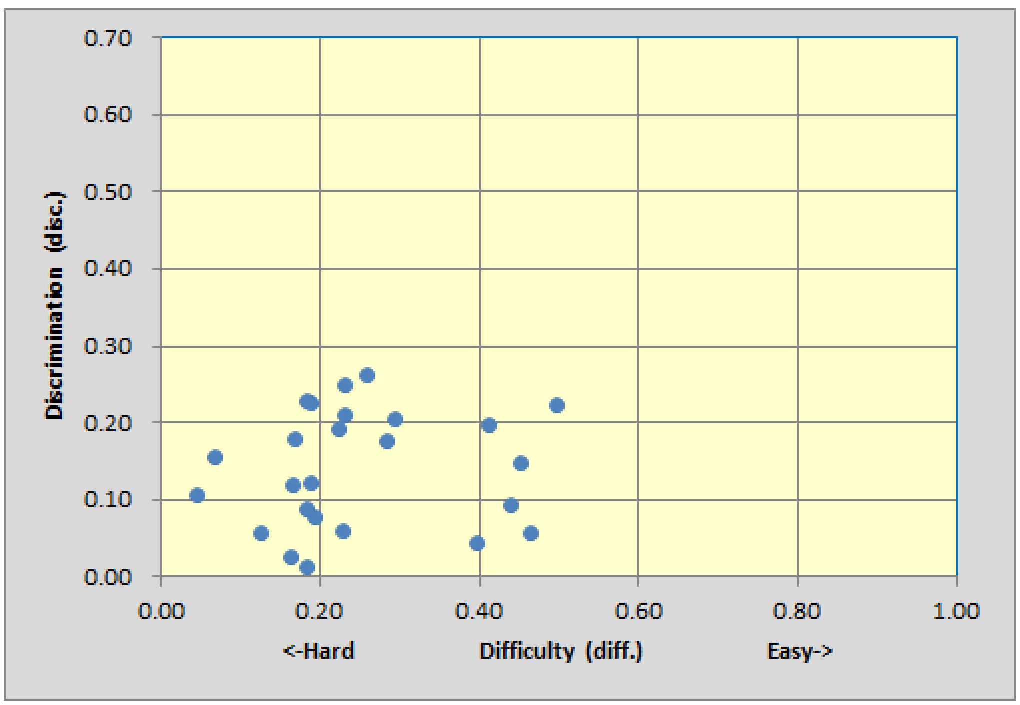
All of these histograms were made by Excel, working under the loving guidance of Lertap. For a reminder of how to get works of art like these, click.

Continue to the next topic for a discussion of results.

2.13.1 Sample results

The scatterplots of item statistics, diff. and disc., looked like this:





The top graph is for the Math subtest; Fisica is at the bottom.

The math subtest has some hard-working items; many of them have disc. values above 0.30. The reliability of this test, the coefficient alpha value found in the Stats1f report, was 0.77.

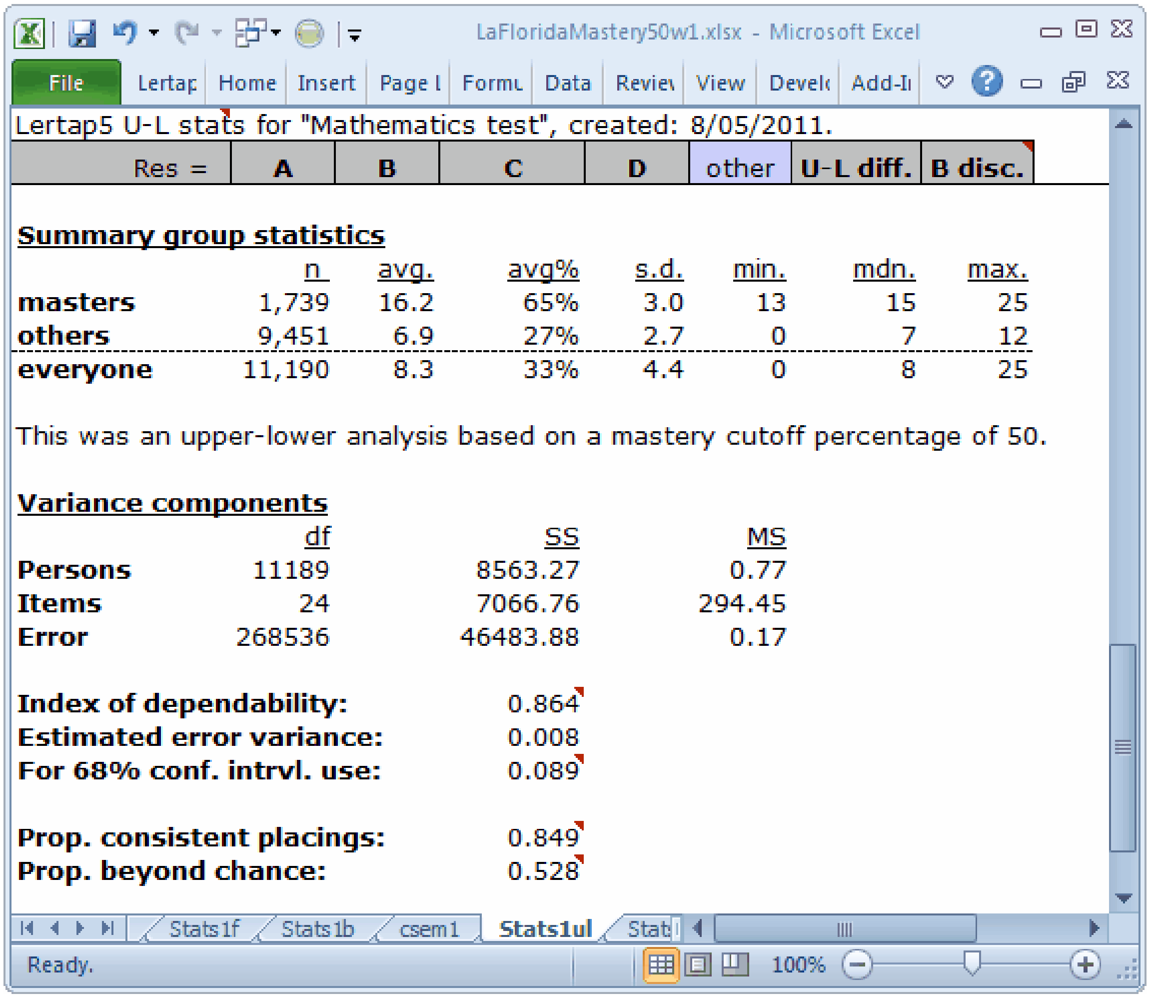
The physics subtest, "Fisica", had many items with disc. values below 0.10, and none above 0.30. Its reliability was just 0.51.

These are marginal reliability figures, especially when it comes to the Fisica subtest.

But what about the mastery test results? Each of these tests has a cut-off score of 50%.

The Statsul reports will have the relevant information. Because there are two subtests, there will be two such reports.

The Stats1ul report has results for the Math subtest:



Lertap5 U-L stats for "Mathematics test", created: 8/05/2011.

Res =	A	B	C	D	other	U-L diff.	B disc.

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
masters	1,739	16.2	65%	3.0	13	15	25
others	9,451	6.9	27%	2.7	0	7	12
everyone	11,190	8.3	33%	4.4	0	8	25

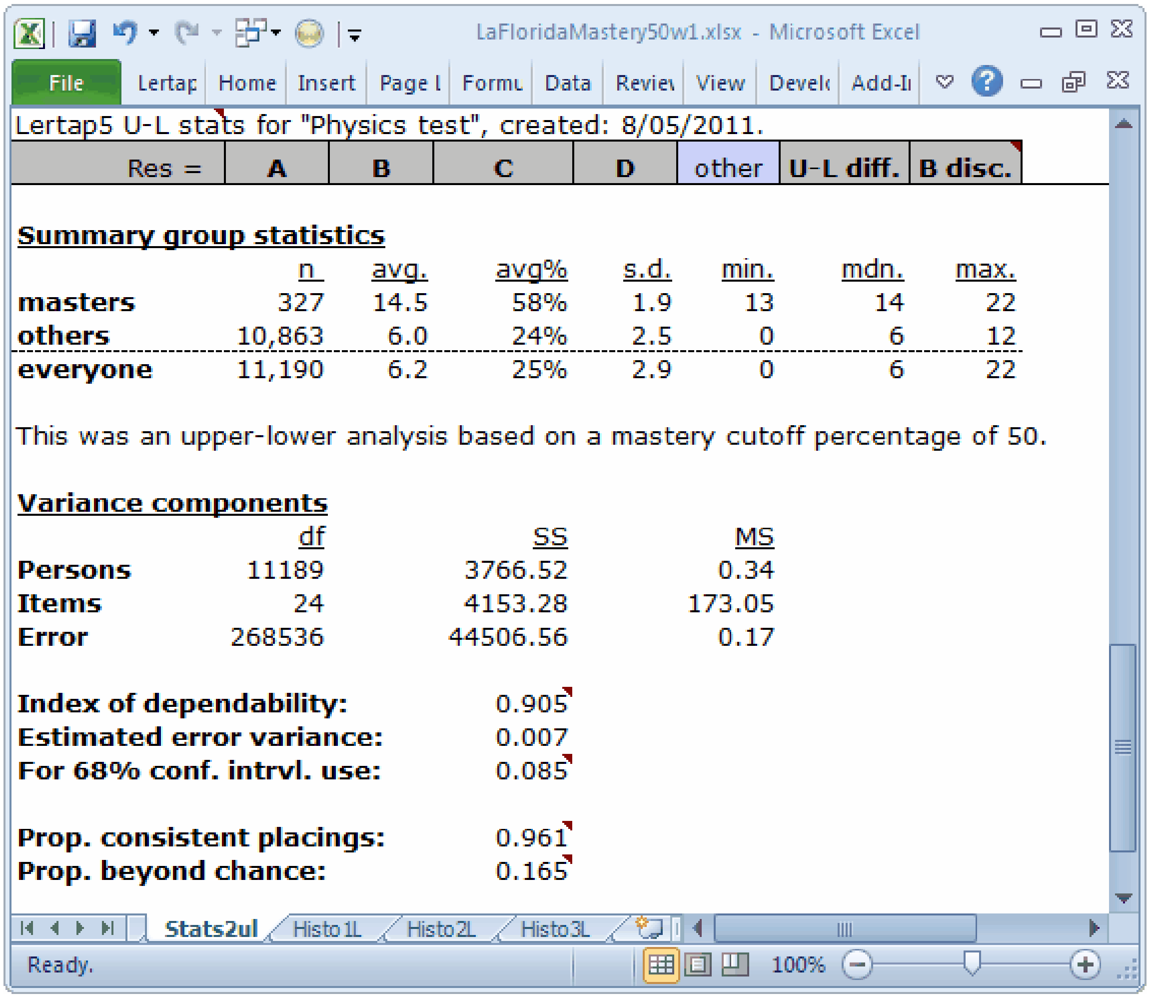
This was an upper-lower analysis based on a mastery cutoff percentage of 50.

Variance components

	df	SS	MS
Persons	11189	8563.27	0.77
Items	24	7066.76	294.45
Error	268536	46483.88	0.17

Index of dependability:	0.864
Estimated error variance:	0.008
For 68% conf. intrvl. use:	0.089
Prop. consistent placings:	0.849
Prop. beyond chance:	0.528

The Stats2ul report has results for Fisica.



Lertap5 U-L stats for "Physics test", created: 8/05/2011.

Res =	A	B	C	D	other	U-L diff.	B disc.

Summary group statistics

	n	avg.	avg%	s.d.	min.	mdn.	max.
masters	327	14.5	58%	1.9	13	14	22
others	10,863	6.0	24%	2.5	0	6	12
everyone	11,190	6.2	25%	2.9	0	6	22

This was an upper-lower analysis based on a mastery cutoff percentage of 50.

Variance components

	df	SS	MS
Persons	11189	3766.52	0.34
Items	24	4153.28	173.05
Error	268536	44506.56	0.17

Index of dependability:	0.905
Estimated error variance:	0.007
For 68% conf. intrvl. use:	0.085
Prop. consistent placings:	0.961
Prop. beyond chance:	0.165

In terms of pass rates, 1739 applicants (15.5%) made it over the bar on the Math subtest, with far fewer, just 327 (3%), topping the Fisica items.

On the Total score, 851 applicants had a combined Math plus Fisica score at 25 or above, for a pass rate of about 7.5%. (Note: a histogram of the Total score was used to determine this number.)

Of interest are the Prop. consistent placings figures for the two subtests. Fisica has particularly weak reliability, but a seemingly (almost) remarkable Prop. consistent placings value of 0.961.

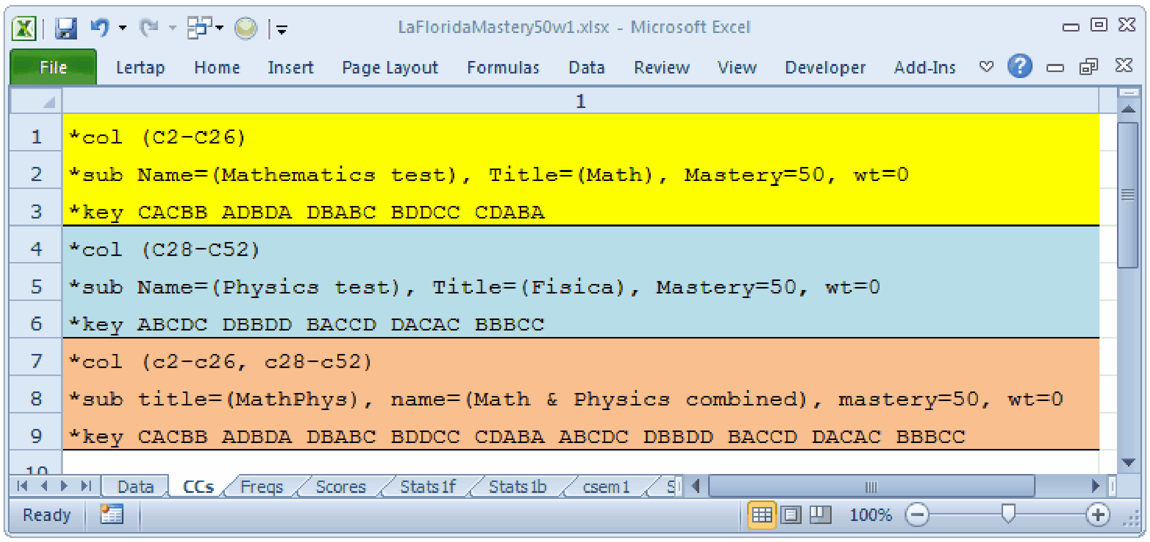
Of course this would not strike you as remarkable had you read this paper about cut-scores and such. What's happened, basically, is that the cut-score on the Fisica test is quite removed from the average test score.

For something interesting to do in your spare time, whiz ahead to the next topic.

2.13.2 Download

Get yourself rigged up with a copy of the La Florida dataset here. Note that you'll get a zip file in this case. The workbook found within the zip file is called "LaFloridaMastery50w1.xlsx".

As to a nifty activity to pursue, here's just one: why not see what would happen with the following CCs lines?



Now we'll have three subtests, the last being all 50 items combined. Does it make sense to add the Math and Fisica items together? Will the reliability of this combined test be higher than than found for the first two subtests? What about Prop. consistent placings for the combined test?

Why do the *sub lines now have wt=0 at the end? We'll answer that one: when there is more than one subtest, Lertap will automatically make a total test score by adding together all the subtest scores, making a global composite score. The wt=0 assignments control the weight each subtest score is given when it comes into this composite. If each subtest has wt=0, no total score will be made. In this case that's entirely appropriate as the third subtest is, in fact, the total.

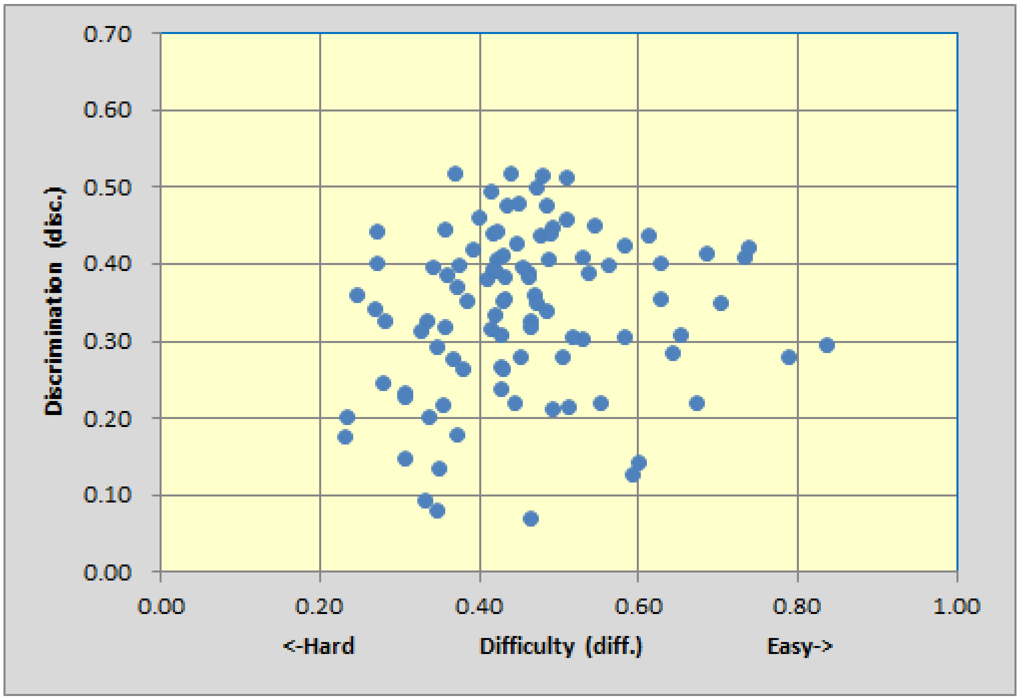
2.14 HalfTime

The results in the HalfTime workbook are from a 100-item mid-semester exam on statistics, administered to undergraduate educational psychology students (n=424).

The exam was meant to discriminate, to identify the strongest students. It was not a mastery test; there was no minimum score required to pass. At the end of the semester, scores on this test were included as part of the final student assessment;

other assessment components included the final exam, and marks from four assignments.

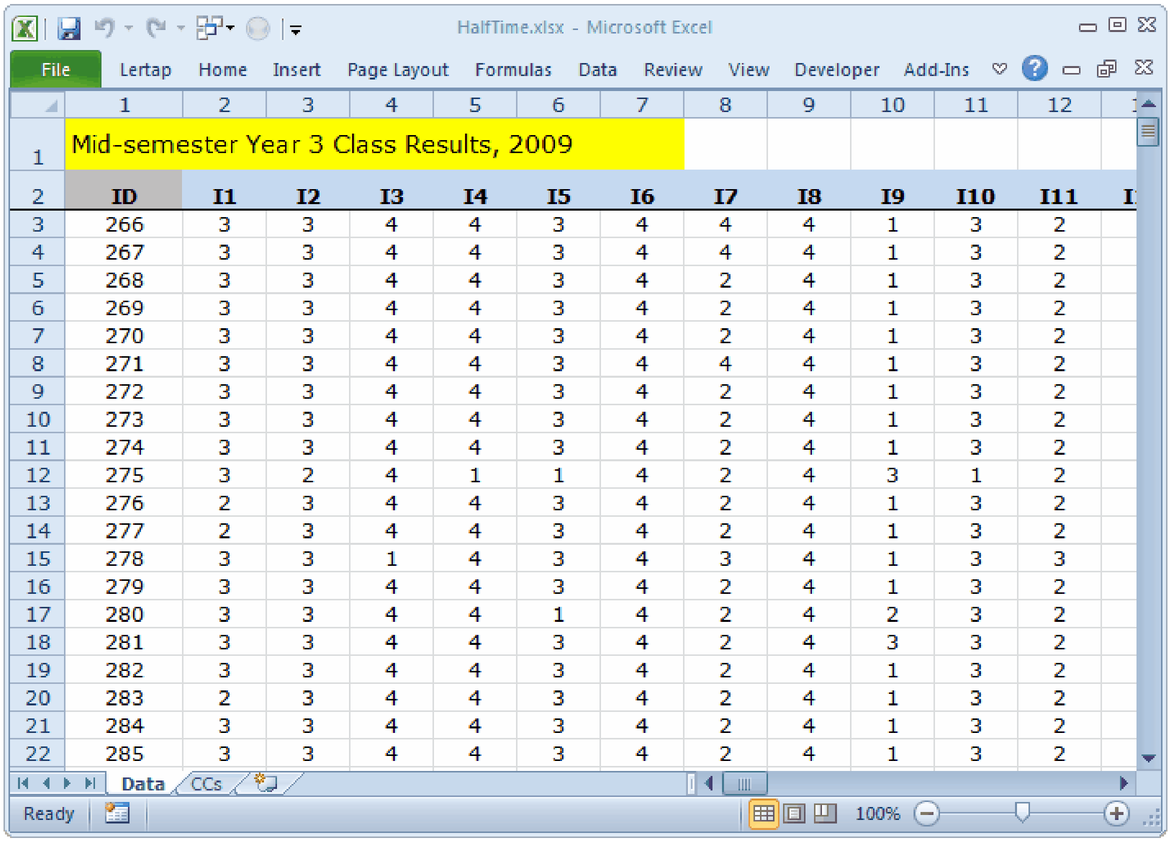
As you will see if you download the workbook, and try things out for yourself, this administration of the test had a reliability of 0.93 (coefficient alpha). Here's the diff./disc. scatter from the bottom of the Stats1b report; admire all the items with disc. values above 0.40, and note the absence of negative disc. items, contributing factors behind the nice alpha value:



What we mostly want to do with the HalfTime workbook is demonstrate how to find **split-half** reliability estimates by using Lertap. It's always been possible to do this in Lertap, even though, in 2005, a reviewer suggested that it wasn't (at that time we said we'd come forth with an example of how to do it, but until now we haven't). Another impetus for the present example stems from the April 2011 release of Iteman 4, another item analysis program available at Assessment Systems Corporation. Iteman 4 has several features which were not available in the previous version, Iteman 3.6. Among these is the computation of a variety of split-half reliability figures. But, you need not use Iteman if you have a split-half interest; Lertap can do the splits, too.

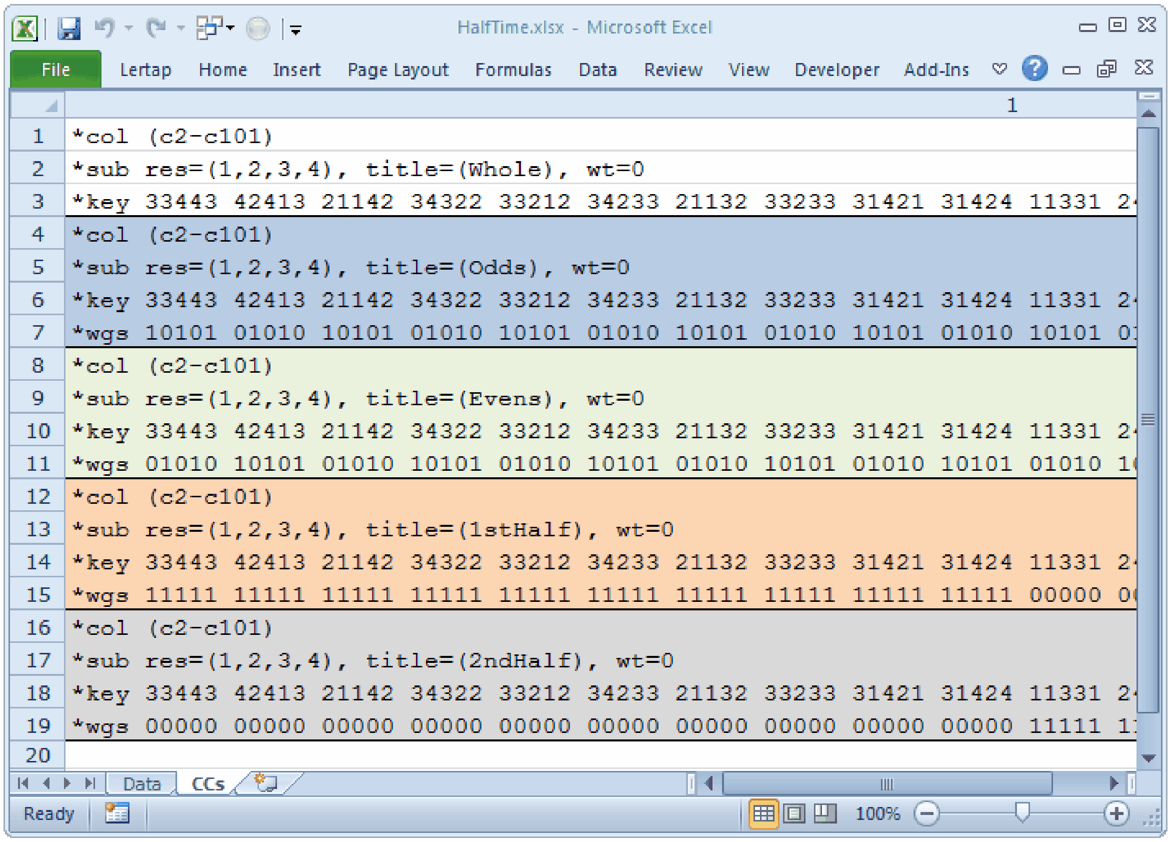
Should you already be wondering if you too might want to be using split-half reliability stuff, we'd suggest no. This is specialist material; not many people have a need for split-half estimates. Coefficient alpha has become the workhorse for estimating reliability (but, before alpha came along, methods for splitting a test into two parts in order to derive a reliability estimate were popular topics indeed).

But wait! Don't go away. You may not ever want to split into halves, but, nonetheless, there might very well be something in this demonstration which you'll find worthwhile as you endeavor to expand your Lertap skills. (*Hope so.*)



Mid-semester Year 3 Class Results, 2009

ID	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11
266	3	3	4	4	3	4	4	4	1	3	2
267	3	3	4	4	3	4	4	4	1	3	2
268	3	3	4	4	3	4	2	4	1	3	2
269	3	3	4	4	3	4	2	4	1	3	2
270	3	3	4	4	3	4	2	4	1	3	2
271	3	3	4	4	3	4	4	4	1	3	2
272	3	3	4	4	3	4	2	4	1	3	2
273	3	3	4	4	3	4	2	4	1	3	2
274	3	3	4	4	3	4	2	4	1	3	2
275	3	2	4	1	1	4	2	4	3	1	2
276	2	3	4	4	3	4	2	4	1	3	2
277	2	3	4	4	3	4	2	4	1	3	2
278	3	3	1	4	3	4	3	4	1	3	3
279	3	3	4	4	3	4	2	4	1	3	2
280	3	3	4	4	1	4	2	4	2	3	2
281	3	3	4	4	3	4	2	4	3	3	2
282	3	3	4	4	3	4	2	4	1	3	2
283	2	3	4	4	3	4	2	4	1	3	2
284	3	3	4	4	3	4	2	4	1	3	2
285	3	3	4	4	3	4	2	4	1	3	2



This CCs worksheet defines how many subtests? How many *col lines are there? Five. There are five subtests.

The *col lines are the same. Each subtest makes use of the same item responses. They're found in columns 2 through 101 of the Data worksheet, that is, c2-c101.

The *sub lines don't differ by much; they just give distinct "titles" to each of the subtests.

The *key lines are all identical.

It's the ***wgs** lines which differ.

What are *wgs lines, anyway? To get the full answer, you might care to look at the reference topic, and find out for yourself.

If you did trouble to have a look at that topic, you'd discover that *wgs is the same as *wts: the entries in the *wgs line correspond to the number of points given for the right answer to an item.

Okay, why isn't there a *wgs line for the first subtest? And, for that matter, why don't the CCs lines seen in the other samples use *wgs lines?

Because they're seldom required. When we score test items on a right-wrong basis, it's common to give one point for the right answer, zero points otherwise. This is Lertap's default action. When this default is the appropriate way to score the test, no *wgs lines are required.

Now, look at our various *wgs lines. They have ones and zeros. Zeros? Yes; we're telling Lertap to give zero points to the right answer for some items. "*Hmmm*", says Lertap to itself. "*If he wants to give zero points for a right answer, I'm just going to ignore the item altogether".*

Items with a zero on the *wgs line are not only "ignored", they are actually excluded from Lertap's reports. This is what we want to do now: exclude some of the items some of the time, depending on the subtest involved. In the split-half world, people devise ways to create two half tests. In this example, the test has 100 items, and we want to make two half tests, each with 50 items. There are a variety of ways in which this may be accomplished. Two of the most common are to split the test on an "odd-even" basis, and/or to split it by first half / second half.

The definition of the second subtest above starts at CCs line 4. We're defining the "odds" half test by (naturally) picking out only the "odd" items: the first, the third, the fifth, the seventh, and so on. These are the items to be scored, with one point for the right answer. And, we want to knock out the even items, so we give them a scoring value of zero. Thus the first *wgs line goes 10101 01010, and so on. It's picking out only the "odd" items (we should say "the odd-numbered items", but items don't always have numbers: I1, I3, I5, I7, and so on is what we mean to say).

The third subtest, starting in CCs line 8, will be our "evens" half test.

The fourth subtest, starting in CCs line 12, will be the "first half" half test. See how its *wgs line is picking out the first 50 items? You do? Good; then it will be obvious to you how we set up the second half, the fifth subtest, starting in CCs line 16.

Okay? Let's get cracking. Time for, you guessed it:

The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

Make it one step?

Getting a bit old, not enough stamina for the old 3-step routine?

Not to worry. Turn on "production mode" and get more output with fewer steps.

Next topic? Just click.

Tidbits:

There's more than one way to set up these subtests, to be sure. For example, we could use CCs lines like this to pick out the odd and even items, making it unnecessary to use *wgs lines:

(Note: the odd items have their responses in even-numbered columns in this sample.)

*col (c2,c4,c6,c8,c10, ...,c98,c100)
*sub res=(1,2,3,4), title=(Odds)
*key 34321 ....

*col (c3,c5,c7,c9,c11, ...,c99,c101) (For the even items.)
*sub res=(1,2,3,4), title=(Evens)
*key 34443 ....

We used *wgs lines above as they greatly ease the work. We got to use the same *col line and the same *key line for each of the subtests.

As mentioned, the *wgs (or *wts) line sets the number of points for the right answer to an item. However, it allows for just a single digit, so the points have to be from 1 to 9. To award more points, or fractions of a point, the *mws line is used. The *mws line is also used to give points for more than one answer.

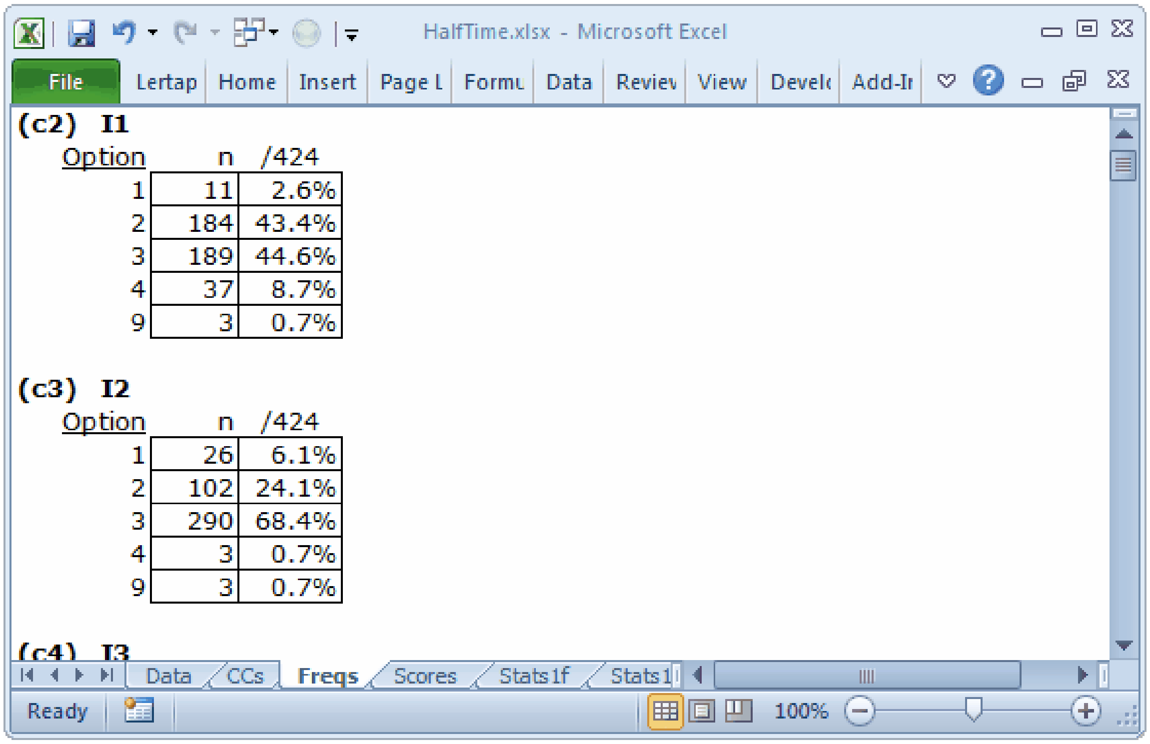
There are more ways to exclude items from a subtest. The *exc line, for example, is one; an example of its use is seen in the CCs worksheet for the LenguaBIg sample.

We like to group entries on CCs lines by 5s, as seen in the *key and *wgs lines above. This makes it easier when we have to, for example, change the key for, say, item I40: it's easier to find the 40th key if we can count by 5s. But, you don't need to group by 5s. You could group by 10s. You could not group at all; maybe you're not a groupie?

2.14.1 Sample results

Always check your data before going out, Mom would say.

Okay, starting with Freqs:



(c2) I1

Option	n	/424
1	11	2.6%
2	184	43.4%
3	189	44.6%
4	37	8.7%
9	3	0.7%

(c3) I2

Option	n	/424
1	26	6.1%
2	102	24.1%
3	290	68.4%
4	3	0.7%
9	3	0.7%

(c4) I3

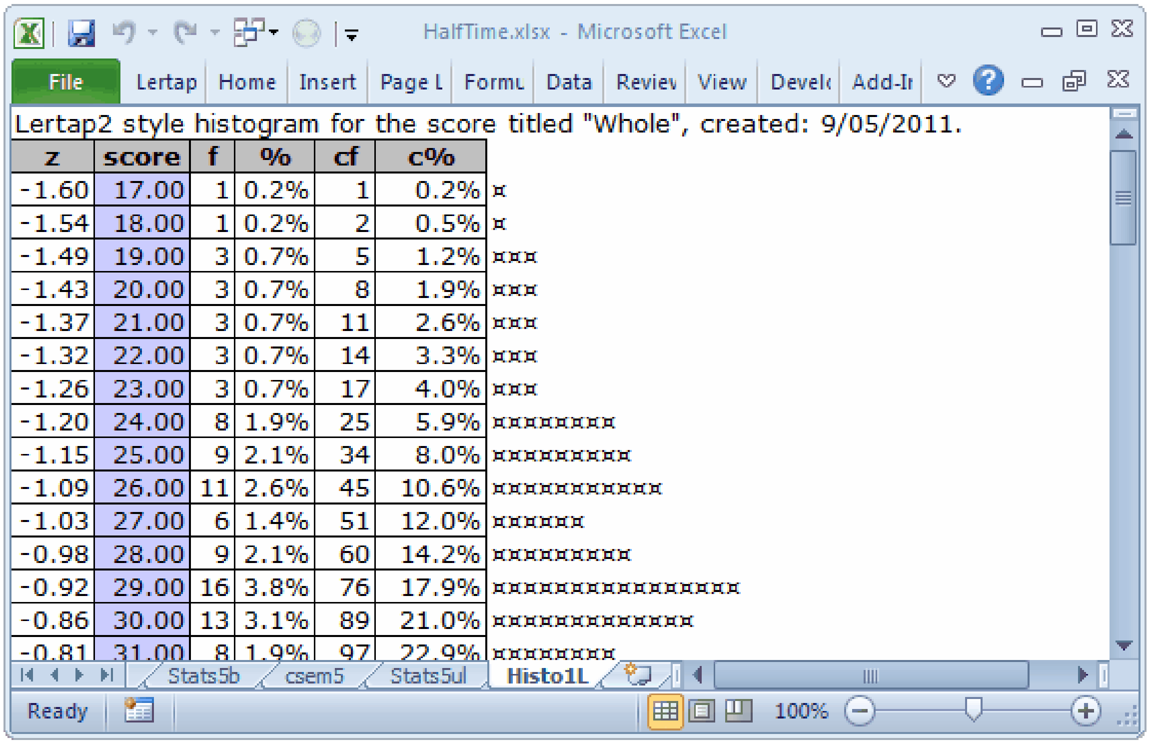
We scrolled down all of the little Freqs tables, looking for "weird" responses, that is, responses which were not 1 2 3 4 (these are the response codes used by each item, sometimes called the "options").

Freqs is trained to report the occurrence of all the 26 letters in the Latin alphabet, upper- and lower-case, as well as the 10 digits, from 0 through 9. If it finds any other characters in the item response columns of the Data worksheet, such as blanks and asterisks, it tallies them in a special row, the ? row.

Almost all of the HalfTime items have "responses" of 9. I1 and I2 had three 9s, as you can see above. In this case, HalfTime, 9s were what the scanner "recorded" whenever it found that a student did not answer an item, or gave more than one answer by shading in more than one bubble on the mark-sense answer sheet.

A few 9s will not discourage us; it is common for students to leave questions unanswered, even when we tell them to guess if they don't know the answer.

What we don't want to find are rows in the Data worksheet which have many, many 9s. Freqs can't tell us this. But, if there are such records, the corresponding test score will be low, possibly even zero. Lertap's traditional (old-style) histogram can indicate if there might be any really low scores, and it's very easy to get:



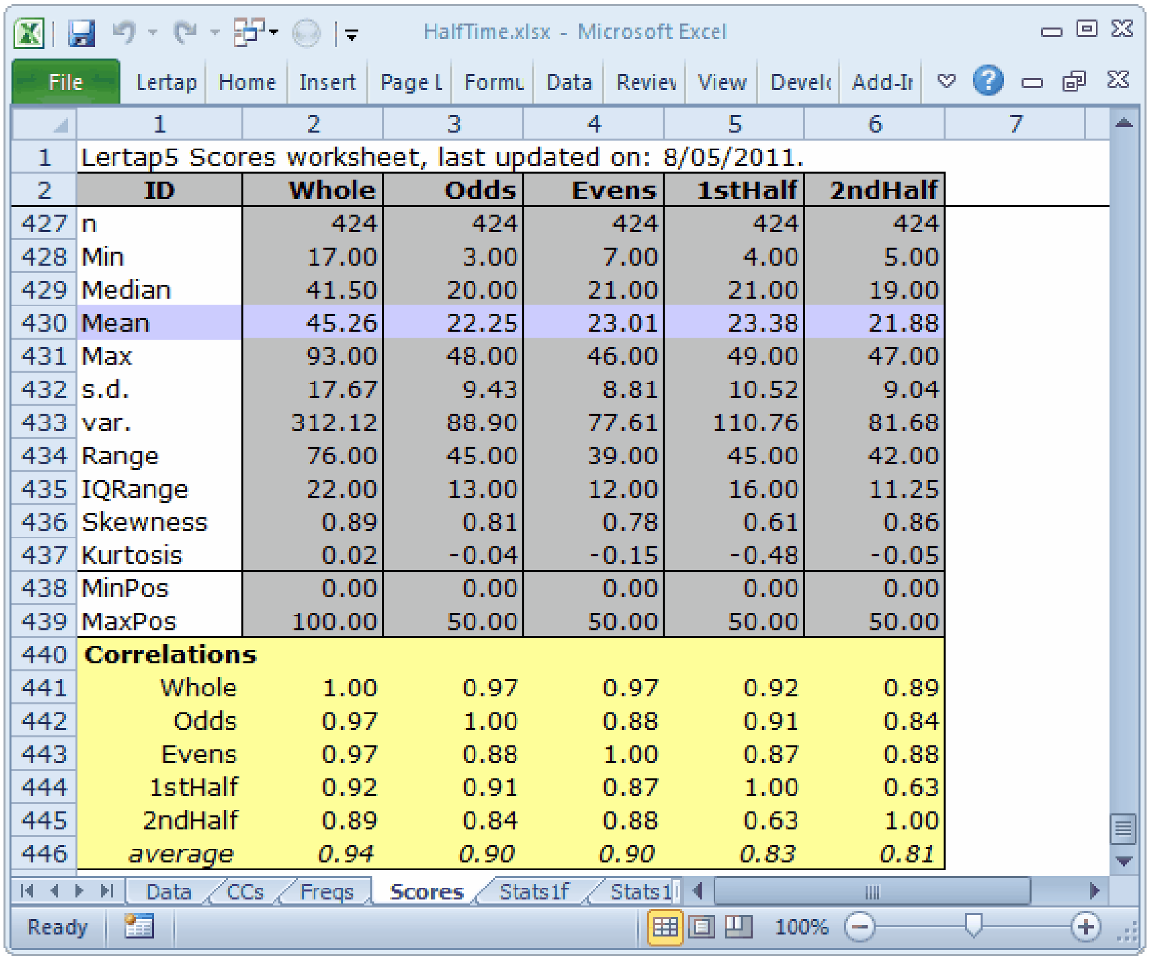
Lertap2 style histogram for the score titled "Whole", created: 9/05/2011.

z	score	f	%	cf	c%	
-1.60	17.00	1	0.2%	1	0.2%	¤
-1.54	18.00	1	0.2%	2	0.5%	¤
-1.49	19.00	3	0.7%	5	1.2%	¤¤¤
-1.43	20.00	3	0.7%	8	1.9%	¤¤¤
-1.37	21.00	3	0.7%	11	2.6%	¤¤¤
-1.32	22.00	3	0.7%	14	3.3%	¤¤¤
-1.26	23.00	3	0.7%	17	4.0%	¤¤¤
-1.20	24.00	8	1.9%	25	5.9%	¤¤¤¤¤¤¤¤
-1.15	25.00	9	2.1%	34	8.0%	¤¤¤¤¤¤¤¤¤
-1.09	26.00	11	2.6%	45	10.6%	¤¤¤¤¤¤¤¤¤¤¤
-1.03	27.00	6	1.4%	51	12.0%	¤¤¤¤¤¤
-0.98	28.00	9	2.1%	60	14.2%	¤¤¤¤¤¤¤¤¤
-0.92	29.00	16	3.8%	76	17.9%	¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤
-0.86	30.00	13	3.1%	89	21.0%	¤¤¤¤¤¤¤¤¤¤¤¤¤
-0.81	31.00	8	1.9%	97	22.9%	¤¤¤¤¤¤¤¤

The Histo1L report's second column, highlighted in blue-gray, shows that the lowest "Whole" HalfTime score was 17. One student had this score. One student had a score of 18. Three had a score of 19. This looks fine; there are no "outliers", low scores which stand on their own -- here, for example, a score of 0 (zero) would have been an outlier, a "weird", unexpected result, probably from a bad results record in the Data worksheet. The M.Nursing sample has an example of a score outlier.

Having followed Mom's advice to check our data, we can now look at some results.

Look at the bottom of the Scores report:



HalfTime.xlsx - Microsoft Excel

Lertap5 Scores worksheet, last updated on: 8/05/2011.

	ID	Whole	Odds	Evens	1stHalf	2ndHalf
427	n	424	424	424	424	424
428	Min	17.00	3.00	7.00	4.00	5.00
429	Median	41.50	20.00	21.00	21.00	19.00
430	Mean	45.26	22.25	23.01	23.38	21.88
431	Max	93.00	48.00	46.00	49.00	47.00
432	s.d.	17.67	9.43	8.81	10.52	9.04
433	var.	312.12	88.90	77.61	110.76	81.68
434	Range	76.00	45.00	39.00	45.00	42.00
435	IQRange	22.00	13.00	12.00	16.00	11.25
436	Skewness	0.89	0.81	0.78	0.61	0.86
437	Kurtosis	0.02	-0.04	-0.15	-0.48	-0.05
438	MinPos	0.00	0.00	0.00	0.00	0.00
439	MaxPos	100.00	50.00	50.00	50.00	50.00
440	**Correlations**					
441	Whole	1.00	0.97	0.97	0.92	0.89
442	Odds	0.97	1.00	0.88	0.91	0.84
443	Evens	0.97	0.88	1.00	0.87	0.88
444	1stHalf	0.92	0.91	0.87	1.00	0.63
445	2ndHalf	0.89	0.84	0.88	0.63	1.00
446	*average*	*0.94*	*0.90*	*0.90*	*0.83*	*0.81*

This was not exactly an easy test. MaxPos, the maximum possible score, was 100 on the "Whole" test, and the median score was 41.50. The medians for the four half tests were in the same ballpark, in the 22 to 23 range.

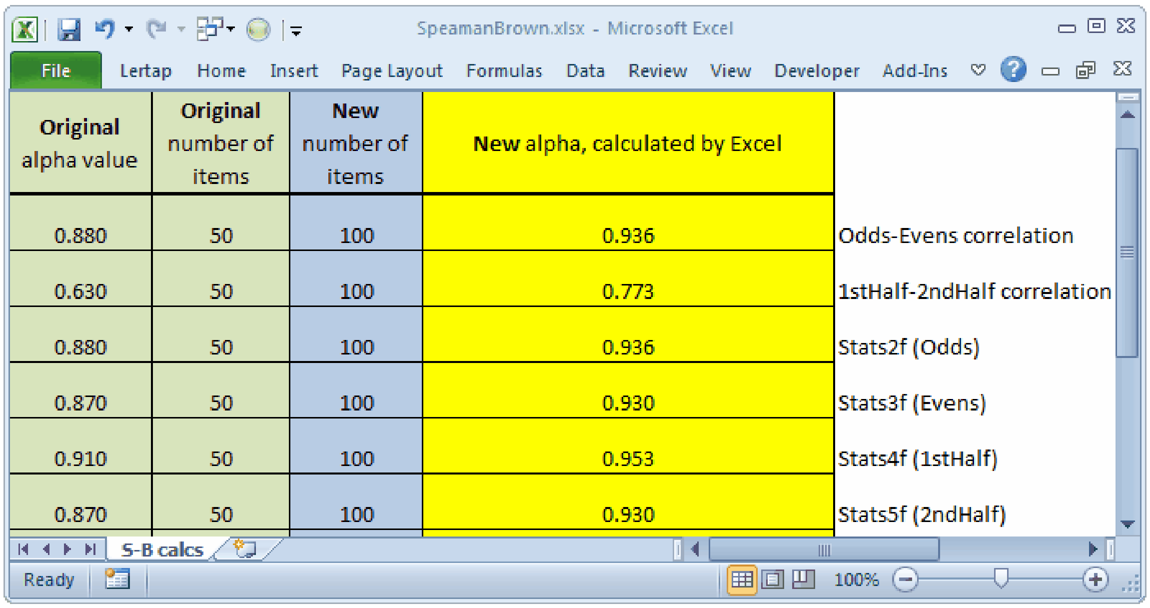
People who work with split-halves will often look at the correlation between respective halves as a reliability figure. This being the case, the split-half reliability for odds-evens would be 0.88. What is surprising is the split-half reliability for 1stHalf-2ndHalf. It's only 0.63.

Here's a summary of the coefficient alpha reliability figures from the various subtests:

Report Name	Subtest Name	No. Items	Alpha
Stats1f	Whole	100	0.93
Stats2f	Odds	50	0.88

Stats3f	Evens	50	0.87
Stats4f	1stHalf	50	0.91
Stats5f	2ndHalf	50	0.87

The reliability of a half test is generally expected to be lower than that for the full test, so we'll trot out our Spearman-Brown calculator to estimate the reliability of each half test if, instead of 50 items, it had 100. (Note: if clicking on the link doesn't get the calculator spreadsheet to download, right-click on it and select "save as".)



Original alpha value	Original number of items	New number of items	New alpha, calculated by Excel	
0.880	50	100	0.936	Odds-Evens correlation
0.630	50	100	0.773	1stHalf-2ndHalf correlation
0.880	50	100	0.936	Stats2f (Odds)
0.870	50	100	0.930	Stats3f (Evens)
0.910	50	100	0.953	Stats4f (1stHalf)
0.870	50	100	0.930	Stats5f (2ndHalf)

Please keep in mind that our main goal here relates to showing how to use Lertap. There may very well be people in the audience who are full-bottle on split-half methods, and we know that they might be saying that the Spearman-Brown formula is not the best way to estimate the reliability of a longer test unless the halves have the same standard deviations (s.d.). When standard deviations are unequal, Rulon's method is preferred over Spearman-Brown (search the internet for "Rulon's method" or "Rulon's formula" or "Rulon split-half").

Of great relevance here is this (Crocker & Algina, 1986, p. 142): **coefficient alpha** "*... is the mean of all possible split-half coefficients that are calculated using the Rulon method ....*".

Well, you okay with this stuff? Why not have a little download of HalfTime, and explore more on your own?

2.14.2 Download

The HalfTime.xlsx workbook may be downloaded by clicking here.

Of all the many possible exercises which might be undertaken with the HalfTime dataset, one could involve getting scatterplots of half-test scores.

For example, we could ask Lertap to ask Excel to plot the "Odds" score against the "Evens" (the correlation of these two scores is 0.88). This is done using the "Scatterplotter". Then, we could compare the plot with another one corresponding to "1stHalf" and "2ndHalf" (correlation of 0.63). Is there some reason for observing such a relatively low correlation among the 1stHalf and 2ndHalf scores?

2.15 Mente2010

A "mature-aged" doctoral student took on the task of developing a new test of what will here be called mental "tenacity".

His goal was to produce an instrument which would rival the quality of two popular, widely-used tests, but be substantially less expensive for people to take (the other tests could cost as much as $155 for a student to sit; our doctoral student, working with backing from a major test publisher, sought to come forth with an instrument which would cost under $40 per sitting).

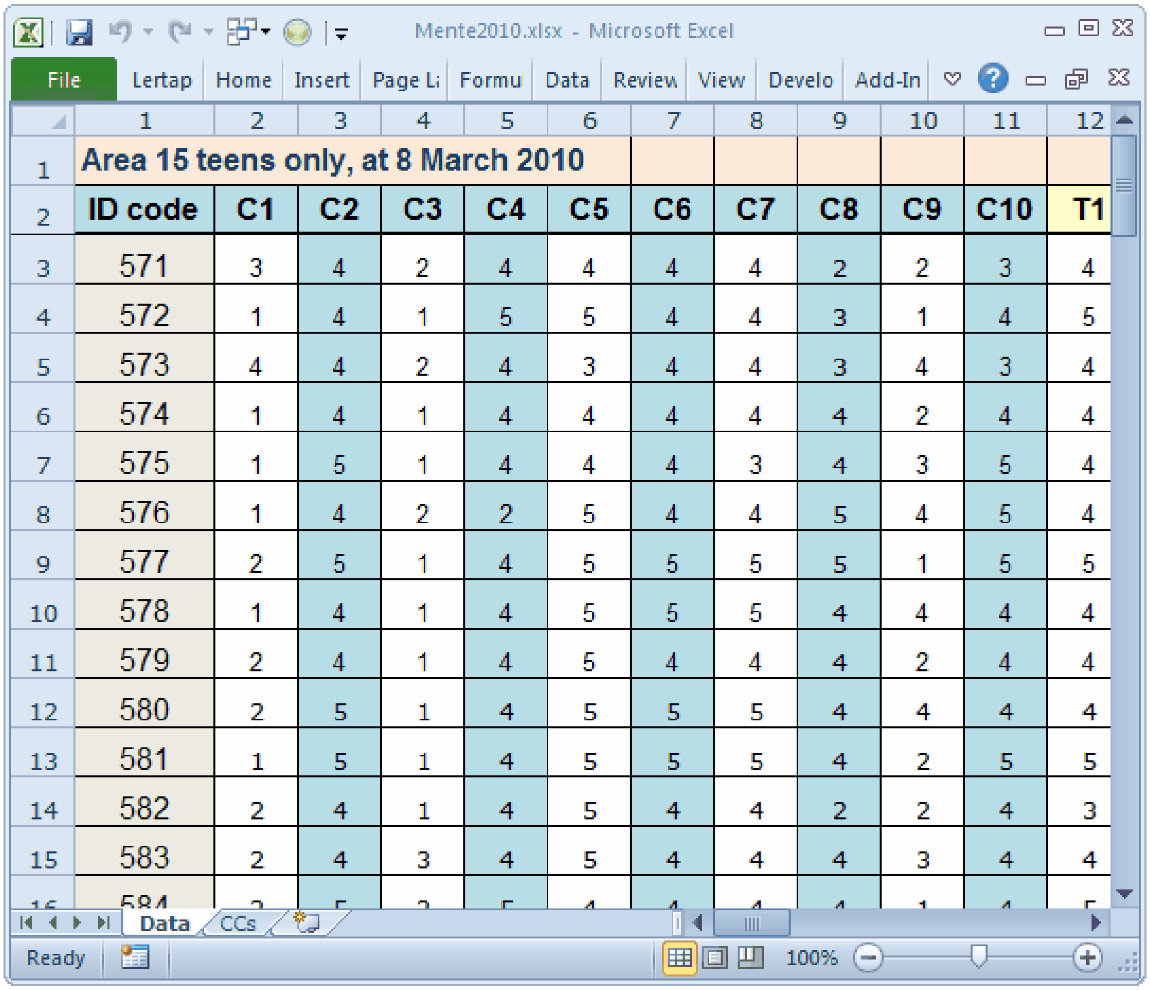
By mental "tenacity" is meant the propensity of a person to persist with a task in the face of adversity. We all know that some people are prone to give up when faced with a new challenge, while others are inclined to persist, even when the going gets really tough. Those who persist are said to be tenacious. Tenacious people would be expected to do well on the doctoral student's new instrument.

The doctoral student developed an initial pool of 80 test items. Sample items would have been similar to these:

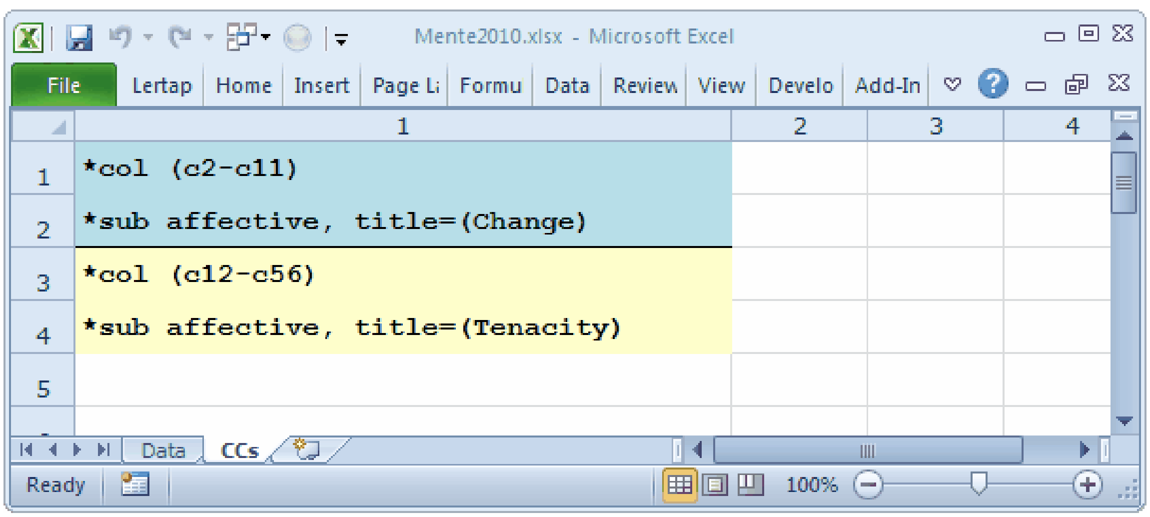
(1)	*I thrive when presented with a difficult task.*	
	1	strongly disagree
	2	disagree
	3	undecided
	4	agree
	5	strongly agree
(2)	*I feel stressed when faced with a challenge I know others find hard to master.*	
	1	strongly disagree
	2	disagree
	3	undecided
	4	agree
	5	strongly agree

After a typical review process in which experts judged the items in the initial pool, a trial instrument using 45 of the 80 items was developed for pilot testing. The doctoral student, his major professor, and the test publisher all recognised that this was still too many items -- it was thought that the final test should have no more than 40 items, hopefully even fewer.

The Data and CCs worksheets for this study are seen below.



	1	2	3	4	5	6	7	8	9	10	11	12
1	Area 15 teens only, at 8 March 2010											
2	ID code	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	T1
3	571	3	4	2	4	4	4	4	2	2	3	4
4	572	1	4	1	5	5	4	4	3	1	4	5
5	573	4	4	2	4	3	4	4	3	4	3	4
6	574	1	4	1	4	4	4	4	4	2	4	4
7	575	1	5	1	4	4	4	3	4	3	5	4
8	576	1	4	2	2	5	4	4	5	4	5	4
9	577	2	5	1	4	5	5	5	5	1	5	5
10	578	1	4	1	4	5	5	5	4	4	4	4
11	579	2	4	1	4	5	4	4	4	2	4	4
12	580	2	5	1	4	5	5	5	4	4	4	4
13	581	1	5	1	4	5	5	5	4	2	5	5
14	582	2	4	1	4	5	4	4	2	2	4	3
15	583	2	4	3	4	5	4	4	4	3	4	4



	1	2	3	4
1	*col (c2-c11)			
2	*sub affective, title=(Change)			
3	*col (c12-c56)			
4	*sub affective, title=(Tenacity)			
5				

Responses to the 45 trial items start in column 12 (c12) of the Data worksheet.

The doctoral student included ten other items in his study, C1 through C10. These items were drawn from an area of the literature which has to do with how people perceive their ability to change: are our mental and physical abilities set, or can we improve on them? (Similar to the "nature" versus "nurture" debate.)

In this case, the ten items included in the pilot test were similar in tone to these:

(C1)	*We can always learn to do new things, even when they may require much effort.*	
	1	strongly disagree
	2	disagree
	3	undecided
	4	agree
	5	strongly agree
(C2)	*Our ability to master things is limited to the set of skills given to us at birth.*	
	1	strongly disagree
	2	disagree
	3	undecided
	4	agree
	5	strongly agree

Questions like these are sometimes used to get an idea of a person's potential to succeed at something which she or he has not done before, such as playing polo, learning a new language, or playing a musical instrument. Some people might say, for example, when presented with the challenge of learning Thai, "*No way, Jose. I have no ability at all for languages*", while someone else might say "*Well, I only speak English. Thai is said to be difficult, but sure, why not, I can give it a try*".

Try this?

2.15.1 Sample results 1

Data analysis was not a strength of the "mature-aged" doctoral student. His major adviser referred the student to the School's IT support group. Someone there had the student set up the Data and CCs sheets, and go through the steps required to get results with Lertap. This the student did while the adviser was away on a brief sabbatical.

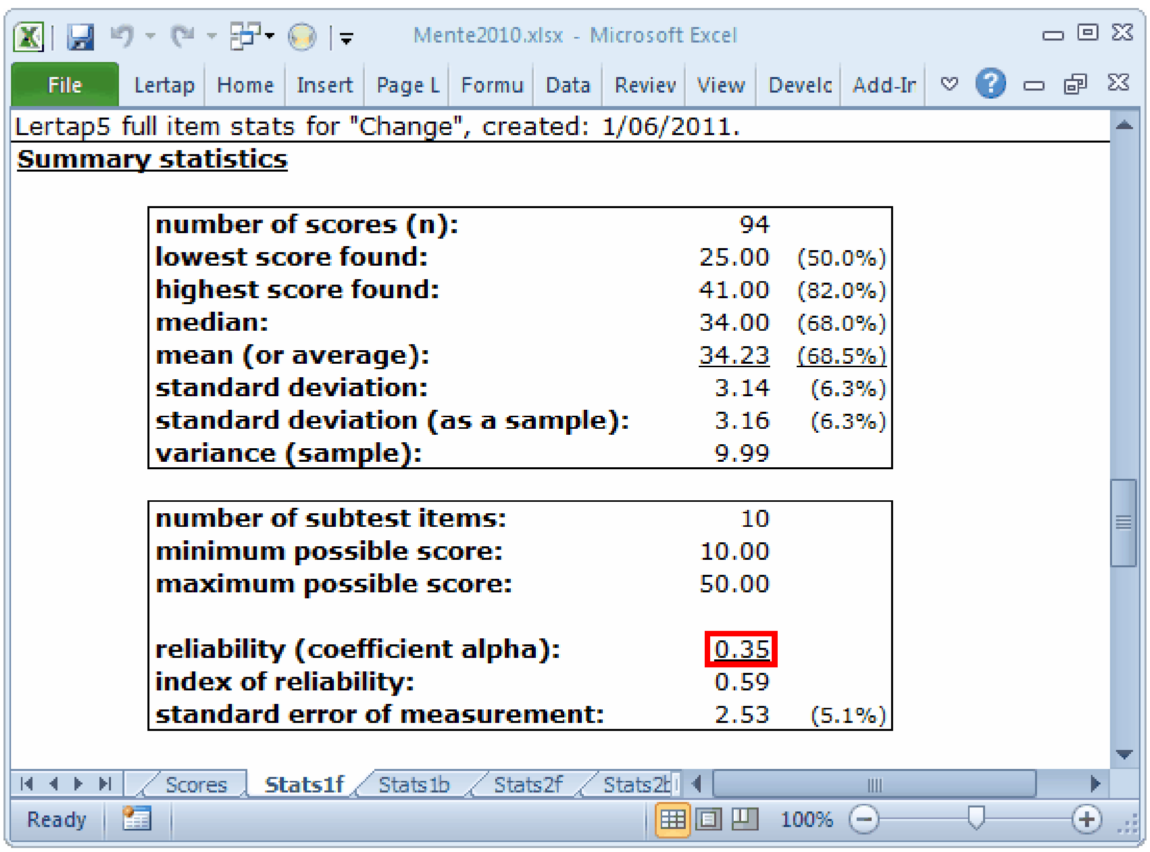
The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

The student was discouraged by what he found. The first subtest, "Change", had a reliability of just 0.35, as seen in the Stats1f report:



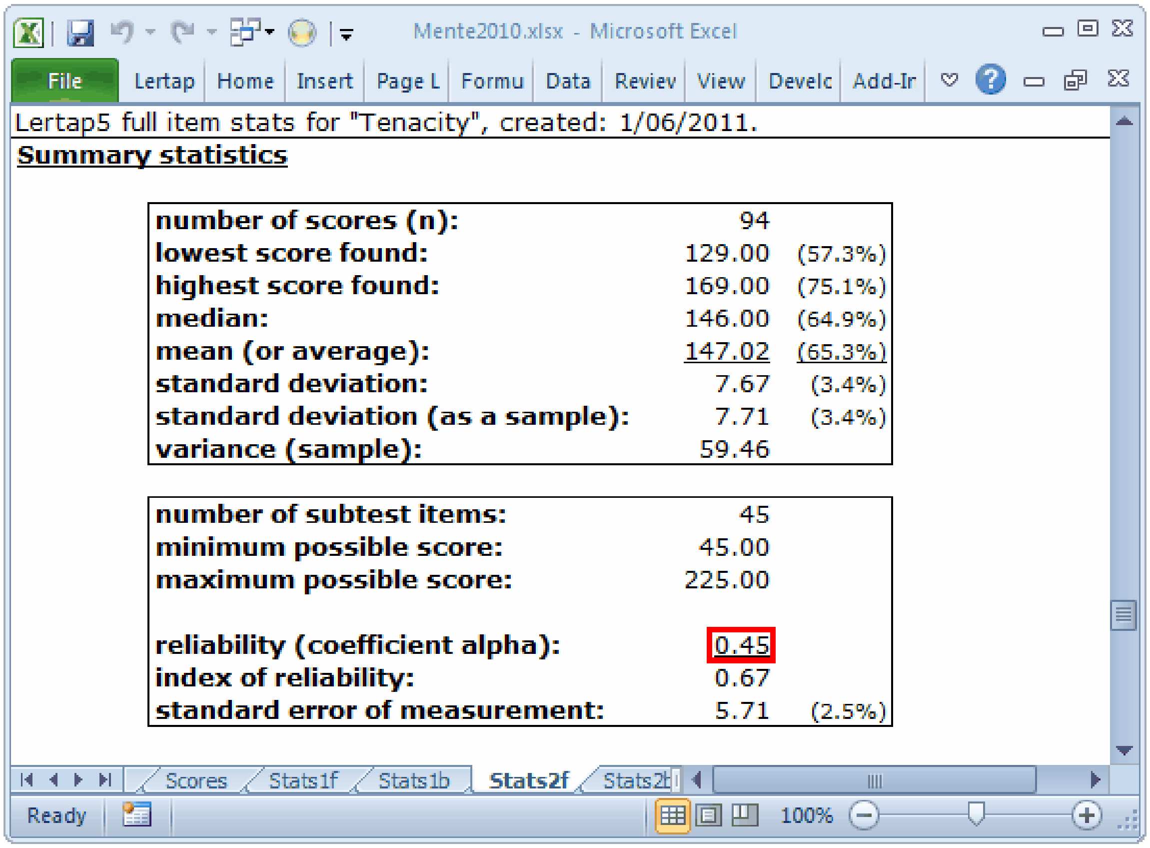
Lertap5 full item stats for "Change", created: 1/06/2011.

Summary statistics

number of scores (n):	94	
lowest score found:	25.00	(50.0%)
highest score found:	41.00	(82.0%)
median:	34.00	(68.0%)
mean (or average):	34.23	(68.5%)
standard deviation:	3.14	(6.3%)
standard deviation (as a sample):	3.16	(6.3%)
variance (sample):	9.99	

number of subtest items:	10	
minimum possible score:	10.00	
maximum possible score:	50.00	
reliability (coefficient alpha):	0.35	
index of reliability:	0.59	
standard error of measurement:	2.53	(5.1%)

The second subtest was not much better:



Lertap5 full item stats for "Tenacity", created: 1/06/2011.

Summary statistics

number of scores (n):	94	
lowest score found:	129.00	(57.3%)
highest score found:	169.00	(75.1%)
median:	146.00	(64.9%)
mean (or average):	147.02	(65.3%)
standard deviation:	7.67	(3.4%)
standard deviation (as a sample):	7.71	(3.4%)
variance (sample):	59.46	

number of subtest items:	45	
minimum possible score:	45.00	
maximum possible score:	225.00	
reliability (coefficient alpha):	0.45	
index of reliability:	0.67	
standard error of measurement:	5.71	(2.5%)

It is commonly agreed that an affective scale should have a minimum alpha value of 0.70 to be useful; some authors suggest 0.80 as a working *minimum*.

Now, coefficient alpha is affected by a number of factors. One of them is test length. The first subtest in this example has only 10 items, but, according to the graduate student, affective scales in this area, "change", will often have a small number of items, yet will return alpha values much higher than 0.35.

Another factor which affects alpha relates to how the item responses are scored. Lertap makes some important assumptions in this regard.

It is assumed that affective items use "response codes" of 1 2 3 4 5. And, Lertap assumes that one point is to be awarded if someone selects the first "response code", two points for the second response code, ..., and five points for the fifth. It's possible to override these assumptions. For example, if other response codes are used, then an Res= declaration is made on the *sub line.

The graduate student, not at all an experienced data analyst, rode into Lertap with the CCs lines seen in the previous topic, implicitly accepting Lertap's default assumptions, and then reporting great dismay with the low alpha values found.

What he should have done, as a first step, was reverse the scoring of his negative items.

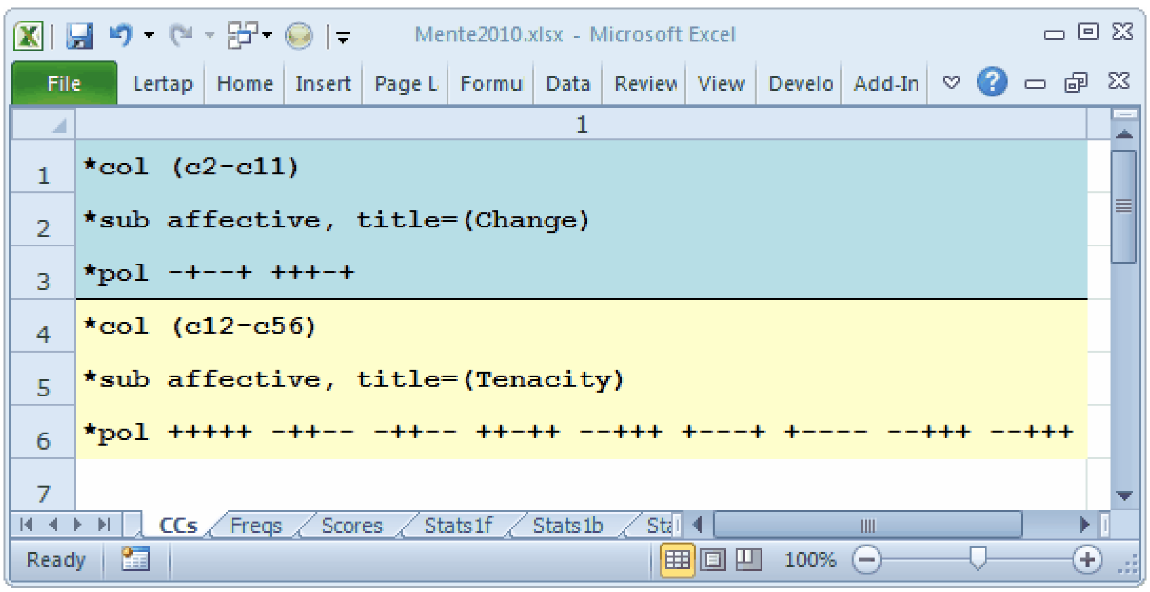
If the first item on the "tenacity" scale was *I thrive when presented with a difficult task,* and the second item was *I feel stressed when faced with a challenge I know others find hard to master,* then he should have reversed the way the second item was scored, giving five points for strongly disagree, and one for strongly agree. He wants a high "tenacity" score to correspond to those who are not adversely affected by a challenge.

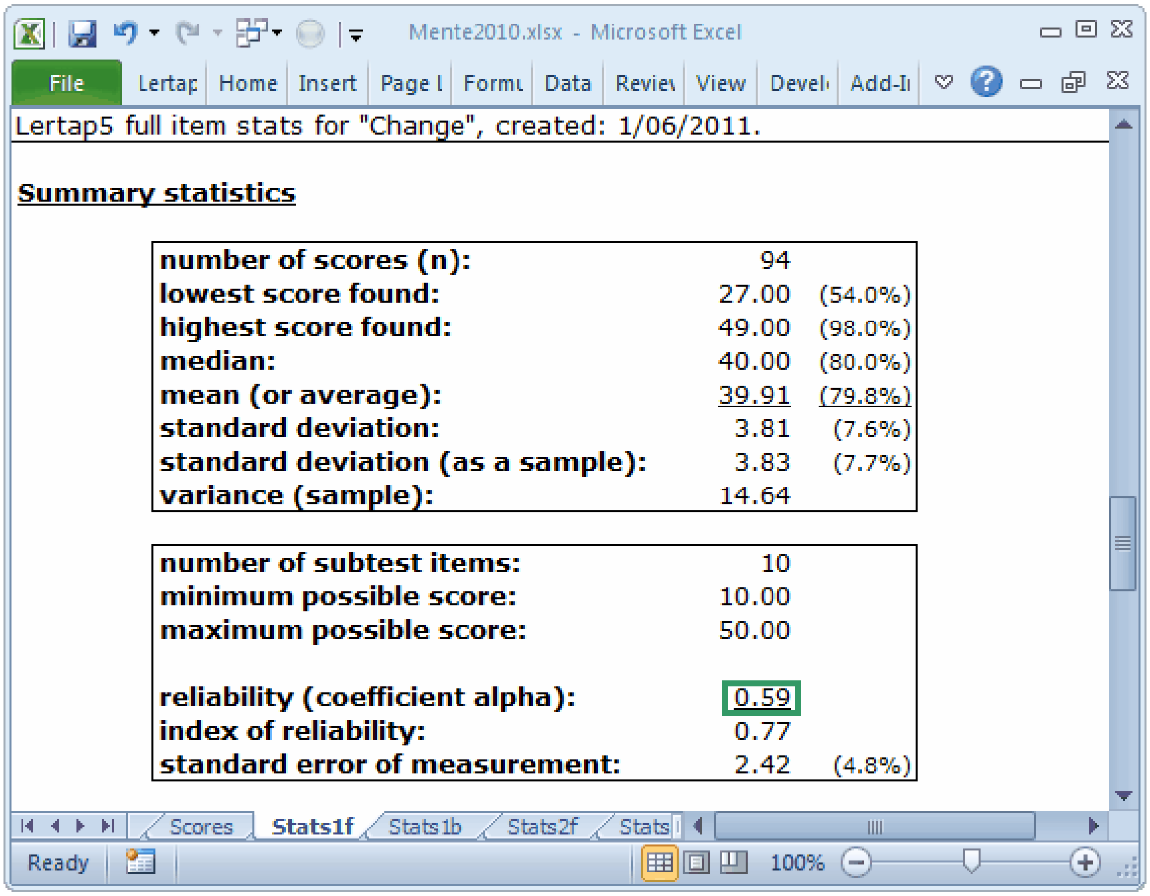
The best way to determine which items should be reverse scored is to look each item in the face and decide it it's negatively phrased. Then a *pol line is used in the CCs sheet to accomplish the reverse scoring. [Page ahead](#) to see examples.

2.15.2 Sample results 2

What a difference a minus sign or two (or three, ...) can make. See for yourself:

The CCs lines were changed by adding appropriate *pol lines:



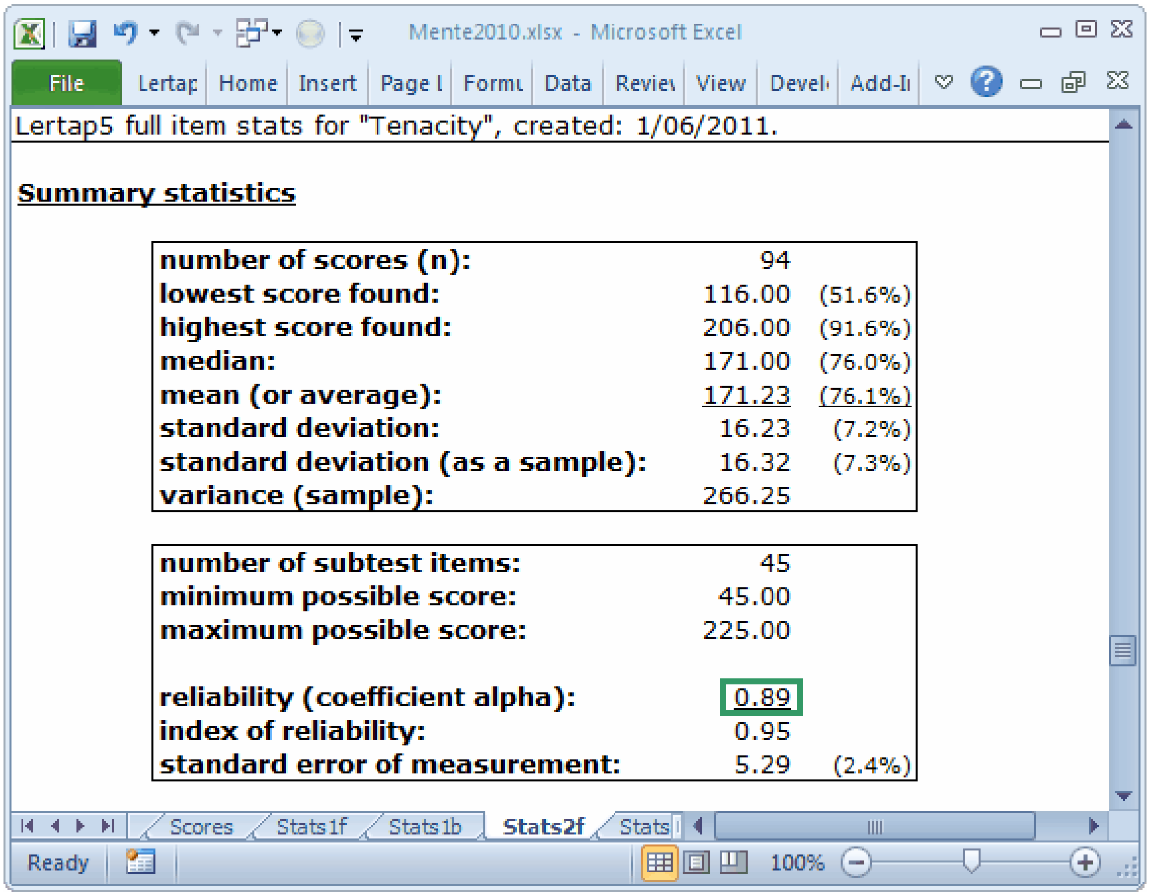


Lertap5 full item stats for "Change", created: 1/06/2011.

Summary statistics

number of scores (n):	94	
lowest score found:	27.00	(54.0%)
highest score found:	49.00	(98.0%)
median:	40.00	(80.0%)
mean (or average):	39.91	(79.8%)
standard deviation:	3.81	(7.6%)
standard deviation (as a sample):	3.83	(7.7%)
variance (sample):	14.64	

number of subtest items:	10	
minimum possible score:	10.00	
maximum possible score:	50.00	
reliability (coefficient alpha):	0.59	
index of reliability:	0.77	
standard error of measurement:	2.42	(4.8%)



Lertap5 full item stats for "Tenacity", created: 1/06/2011.

Summary statistics

number of scores (n):	94	
lowest score found:	116.00	(51.6%)
highest score found:	206.00	(91.6%)
median:	171.00	(76.0%)
mean (or average):	171.23	(76.1%)
standard deviation:	16.23	(7.2%)
standard deviation (as a sample):	16.32	(7.3%)
variance (sample):	266.25	

number of subtest items:	45	
minimum possible score:	45.00	
maximum possible score:	225.00	
reliability (coefficient alpha):	0.89	
index of reliability:	0.95	
standard error of measurement:	5.29	(2.4%)

Quite an improvement, eh? Both reliability figures have gone up.

That for the "tenacity" scale, 0.89, is especially nice. The scale still has 45 items, more than desired, but the chances of ending up with a scale having just 40 items, and a reliability of possibly 0.80, would seem to be good.

How to determine which of the 45 items might be weeded out is something many people would attempt by using **factor analysis**.

Alas, Lertap does not yet have a factor analysis routine. However, it is able to extract the first **principal component** from the interitem correlation matrix. This is done via the "Item scores and correlations" option found under the Run menu. A click here will zip you off to related help topics.

(The **SPSS** package is often used for factor and components analysis. Jump to a topic which discusses importing Lertap data with SPSS.)

The screen snapshot below is an example of the "**IStats**" report created by this option.

Mente2010.xlsx - Microsoft Excel

Lertap5 IStats matrix, last updated on: 1/06/2011.

ID code	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
eigens	9.47	3.24	2.45	2.27	1.91	1.80	1.64	1.51	1.42	1.31	1.22	1.21	1.13
percent	21.0%	7.2%	5.5%	5.0%	4.2%	4.0%	3.6%	3.3%	3.1%	2.9%	2.7%	2.7%	2.5%
p-comp1	0.36	0.56	0.67	0.55	0.63	0.34	0.50	0.55	0.32	0.38	0.51	0.58	0.60

SMC bands

.00:
.10:
.20:
.30:
.40: T45
.50: T1 T10 T11 T18 T19 T21 T22 T24 T26 T29 T35 T41 T43 T44
.60: T2 T4 T6 T7 T8 T9 T12 T13 T14 T15 T17 T23 T25 T27 T28 T30 T31 T33 T34 T36 T38 T39 T40 T42
.70: T3 T5 T16 T20 T32 T37
.80:
.90:

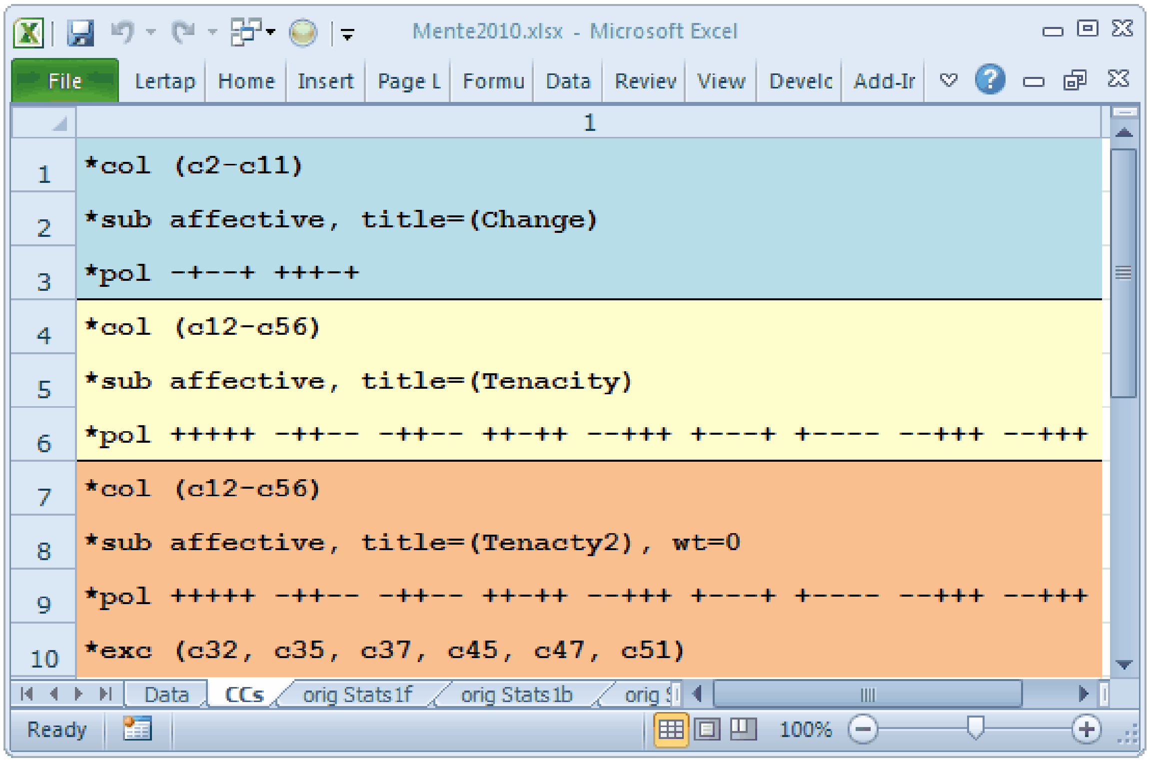
P-Comp1 bands

.00:
.10: T21 T24 T26 T40
.20: T34 T36
.30: T1 T6 T9 T10 T14 T22 T27 T35 T43 T44
.40: T17 T18 T28 T30 T33 T37 T41 T42
.50: T2 T4 T7 T8 T11 T12 T15 T16 T19 T20 T23 T25 T29 T31 T32 T38 T39 T45
.60: T3 T5 T13
.70:
.80:
.90:

Stats1f Stats1b Stats2f Stats2b IStats orig Stats

There are six items whose correlations with the first principal component are below 0.30. Their item ID numbers are seen in the **P-Comp1 bands**.

The following CCs lines include a new subtest, "Tenacty2", with a *exc line which serves to eliminate these six items from the scale (the first item to be omitted is T21, which is located in c32 of the Data worksheet).

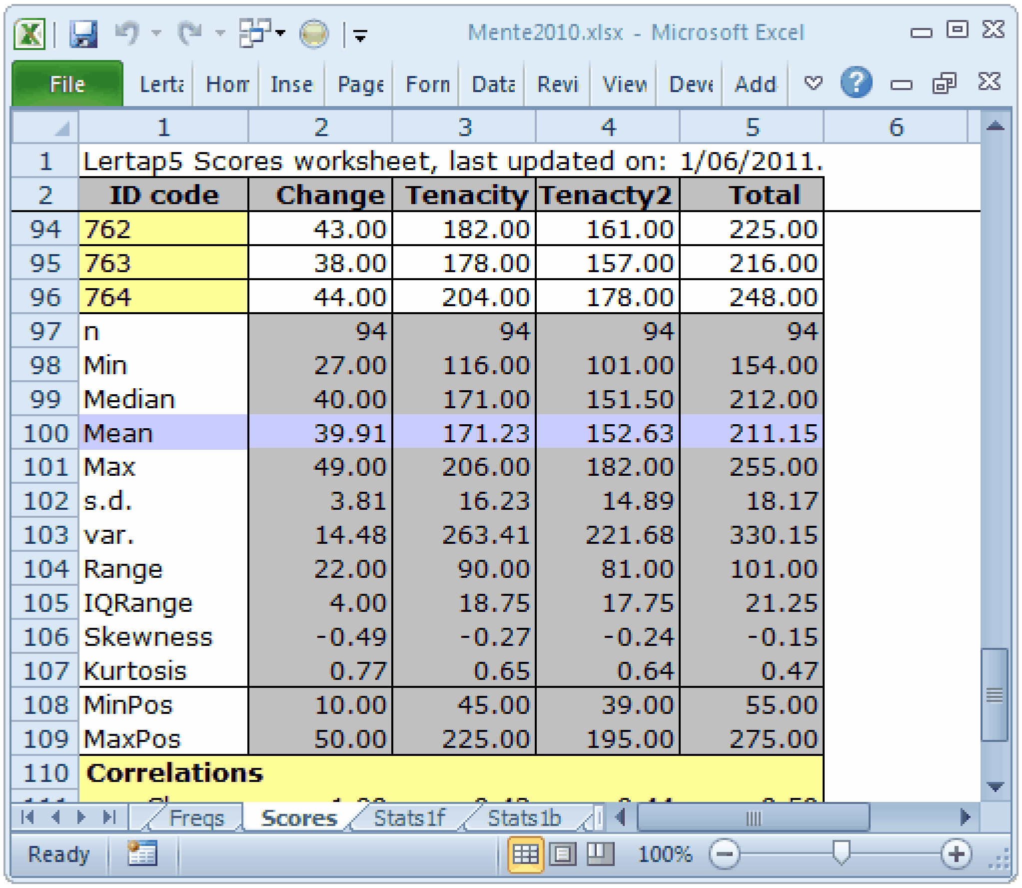


Wouldn't you like to know what the reliability of the new subtest was? We've eliminated six items, now running with a scale having just 39 items, and the reliability has ... has what? Do it yourself. Download the workbook, and give it a go. You'll be surprised if you do; the reliability actually *increases*.

Tidbits:

Why did we call the new subtest "Tenacty2" instead of "Tenacity2"? Because we know that Lertap allows a maximum of 8 characters for the title= assignment. If the title is longer than eight characters, it gets truncated. "Tenacity2" would have shown up as just "Tenacity" in Lertap's various reports.

What's accomplished by using wt=0 on Tenacty2's *sub line? It keeps the new "Tenacty2" score from entering into the Total composite score value in the Scores worksheet:



Lertap5 Scores worksheet, last updated on: 1/06/2011.

ID code	Change	Tenacity	Tenacty2	Total
762	43.00	182.00	161.00	225.00
763	38.00	178.00	157.00	216.00
764	44.00	204.00	178.00	248.00
n	94	94	94	94
Min	27.00	116.00	101.00	154.00
Median	40.00	171.00	151.50	212.00
Mean	39.91	171.23	152.63	211.15
Max	49.00	206.00	182.00	255.00
s.d.	3.81	16.23	14.89	18.17
var.	14.48	263.41	221.68	330.15
Range	22.00	90.00	81.00	101.00
IQRange	4.00	18.75	17.75	21.25
Skewness	-0.49	-0.27	-0.24	-0.15
Kurtosis	0.77	0.65	0.64	0.47
MinPos	10.00	45.00	39.00	55.00
MaxPos	50.00	225.00	195.00	275.00
Correlations				

Principal components analysis is discussed in one of the Lertap "erudite epistles". (It's the "*About eigenvalues, scree tests, and coefficient alpha*" paper; Download it if you'd like. Read about Lertap and principal components here.)

2.15.3 Download

The Mente2010.xlsx workbook may be downloaded by clicking here.

There would be several activities which could be pursued with this dataset.

What was the correlation between the "change" score, and the "tenacity" score? Did people in the pilot sample (n=94) with high "tenacity" scores also have high "change" scores?

Are there some items in the "change" scale with particularly high correlations with the "tenacity" score? (This could be answered by using the tenacity score as an external criterion for the "change scale". There are two examples in this document which would serve as examples: LenguaBIg and The Cook's Tour.)

How about exporting the item scores from the tenacity scale's IStats matrix to a factor analysis program? How many items could you knock out and still keep the scale's reliability figure up around the 0.80 level? (You might use **SPSS** for this. To see how, click here.)

2.16 StuIQ

StuIQ is another real-life example, based on the development of a scholastic **aptitude test** for use with senior high school students.

Two forms of the test were developed, each with 70 items, each having a mixture of multiple-choice and constructed-response items.

Questions investigated: the reliability of each form as a total test, the reliability of each form's multiple-choice items as subtests, the reliability of the constructed-response subtests, **parallel-forms reliability**, and **practice effects**.

A description of the study, "**PFExample1**", is available in this PDF file. It's a relatively complex study as the item responses were coded in a unique manner. Note that the Excel 2003 version of Lertap was used in this study.

A link to the complete dataset is available on the following page.

2.16.1 Download

The **StuIQ.xlsx** dataset is available here.

The corresponding technical guide, PFExample1.pdf, is available here.

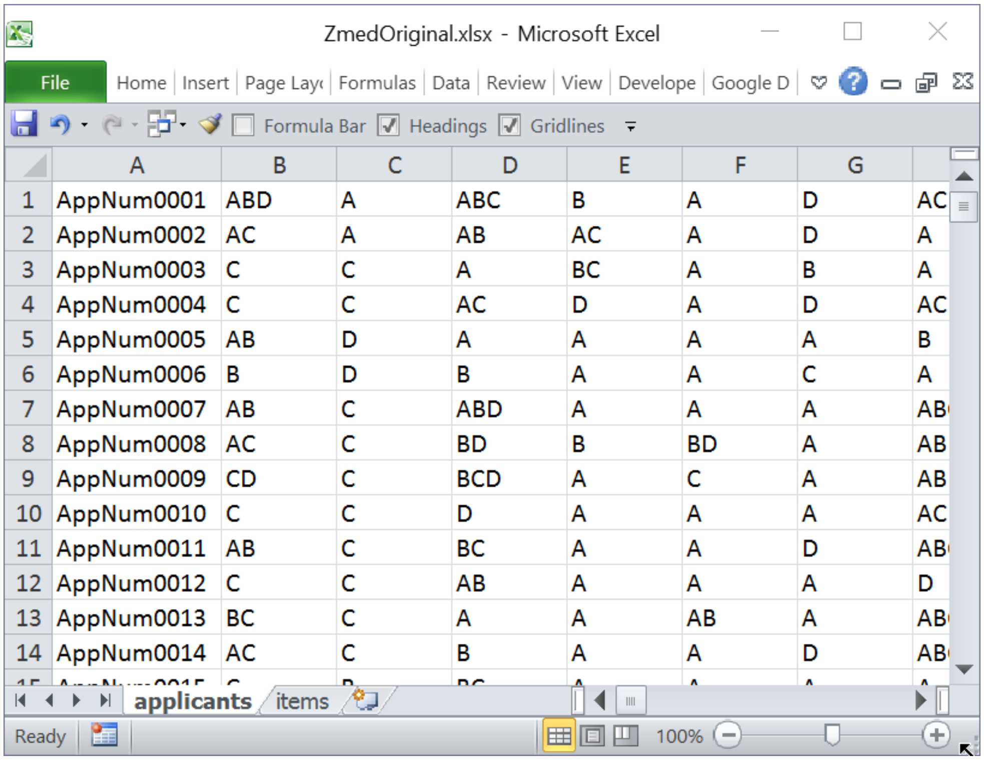
2.17 Zmed

Here's an interesting sample from a major medical school in Europe.

School staff developed a 100-item multiple-choice admissions screening test. The results in the downloadable Excel workbook are from 2,470 applicants who sat the test in early 2015.

This dataset is interesting as many of the items allowed for multiple responses. In fact, to get an item right, on some items students had to choose, for example, option B and option D. Some items even required the selection of three options in order to be marked correct.

Here's a snippet of the original item response worksheet, "applicants".



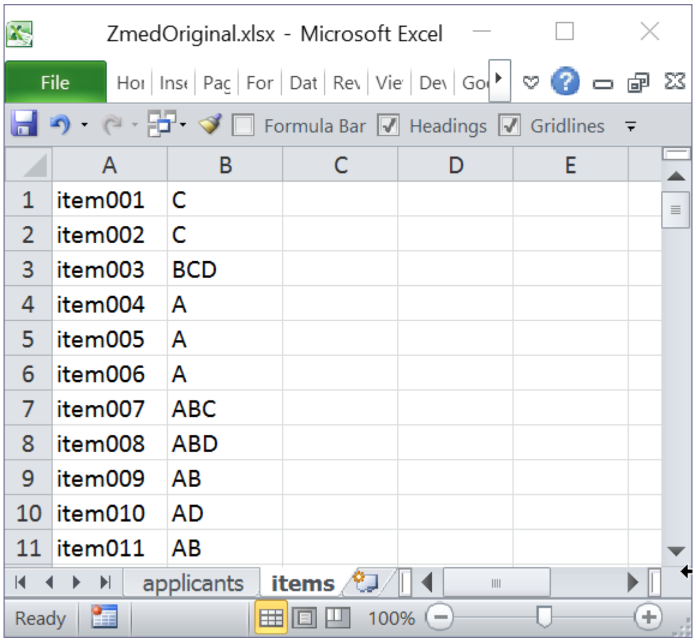
	A	B	C	D	E	F	G	H
1	AppNum0001	ABD	A	ABC	B	A	D	AC
2	AppNum0002	AC	A	AB	AC	A	D	A
3	AppNum0003	C	C	A	BC	A	B	A
4	AppNum0004	C	C	AC	D	A	D	AC
5	AppNum0005	AB	D	A	A	A	A	B
6	AppNum0006	B	D	B	A	A	C	A
7	AppNum0007	AB	C	ABD	A	A	A	AB
8	AppNum0008	AC	C	BD	B	BD	A	AB
9	AppNum0009	CD	C	BCD	A	C	A	AB
10	AppNum0010	C	C	D	A	A	A	AC
11	AppNum0011	AB	C	BC	A	A	D	AB
12	AppNum0012	C	C	AB	A	A	A	D
13	AppNum0013	BC	C	A	A	AB	A	AB
14	AppNum0014	AC	C	B	A	A	D	AB

The screen snapshot above was obtained when using Excel 2010 with its ribbon collapsed. The ribbon's tabs are of course still displayed; they start with File (in green), followed by Home, Insert, and so on. Immediately beneath the tab titles is the "QAT" row of selected controls, including Formula Bar, Headings, and Gridlines. QAT means "quick-action toolbar".

Column A contains an ID code for each applicant. Column B contains the responses of each student to the first item. On the first item, the first applicant, "AppNum0001", selected three options, A, B, and D, while on the second item, s/he selected option A only.

From just looking at the student responses above, it would seem that the first item used multiple responses, as did the third (column D), but not, it would seem, the second item -- column C shows just a single response selection.

Now look at the "items" worksheet. It provides the item "keys".



The name of each item is displayed in column A, with the correct answer, the "keys", found in the adjacent column, B.

Now we can see that the first item did not require multiple responses. The third item did -- in order to get this item correct, a student had to select options B, C, and D.

Alright, we want an item analysis and a test score for each student using Lertap 5.

Can do? Yes, of course, but in its present condition this workbook is quite far from being ready for Lertap.

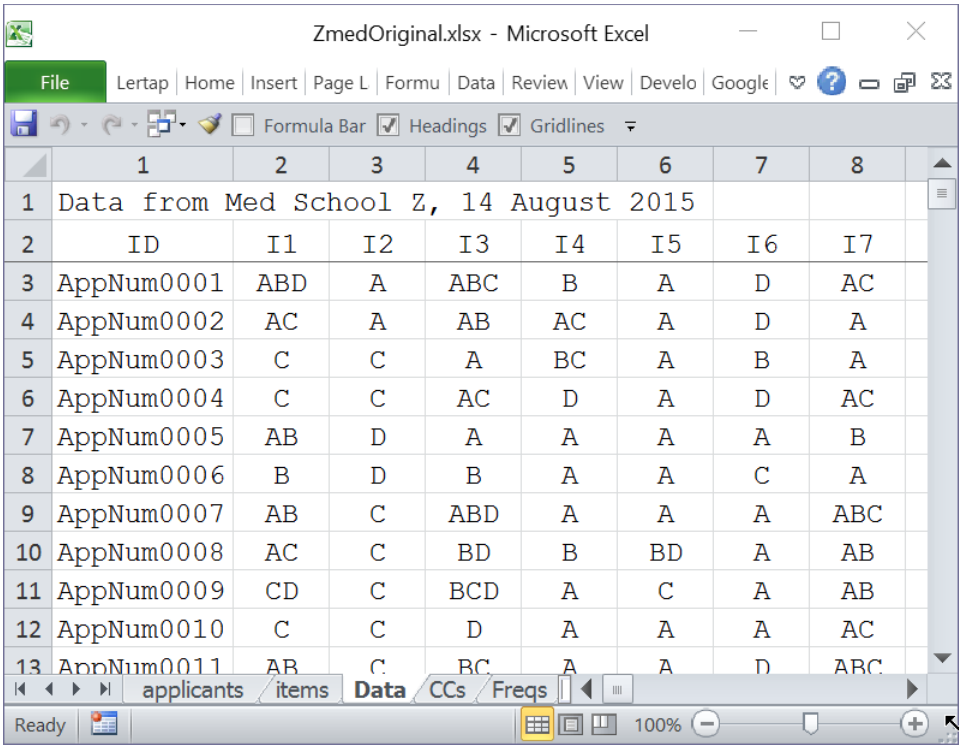
Lertap wants a Data worksheet, and a "CCs" worksheet too. Each of these two worksheets requires a special format and, most importantly, since Lertap 5 is reluctant to process items which allow for multiple responses, it will be necessary to recode each of the items so that the Data worksheet has item responses which are just a single character.

Page forward if you will.

2.17.1 Making ready

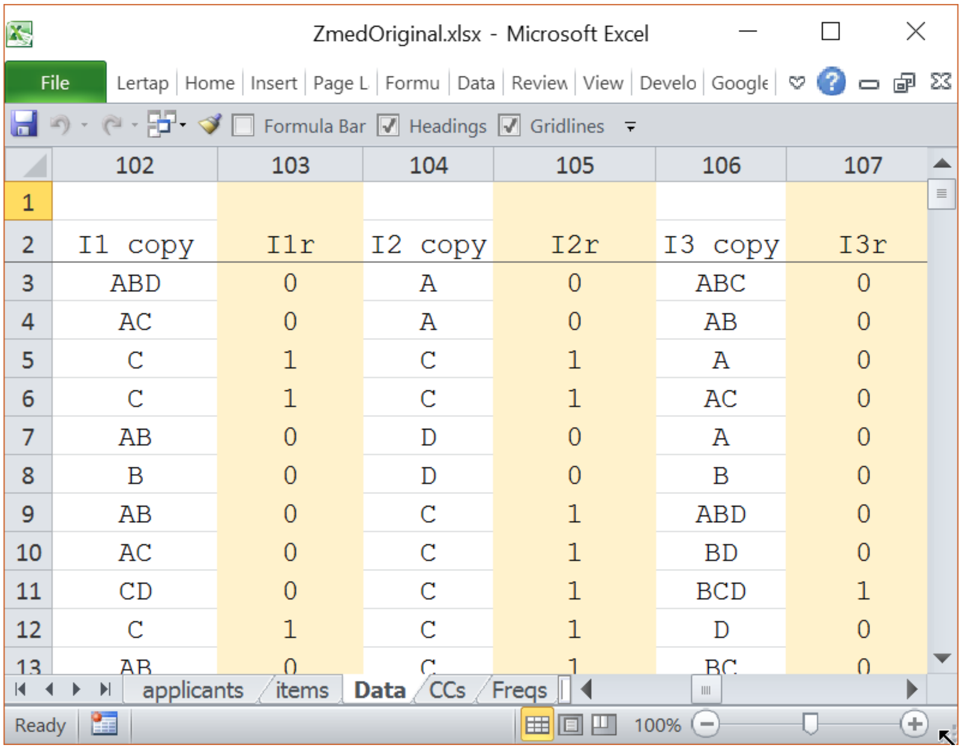
The requirements of a Lertap Data worksheet are set out in this topic.

The screen snapshot below displays the first eight columns of an appropriate Data worksheet. Note that the row of ribbon tab titles now includes a Lertap tab after the File tab. This indicates that the Lertap 5 workbook, Lertap5.xlsm, has been opened. Lertap prefers to have column headings which are numbers instead of letters, and it has accordingly changed the column headers, as you can see. What was column "A" is now column "1". (Refer to this topic for more information if wanted, and look for "Ref. style".)



	1	2	3	4	5	6	7	8
1	Data from Med School Z, 14 August 2015							
2	ID	I1	I2	I3	I4	I5	I6	I7
3	AppNum0001	ABD	A	ABC	B	A	D	AC
4	AppNum0002	AC	A	AB	AC	A	D	A
5	AppNum0003	C	C	A	BC	A	B	A
6	AppNum0004	C	C	AC	D	A	D	AC
7	AppNum0005	AB	D	A	A	A	A	B
8	AppNum0006	B	D	B	A	A	C	A
9	AppNum0007	AB	C	ABD	A	A	A	ABC
10	AppNum0008	AC	C	BD	B	BD	A	AB
11	AppNum0009	CD	C	BCD	A	C	A	AB
12	AppNum0010	C	C	D	A	A	A	AC
13	AppNum0011	AB	C	BC	A	A	D	ABC

Lertap has a "recoder" which may be used to create item response columns that have just a single character. As an example, the correct answer to the first item, now called "Item 1", is C. We'll recode it by using the "recoder"; then we'll also use the recoder on item I2, correct answer C, and I3, correct answer BCD. Behold:



ZmedOriginal.xlsx - Microsoft Excel

	102	103	104	105	106	107
1						
2	I1 copy	I1r	I2 copy	I2r	I3 copy	I3r
3	ABD	0	A	0	ABC	0
4	AC	0	A	0	AB	0
5	C	1	C	1	A	0
6	C	1	C	1	AC	0
7	AB	0	D	0	A	0
8	B	0	D	0	B	0
9	AB	0	C	1	ABD	0
10	AC	0	C	1	BD	0
11	CD	0	C	1	BCD	1
12	C	1	C	1	D	0
13	AB	0	C	1	BC	0

applicants / items / Data / CCs / Freqs

Column 102 has a copy of the original responses to the first item, while column 103 shows the recoded responses using a column header of "I1r", for "I1 recoded".

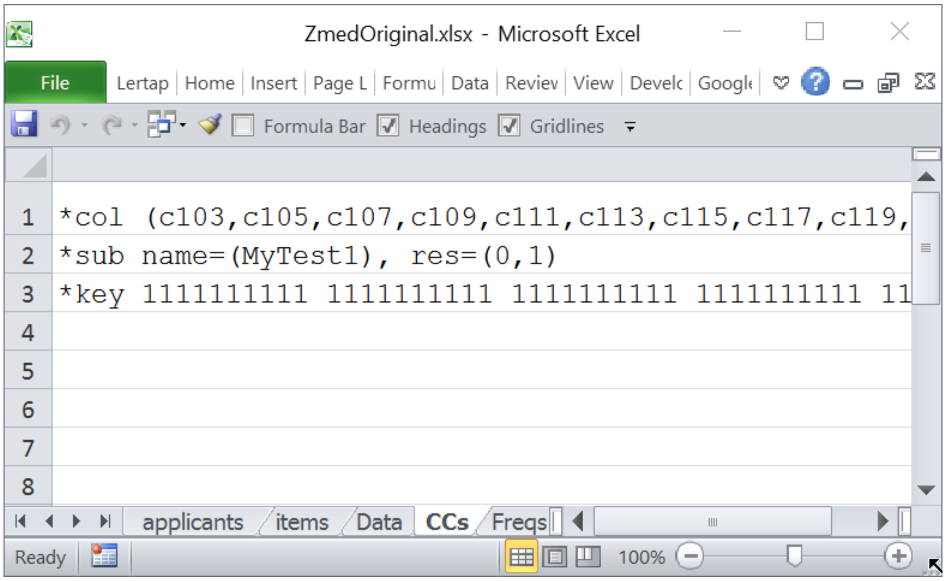
If a student got I1 correct by selecting option C, her/his recoded response was a 1 (one), with any other response, or responses, becoming a 0 (zero). This is "binary" item coding; it could also called "right-wrong" coding, or "dichotomous" item coding.

The right answer to the second item, I2, was also C. The third item's correct response was BCD. Looking down the I3r column, column 107, we see that only one of the students shown above got the item correct.

We applied the recoder 100 times in order to get binary item responses for each of the items.

Was this fun? Not! It's easy to use the recoder, but this is a job that begs for a special macro. Such macros involve a bit of computer programming, but they're often fairly straightforward to develop. Please refer to this topic for more information.

So much for getting the Data worksheet ready, now for the CCs sheet:



The *col line refers to the location of each of the recoded items. The res=(0,1) assignment on the *sub line tells Lertap that binary item response codes are to be found in these columns, while the *key line simply says that the right answer to each recoded item was 1 (one).

If you download this dataset and take it for a spin by using the "Interpret" and "Elmillon" options, you'll find that the test had excellent properties, so good that the medical school could cut back on the number of test items, resulting in a shorter test but one probably still having adequate measurement precision for the purpose of an admissions exam.

In case you're new to Lertap 5, and do not have an idea of how the "Interpret" and "Elmillon" options are used, an adequate reference might be this one. It's possible to nudge Lertap 5 into a "Production mode" which simplifies how Lertap 5 options are called into action -- click here to read about it.

Page forward.

2.17.2 Download

A simple click here will initiate a download of the "**ZmedOriginal.xlsx**" Excel workbook ready to work with Lertap5.

A click here will (should) lead to a download of a file called "**Zmed-IScores.csv**". This file is a "comma-separated values" file containing a header row (with item titles) followed by rows of item scores, one for each person, with each item scored (0,1) -- zero if the person/student got the item wrong, and one if s/he got the item right. (If an item response was missing, it will be coded as a zero.) csv files are often used with an R package, and/or with a data analysis system such as JASP.

Once the ZmedOriginal.xlsx workbook has been copied to your computer, get Lertap 5 going by opening the Lertap5.xlsm file. Once it's open, return to the ZmedOriginal.xlsx workbook and use the "Interpret" and "Elmillon" options to get results. Note the excellent reliability (coefficient alpha) and quite nice distribution of test scores.

As to suggested exercises, well, this was an admissions test, and as such we'd normally expect a cut-off score to be applied. The Mastery= assignment is used to set cut-off scores. What you could do is experiment with various Mastery= settings to see how they affect the pass rate (number of students passing the test), and the "Estimated number of incorrect classifications".

Suppose the medical school had places for just 500 students; where should the cut-off score be placed so that only 500 of the 2,470 applicants pass?

Try reducing the number of items and see how that impacts test reliability (coefficient alpha) and the "Estimated number of incorrect classifications". Could we bring the test down to a 70-item exam and still have adequate reliability while not raising the "Estimated number of incorrect classifications" by too much?

How to reduce the number of items? There are several ways, the easiest being to use a *exc line in the CCs sheet, after the *key line. You could do this rapidly by copying the *col line, pasting the copy after *key, changing "*col" to *exc" and then simply deleting some of the column numbers found in the parentheses after what will be *exc, perhaps by starting from the right end (keep the first items, delete only some of the last ones).

Intrigued by these questions but not sure how to go about answering them? Try dropping an email to lertap5@gmail.com.

2.18 Test13

This dataset stems from work undertaken decades ago in an Asian country, well before electronic calculators became commonplace.

It has been included in these sample datasets largely because the original test items are available for study. While this is also true of the FIMS dataset, in this case all items are straightforward multiple-choice cognitive questions, each with four options. In such cases Lertap5's CCs lines are particularly simple to create and (hopefully) understand.

An additional feature is the inclusion of a categorical variable, "Zone", leading to the possible use of Lertap5's group differences options. It may be seen that students in one or two of the four country zones appeared to be weak.

A presentation/discussion of Test13 items and an analysis of results, suggested as perhaps a useful reference for use with test and measurement classes is here.

The next page has links to a variety of downloadable files. The links lead to Excel workbooks with student responses, and to data sets and reports involving both **CTT** (classical test theory, Lertap5's forte) and **IRT** (item response theory).

Note: the country the test was originally used in has here been disguised as Australia.

2.18.1 Downloads

The actual multiple-choice items used in this very basic 13-item mathematics test may be seen here.

An Excel workbook with item responses from **2,976** students, "**Test13.xlsx**", may be downloaded by clicking here. It includes a "zone" code for each of the students which indicates if they were from the north, south, east, or west of the country. The zone code can be used to breakout score and item responses using two standard Lertap5 options (breakout test scores, and breakdown item responses).

A data subset with responses from the **2,727** students who gave answers to at least 8 of the 13 items is available here. This subset excludes 249 students who failed to answer 5 or more of the 13 items. Unanswered questions are coded as 9 in Test13 data.

A presentation/discussion of Test13 items and an analysis of results, suggested as perhaps a useful reference for use with test and measurement classes is here.

A comma-separated values ("csv") file of item scores may be downloaded from this link. This file has responses from those **2,584** students with test scores in the range of 2 to 12 (students with extreme scores of 0, 1, and 13 have been omitted). This file is meant to be used with R and RMD scripts, such as Rasch-Analysis-Tam.Rmd. The name of the file, Omega-IScores.csv, relates to this topic. The Excel workbook corresponding to this subset of 2,584 students may be found here.

An item analysis report created by using the above-mentioned script, Rasch-Analysis-TAM.Rmd, is found in this pdf file; the data used in the analysis came from the 2,584 students having test scores ranging from 2 to 12.

The original source for Test13 is "*A Course on Test and Item Analysis*" by Prof. Margaret Wu. Chapter 7 introduces Test13 and discusses how to access a DAT file having the original item responses and associated data. The file's name is Maths13.dat.

Comments on using R with Lertap5 are found here. Lertap5 has its own Rasch module; read about it here.

Two of the 13 test items performed somewhat poorly, having weak discrimination. This Excel workbook is set up with two Lertap5 subtests, one with all 13 items, and one with just the best-performing 11 items. Running Lertap5's Interpret and Elmillon options will create Stats1b and Stats2b worksheets which may be compared; a look at the Stats1f and Stats2f worksheets is also worthwhile (compare the reliability and standard error of measurement values seen in the two worksheets - results are a bit better without the two weak items).

3 The Cook's Tour

The "Cook's tour" is an introduction to Lertap which first appeared in Chapter 2 of the manual.

An abbreviated version of the tour appears in the online Lertap help system, which you can make your way to here.

In this website we have both amalgamated and updated the tour. It begins here ...,

Ready to get rolling?

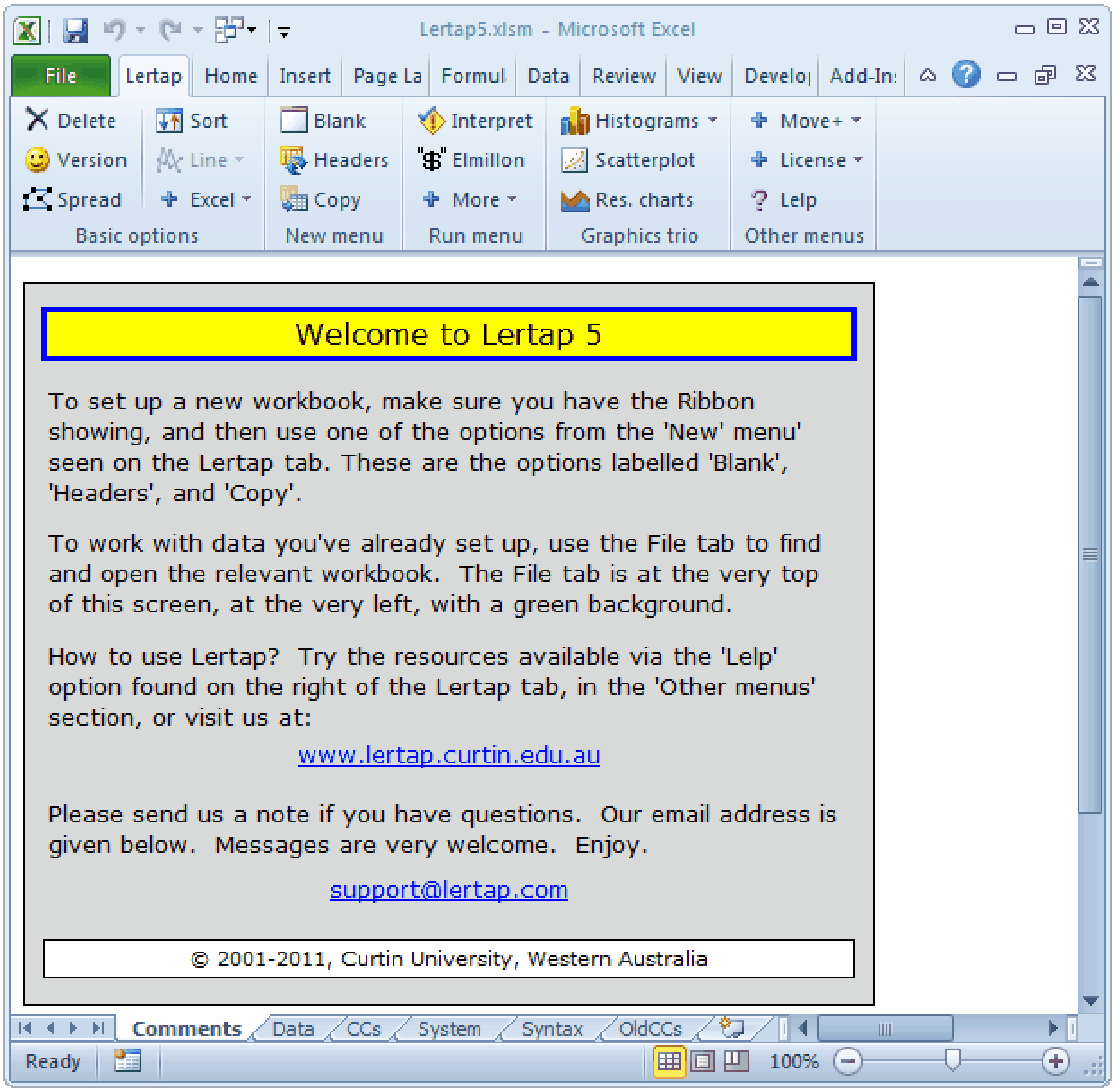
Open Lertap5.xlsm

Find the Lertap5.xlsm file on your computer, and open it.

This file contains **"macros"**. Lots of them. The macros contain the computer programming code 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.xlsm, 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.

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



Here are some things to note at this point:

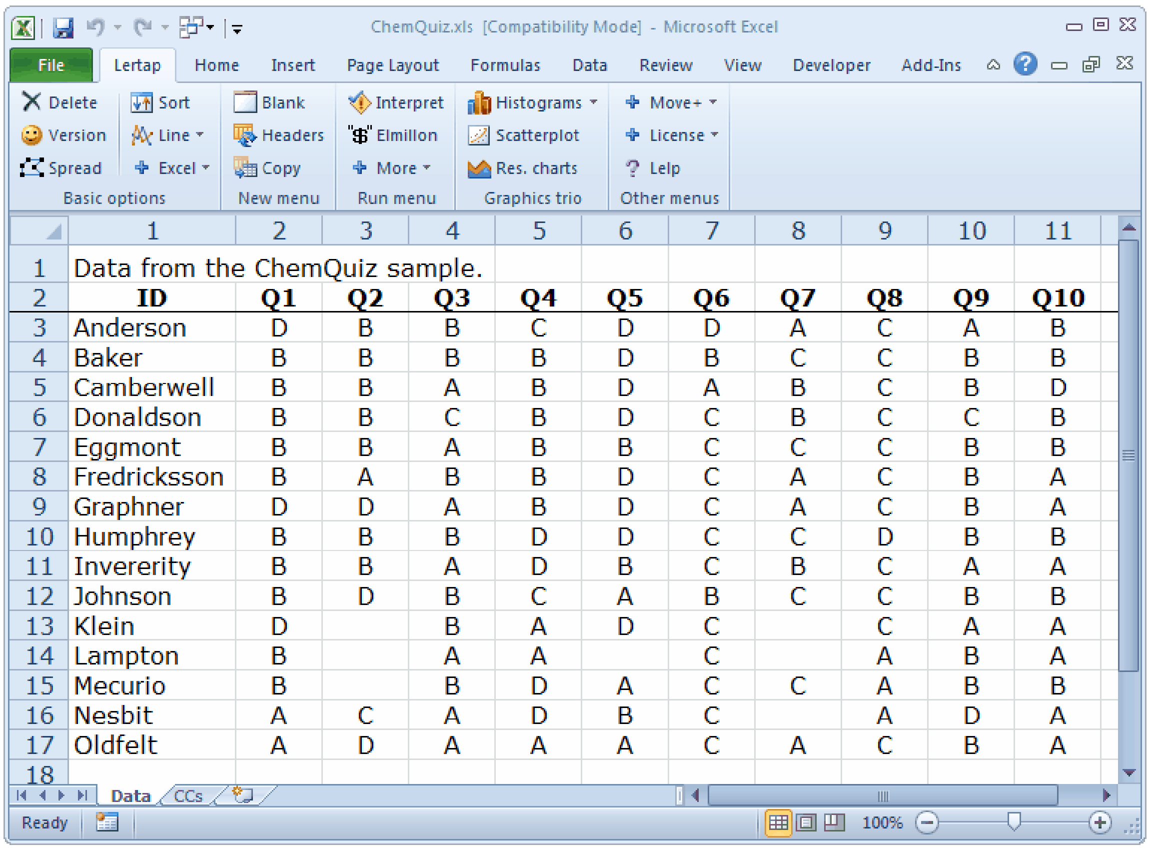
1. The Excel **ribbon** is on display. It has 11 visible tabs at the top, ranging from File to Add-Ins (your ribbon may have different tabs).

2. The **Lertap tab** is open. It shows five groups of icons, from "Basic options" to "Other menus".

3. The name of the workbook is seen at the very top: Lertap 5.xlsm — Microsoft Excel. The icons to the left of the name are said to be on "The quick-access toolbar", or **QAT**. You can put your own favourite options on the QAT quite easily; search Excel help to find out how.

4. The workbook has six (6) worksheets. The names of each worksheet are displayed on the tabs at the bottom of the screen, going from "Comments" on the left, to "OldCCs" on the right.

Your screen will differ when you're working with your own workbook. For example, here's how a chemistry teacher's workbook looked when he set about processing student responses to a short quiz:

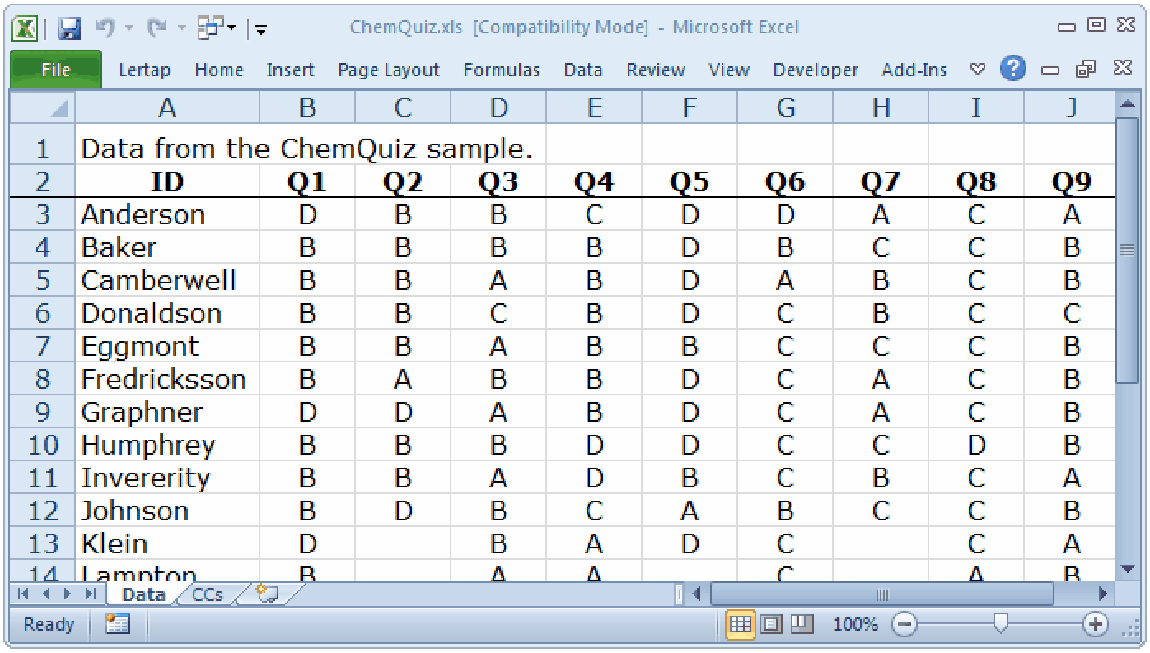


	1	2	3	4	5	6	7	8	9	10	11
1	Data from the ChemQuiz sample.										
2	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
3	Anderson	D	B	B	C	D	D	A	C	A	B
4	Baker	B	B	B	B	D	B	C	C	B	B
5	Camberwell	B	B	A	B	D	A	B	C	B	D
6	Donaldson	B	B	C	B	D	C	B	C	C	B
7	Eggmont	B	B	A	B	B	C	C	C	B	B
8	Fredricksson	B	A	B	B	D	C	A	C	B	A
9	Graphner	D	D	A	B	D	C	A	C	B	A
10	Humphrey	B	B	B	D	D	C	C	D	B	B
11	Invererity	B	B	A	D	B	C	B	C	A	A
12	Johnson	B	D	B	C	A	B	C	C	B	B
13	Klein	D		B	A	D	C		C	A	A
14	Lampton	B		A	A		C		A	B	A
15	Mecurio	B		B	D	A	C	C	A	B	B
16	Nesbit	A	C	A	D	B	C		A	D	A
17	Oldfelt	A	D	A	A	A	C	A	C	B	A
18											

In this screen snapshot,the Excel ribbon is still on display, and the Lertap tab is still open. The workbook's name is ChemQuiz.xls. The workbook presently has two worksheets, "Data" and "CCs".

The [Compatibility Mode] text after the workbook name reflects the fact that the workbook's extension is xls, not xlsx. The xls extension for a workbook was used for many years by Excel, becoming outdated when Office 2007 emerged. Compatibility mode means that Excel has recognised the old workbook format, and made adjustments to how it runs in order to accommodate the former format.

A note about the "**reference style**" setting in Excel: column headers may be numbers, as shown above, or letters, as shown below.



	A	B	C	D	E	F	G	H	I	J
1	Data from the ChemQuiz sample.									
2	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
3	Anderson	D	B	B	C	D	D	A	C	A
4	Baker	B	B	B	B	D	B	C	C	B
5	Camberwell	B	B	A	B	D	A	B	C	B
6	Donaldson	B	B	C	B	D	C	B	C	C
7	Eggmont	B	B	A	B	B	C	C	C	B
8	Fredricksson	B	A	B	B	D	C	A	C	B
9	Graphner	D	D	A	B	D	C	A	C	B
10	Humphrey	B	B	B	D	D	C	C	D	B
11	Inverity	B	B	A	D	B	C	B	C	A
12	Johnson	B	D	B	C	A	B	C	C	B
13	Klein	D		B	A	D	C		C	A
14	Lampton	B		A	A		C		A	B

Excel's reference style setting determines whether numbers or letters are used as column headers. Lertap prefers to use numbers. When Excel is using numbers for column headings, it is said to be using the "R1C1" reference style; when it's using letters, it's in the "A1" reference style.

R1C1 and A1 indicate how the first cell on a worksheet is reference. R1C1 means row 1, column 1, while A1 means column A, row 1.

You can change the reference style quickly if the Lertap tab is showing. Go to this page to see how; the option to use is called "Ref. style". If the Lertap tab is not showing, the reference style is changed by using the File tab, then Options, then Formulas.

Next?

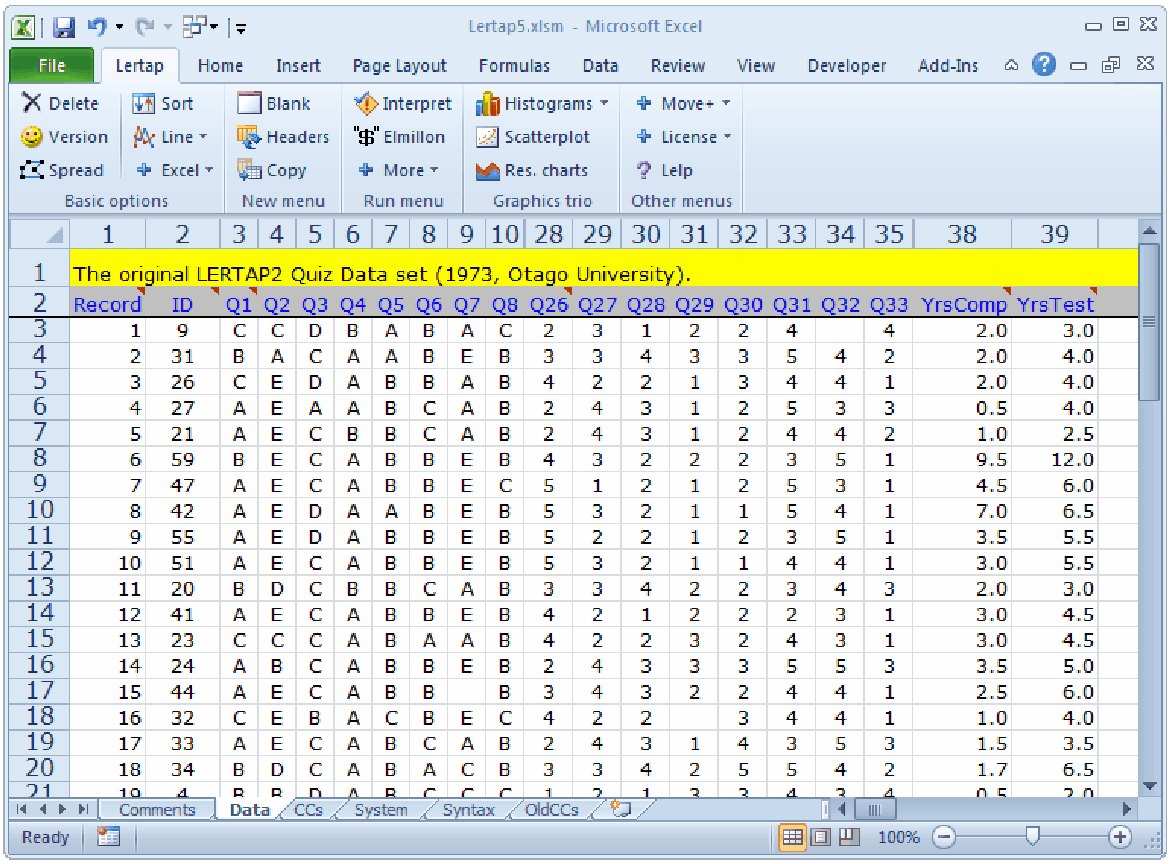
Notes:

Excel will also work with "OpenDocument" spreadsheets, ones having an extension of "ods". A popular spreadsheet system using the OpenDocument format is available from OpenOffice.org, a competitor to Microsoft Office. **Beware**: should you install OpenOffice on your computer, the Lertap5.xlsm workbook will not work with it; the macro language used by OpenOffice is not entirely compatible with that used by Lertap.

3.1 The Lertap5.xlsm worksheets

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

The **Comments** sheet is the first sheet to show when the Lertap5.xlsm file is opened. It isn't much, just a welcoming title page (see it here). 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.lertap5.com/lertap/.



	1	2	3	4	5	6	7	8	9	10	28	29	30	31	32	33	34	35	38	39
1	The original LERTAP2 Quiz Data set (1973, Otago University).																			
2	Record	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	YrsComp	YrsTest
3	1	9	C	C	D	B	A	B	A	C	2	3	1	2	2	4		4	2.0	3.0
4	2	31	B	A	C	A	A	B	E	B	3	3	4	3	3	5	4	2	2.0	4.0
5	3	26	C	E	D	A	B	B	A	B	4	2	2	1	3	4	4	1	2.0	4.0
6	4	27	A	E	A	A	B	C	A	B	2	4	3	1	2	5	3	3	0.5	4.0
7	5	21	A	E	C	B	B	C	A	B	2	4	3	1	2	4	4	2	1.0	2.5
8	6	59	B	E	C	A	B	B	E	B	4	3	2	2	2	3	5	1	9.5	12.0
9	7	47	A	E	C	A	B	B	E	C	5	1	2	1	2	5	3	1	4.5	6.0
10	8	42	A	E	D	A	A	B	E	B	5	3	2	1	1	5	4	1	7.0	6.5
11	9	55	A	E	D	A	B	B	E	B	5	2	2	1	2	3	5	1	3.5	5.5
12	10	51	A	E	C	A	B	B	E	B	5	3	2	1	1	4	4	1	3.0	5.5
13	11	20	B	D	C	B	B	C	A	B	3	3	4	2	2	3	4	3	2.0	3.0
14	12	41	A	E	C	A	B	B	E	B	4	2	1	2	2	2	3	1	3.0	4.5
15	13	23	C	C	C	A	B	A	A	B	4	2	2	3	2	4	3	1	3.0	4.5
16	14	24	A	B	C	A	B	B	E	B	2	4	3	3	3	5	5	3	3.5	5.0
17	15	44	A	E	C	A	B	B		B	3	4	3	2	2	4	4	1	2.5	6.0
18	16	32	C	E	B	A	C	B	E	C	4	2	2		3	4	4	1	1.0	4.0
19	17	33	A	E	C	A	B	C	A	B	2	4	3	1	4	3	5	3	1.5	3.5
20	18	34	B	D	C	A	B	A	C	B	3	3	4	2	5	5	4	2	1.7	6.5

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 the first version of Lertap, a version which appeared in 1973. (In the snapshot above, only the first eight cognitive questions, Q1 through Q8, are showing.)

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 R3C34 (row 3, column 34) for example, appears to have nothing in it. The respondent whose answers were recorded in row 3 of the Data worksheet did not answer Q32.

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.

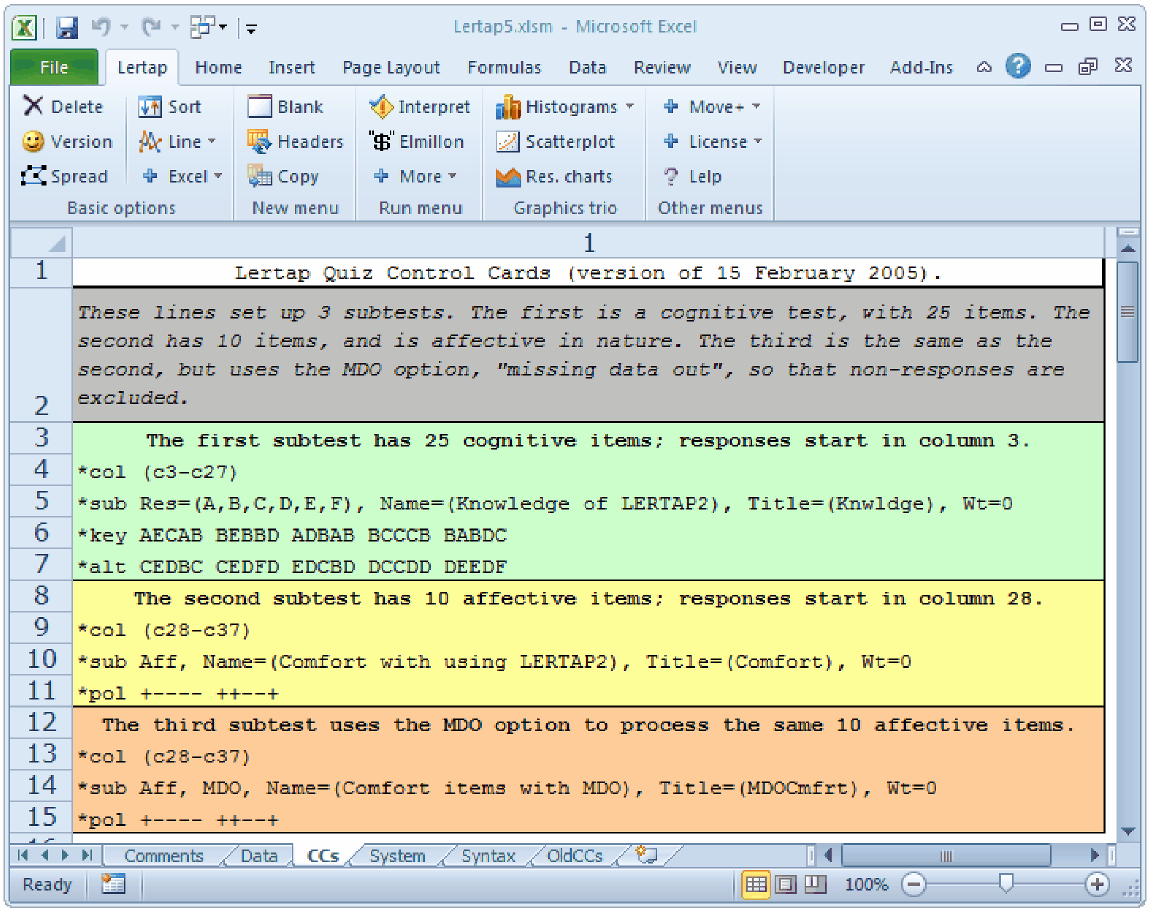
Before looking at the next sheet, CCs (for "Control Cards"), it will be worthwhile to summarise why we'd want to use Lertap to analyse the results seen in the Data worksheet. We could think of some "research questions" which we'd like to have answers to.

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. (Please see the **Tidbits** below. This particular CCs worksheet is somewhat complex, using lines and options which are by no means always required; for example, the *alt line is not often used; it's not common to have a workbook which has both cognitive and affective item responses.)



The **CCs** sheet shown above has 15 rows of information. Only rows which begin with an asterisk, * , are actually used by Lertap; the other rows are simply comments. In this example, rows 1, 2, 3, 8, and 12 are comments.

Rows 4 through 7 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, with Res=(A,B,C,D,E,F), 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.

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; this item used just the first 3 of the possible response letters, which would be A B C. The correct answer to the second item was E, and this item used the first 5 of the possible response letters, A B C D E. The actual items are seen below.

(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

The entries on the *key, *alt, and *pol lines have been grouped in sets of five, with a space between each set. It is not necessary to group by fives. The *key line, for example, could have been

*key AECABBEBBDADBABBCCCBBABDC

The *col (c28-c37) line in row 9 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.Two examples of affective items similar to those actually used are shown below; question (26) was a positive statement, while (27) was negative.

(26) Lertap seems powerful, but simple to use.

1 strongly disagree

2 disagree

3 undecided

4 agree

5 strongly agree

(27) The Lertap manual is lacking detail in many important areas.

1 strongly disagree

2 disagree

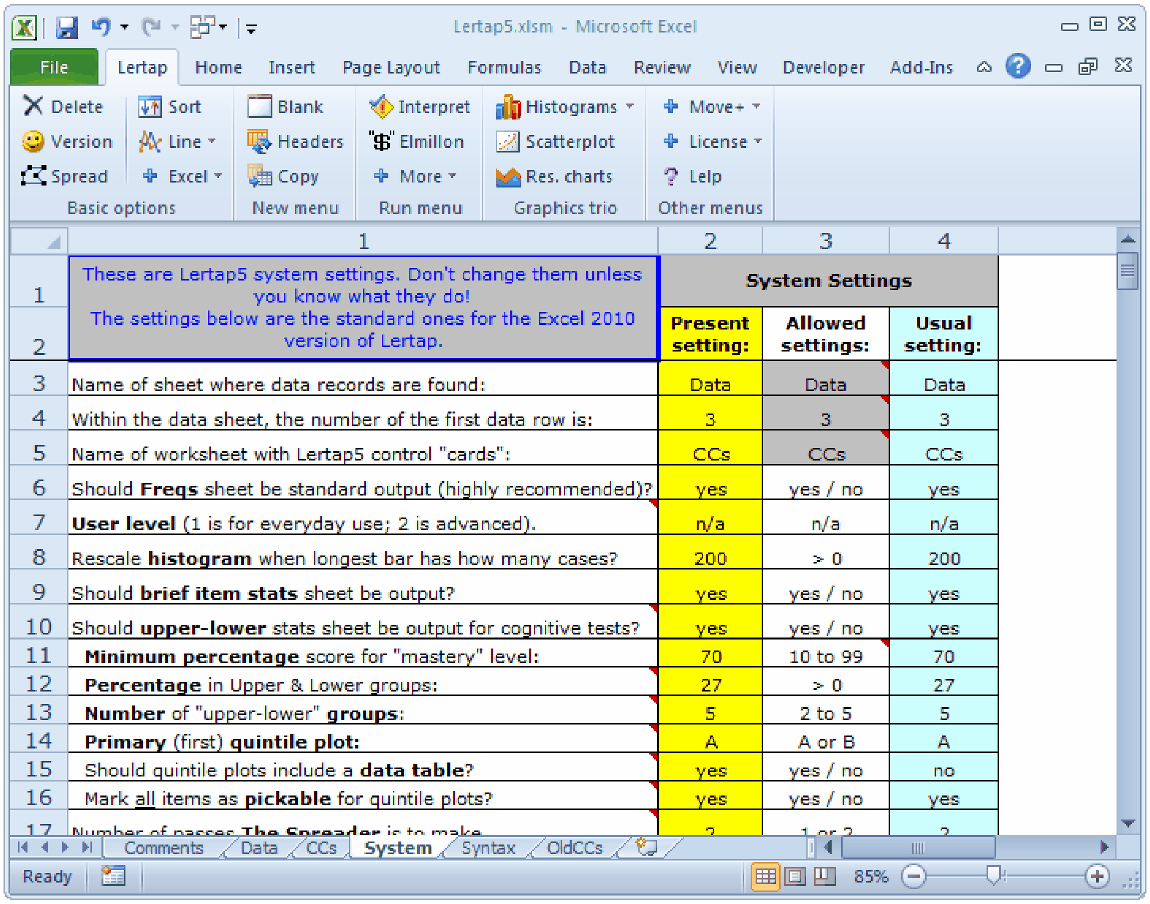
3 undecided

4 agree

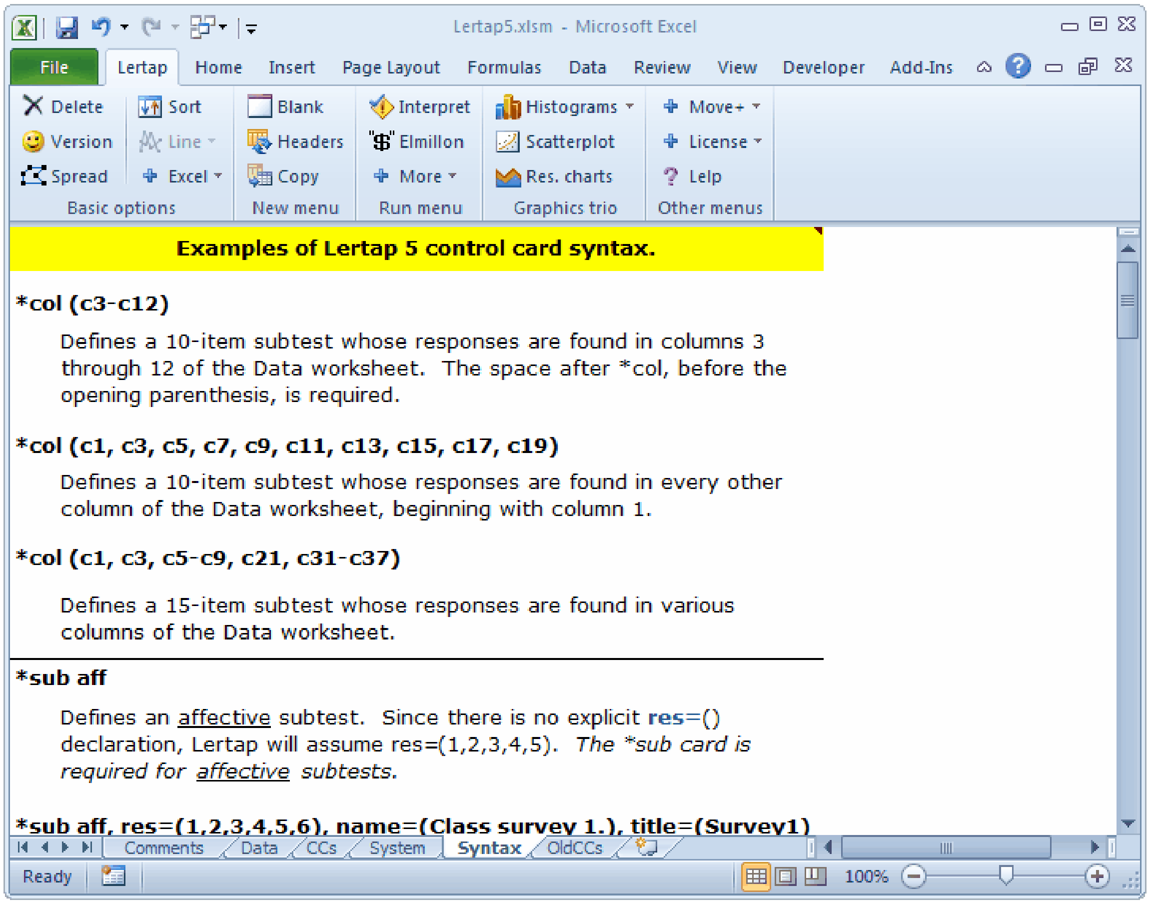
5 strongly agree

In these little examples, note how the cognitive items, (1) and (2) above, use letters as **response codes**, while the affective items, (26) and (27) use digits. In current versions of Lertap 5, an item may use as many as 26 responses codes. Cognitive items do not have to use letters for their response codes—digits are fine. But, having said this, it is important to note that, unless you say otherwise by using Res= on a *sub line, Lertap assumes that cognitive items use A B C D as response codes, and that affective items use 1 2 3 4 5. These are the **default** Res= settings. In this example, an Res= declaration was required for the first subtest (a cognitive subtest), but not for the second and third subtests (affective subtests).

Finally, the *col (c28-c37) line in row 14 signals to Lertap that there's a third subtest to be processed. In fact, this subtest consists of the same items found in the second subtest (row 14 is the same as row 9), but now the items are being scored in a different fashion, using the MDO option. Having items scored in different ways is not too uncommon; in mastery testing, for example, the same *col, *sub, and *key lines may be used repeatedly, with different Mastery= settings on the various *sub lines used to examine the effects of using different cutscores.



The **System** worksheet is where Lertap stores a variety of important settings. Much more about this worksheet may be found in another Lertap help file.



The fifth of the 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.

The last worksheet, **OldCCs**, may be useful to veteran Lertap 5 users, people who may have started to use Lertap 5 and its printed manual when it first appeared in the year 2001. At that time the CCs worksheet was a bit different to the one seen now.

Summary

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

Of these, 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 Lertap Excel workbook looks like, and they are used as the main ingredients in the "Cook's tour".

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

Tidbits:

The CCs worksheet in the Lertap5.xlsm workbook is perhaps unnecessarily complex for beginning Lertap users. Much more elementary examples start here. The recommended reference for boning up on CCs worksheets and their control lines is the on-line help system.

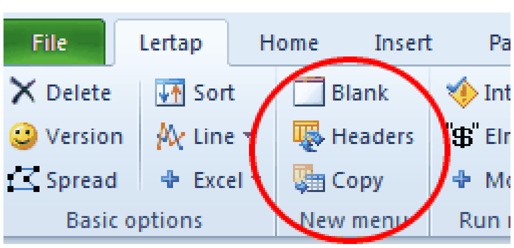
3.2 Setting up a Lertap workbook

If you have opened the Lertap5.xlsm workbook there on your computer, and if you follow the instructions in this section, you'll begin to roll through the "**Cook's Tour**" of Lertap, interacting with your computer and being gobsmacked by Lertap and Excel.

A "**Lertap workbook**" is any workbook which has at least two worksheets, one named Data, and one named CCs. A Lertap workbook must have these two worksheets. It may have other worksheets too, but, to repeat, it has to have a worksheet called Data, and it must have a worksheet called CCs.

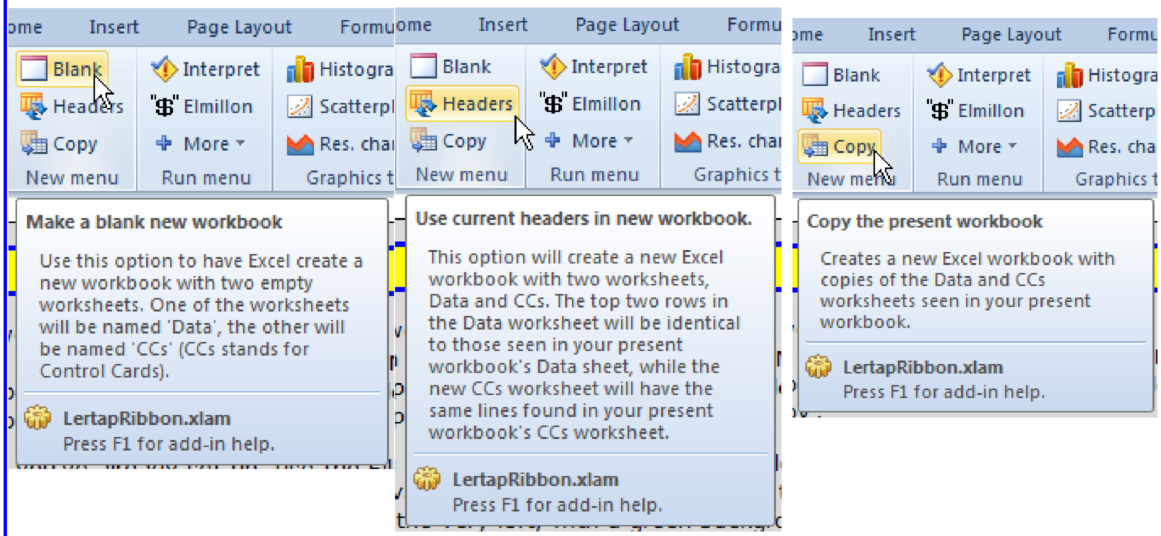
The Data sheet will (must!) always have general information in row 1 and row 2. Row 1 usually has a title of some sort; this may be anything you want. Row 2 will have the column headers. Data records begin in row 3.

The easiest way to set up a new Lertap workbook is to use options from the "New menu" on the Lertap tab:



To find out what the three options in the New menu do, let your mouse hover above them. Here's what you might see:

Blank	Headers	Copy



Two of these options, Headers and Copy, make reference to the "present workbook". This is nothing but the active workbook, the one you're working on at the moment. **Beware**: if the present workbook does not have Data and CCs worksheets, these options may fail.

Let's suppose that the active workbook is Lertap5.xlsm. A click on **Copy** will create a new Lertap workbook.

Try it. Make Lertap5.xlsm the active workbook.

Mouse over to the New menu on the Lertap tab.

Click on **Copy**

Your new workbook will have just two worksheets, copies of the Data and CCs sheets seen in the Lertap5.xlsm file. Excel probably named the new workbook “Book1”, or “Book2”, or something like that.

So, 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 the various lines of job definition statements? Great! Good on you.

The Cook's Tour

Okay. Let's put Lertap through some of its hoops. Come on our "Cook's Tour" of Lertap by cruising through the following topics.

Tidbit:

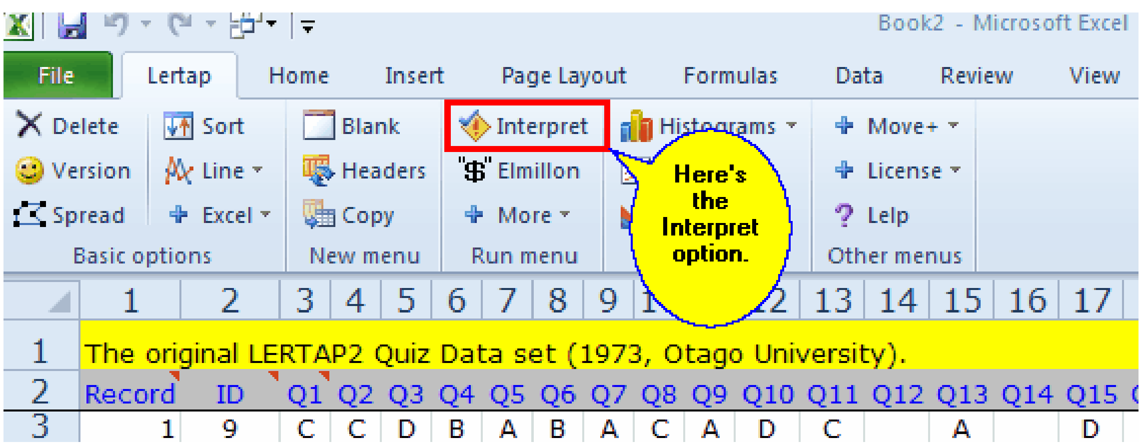
There are several ways to prepare data for Lertap. Find out more now.

3.3 Interpret CCs lines

Our Cook's Tour begins by asking Lertap if it can understand those ever-so-nifty, eye-catching lines found in the CCs worksheet.

We assume that you have used the Copy option on the New menu in order to get set up with some data to play with. If you haven't, please double back to the previous topic, follow the instructions there, and then come back here. You need some data to work with, and the previous topic shows you how to get some, and quickly.

Click on **Interpret**



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 asks Excel to add some new worksheets to the workbook. For each of the *col lines found in the CCs sheet, Lertap creates a new "**Sub**" sheet. Once it's done this, it then creates a single "**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. Is there a limit to the number of worksheets? Sure. To get a better idea as to what the limit is, ask the internet. It depends on the amount of memory available to Excel, but should be on the order of 250, very possibly much more.

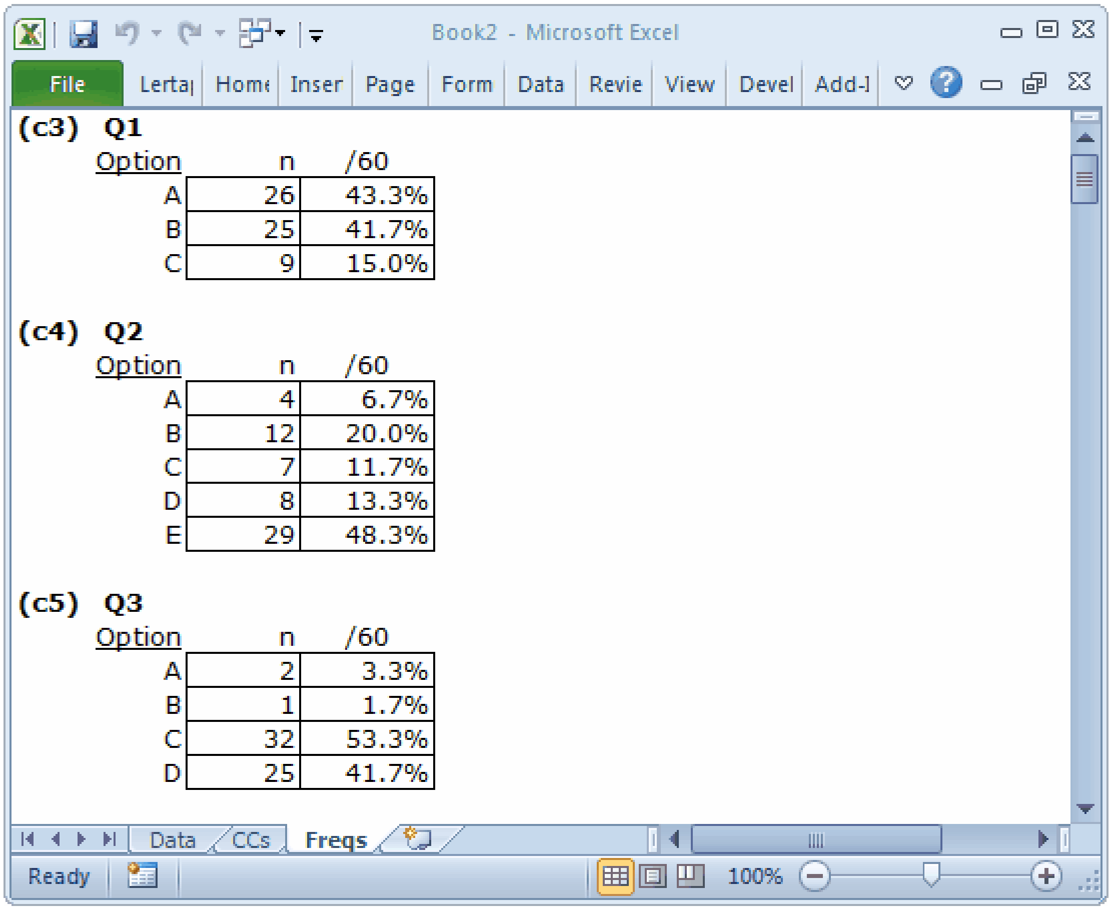
Worksheets which are not used very much, or which contain data used only by the workbook's macros, may be hidden easily: right-click on the worksheet's tab at the bottom of the Excel screen, and select the hide option. Should you try this, you may find an option to "unhide" worksheets if the workbook has some hidden ones. Lertap always hides the Sub worksheets automatically.

The Sub sheets made by the Interpret option are "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 ever ever change their contents.

Freq ahead, the coast is clear.

3.4 The Freqs worksheet

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



(c3) Q1

Option	n	/60
A	26	43.3%
B	25	41.7%
C	9	15.0%

(c4) Q2

Option	n	/60
A	4	6.7%
B	12	20.0%
C	7	11.7%
D	8	13.3%
E	29	48.3%

(c5) Q3

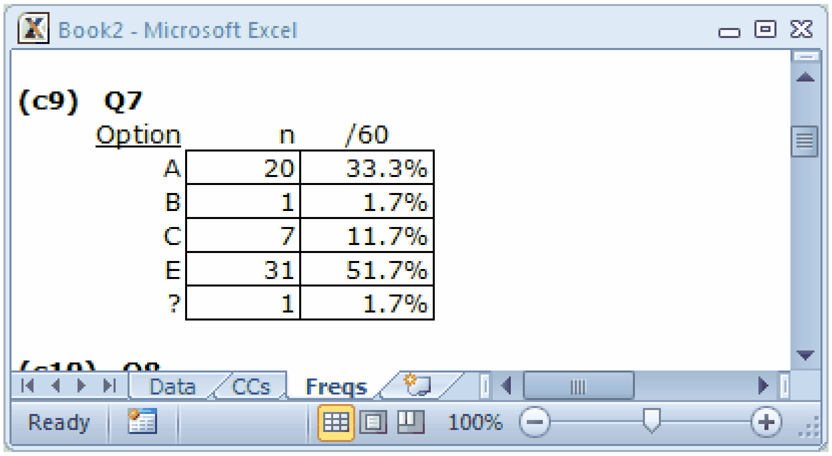
Option	n	/60
A	2	3.3%
B	1	1.7%
C	32	53.3%
D	25	41.7%

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 **(c3)** next to **Q1** indicates that these responses were found in column 3 of the Data worksheet.

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:



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.

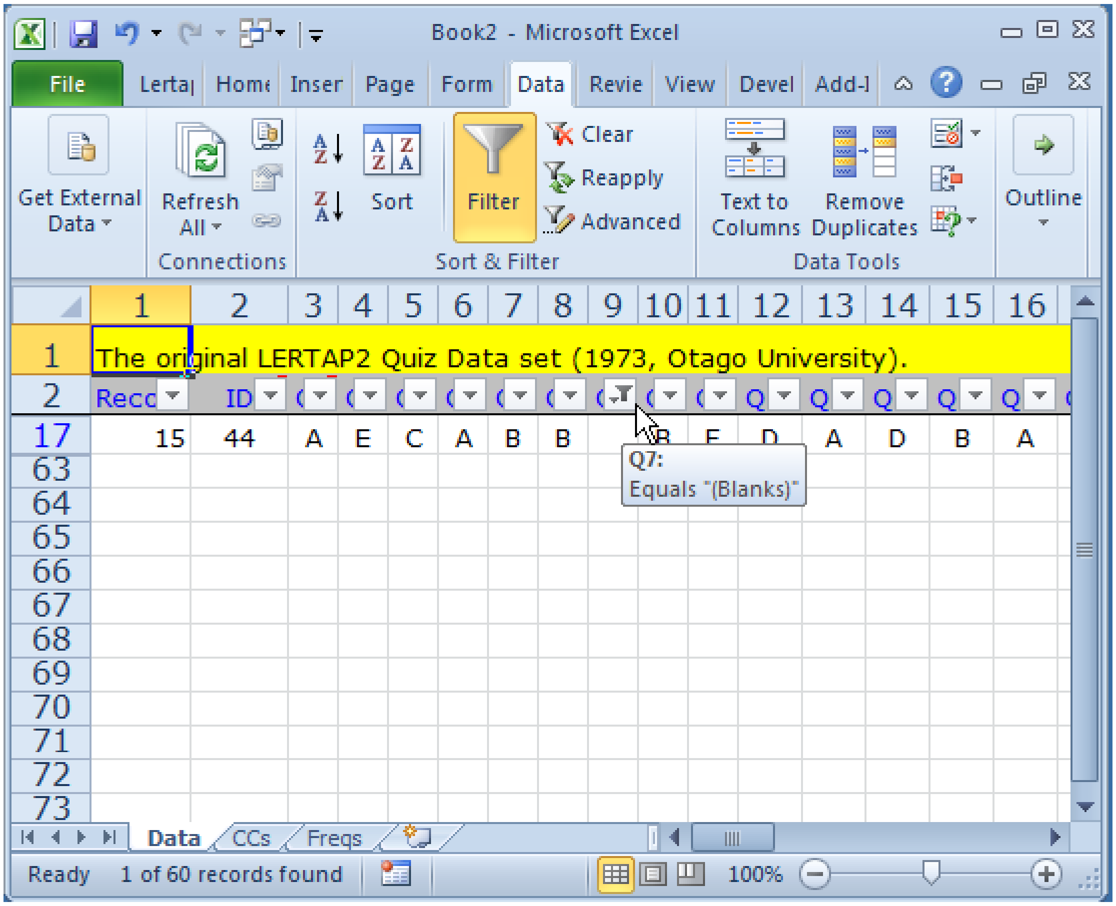
Why doesn't Freqs show a "D" for Q7? Because nobody chose that option.

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? The Data tab on the Excel ribbon has a "**Filter**" option, a powerful resource which will let you rapidly find records which have "weird" results.

Here's a snap of Filter in action. We asked it to find all records with a blank in column 9. The screen snapshot shown here looks messy, but try Filter for yourself. It's easy to use.



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

Enter Elmillon!

3.5 Elmillon item analysis

Elmillon? What is it? Ah-ha, caught you sleeping! Were you the diligent, conscientious reader we assumed you were, there would be no need to ask this question. The answer is given in an earlier topic.

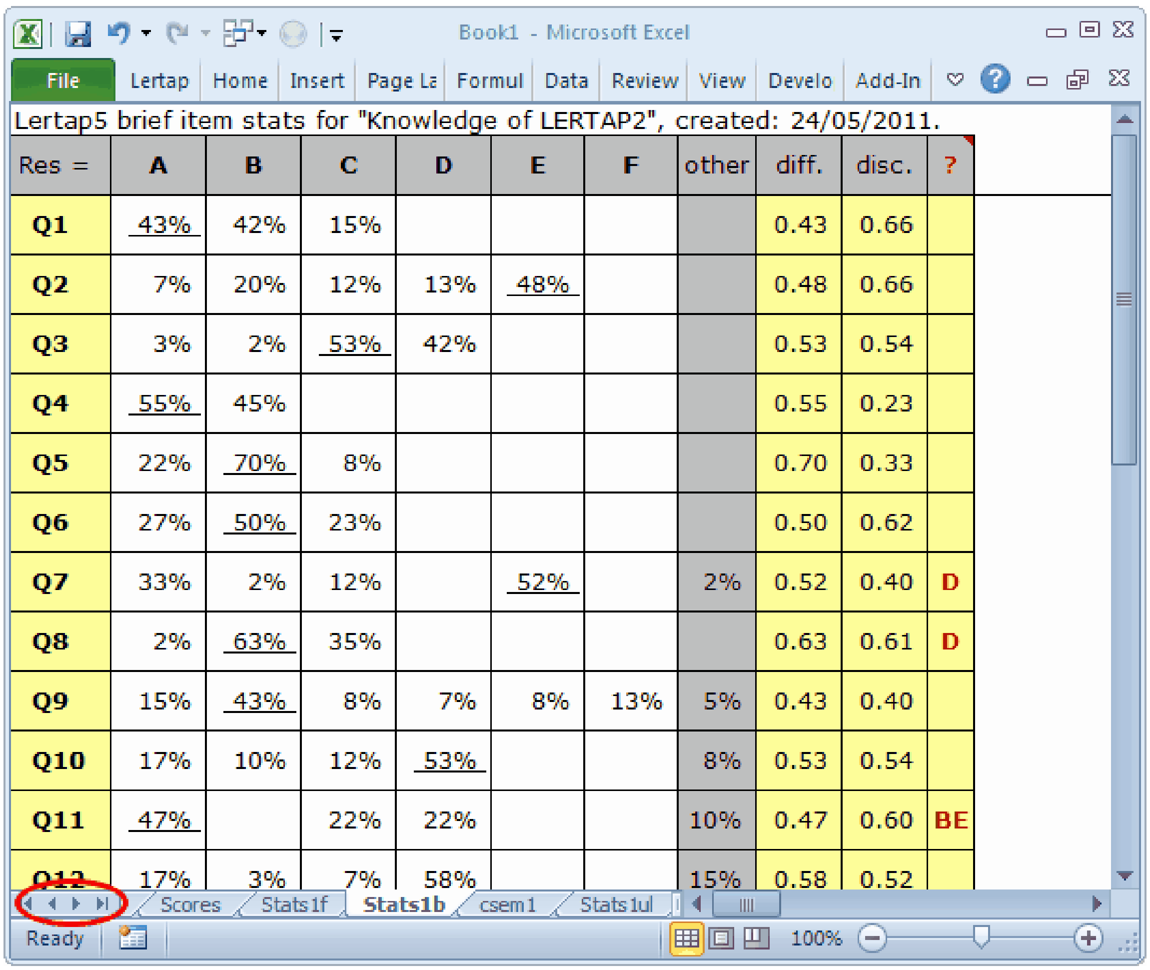
Click on **Elmillon**

After you clicked on Elmillon, 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.

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

Lertap went off to have a read of all the data it stored in those secret (hidden) Sub worksheets which were by-products of the "Interpret" stage. Then it read through the Data worksheet two or three times, writing a scores worksheet, four summary results worksheets for the cognitive test, and two results sheets for each of the affective tests. *Sheet it in, Lertap.*

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



Lertap5 brief item stats for "Knowledge of LERTAP2", created: 24/05/2011.

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	
Q11	47%		22%	22%			10%	0.47	0.60	BE
Q12	17%	3%	7%	58%			15%	0.58	0.52	

See the tabs named Scores, Stats1f, Stats1b, csem1, and Stats1ul? At this point the workbook has twelve visible (unhidden) worksheets, but only five tabs fit into the little viewing area at the bottom of the screen. You can see the other tabs by using the little scroll arrows circled in red, lower left of of the picture, above where the Status Bar says "Ready".

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 best 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 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. When refreshed, proceed.

3.6 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.

3	The first subtest has 25 cognitive items; responses start in column 3.
4	*col (c3-c27)
5	*sub Res=(A,B,C,D,E,F), Name=(Knowledge of LERTAP2), Title=(Knwldge), Wt=0
6	*key AECAB BEBBD ADBAB BCCCB BABDC
7	*alt CEDBC CEDFD EDCBD DCCDD DEEDF
8	The second subtest has 10 affective items; responses start in column 28.
9	*col (c28-c37)
10	*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort), Wt=0
11	*pol +---- ++--+
12	The third subtest uses the MDO option to process the same 10 affective items.
13	*col (c28-c37)
14	*sub Aff, MDO, Name=(Comfort items with MDO), Title=(MDOCmfrt), Wt=0
15	*pol +---- ++--+

There are three *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. The third points to the same items as the second; in this case, the third subtest uses the same items as the second subtest, but the items are scored differently.

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 *wts and *mws lines in the CCs worksheet.

(Note: the discussion of the MathsQuiz sample has an example which uses *mws lines. *wts lines, also known as *wgs lines, are mentioned here.)

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.

Lertap's default assumption is 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 26. 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" scoring.

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.

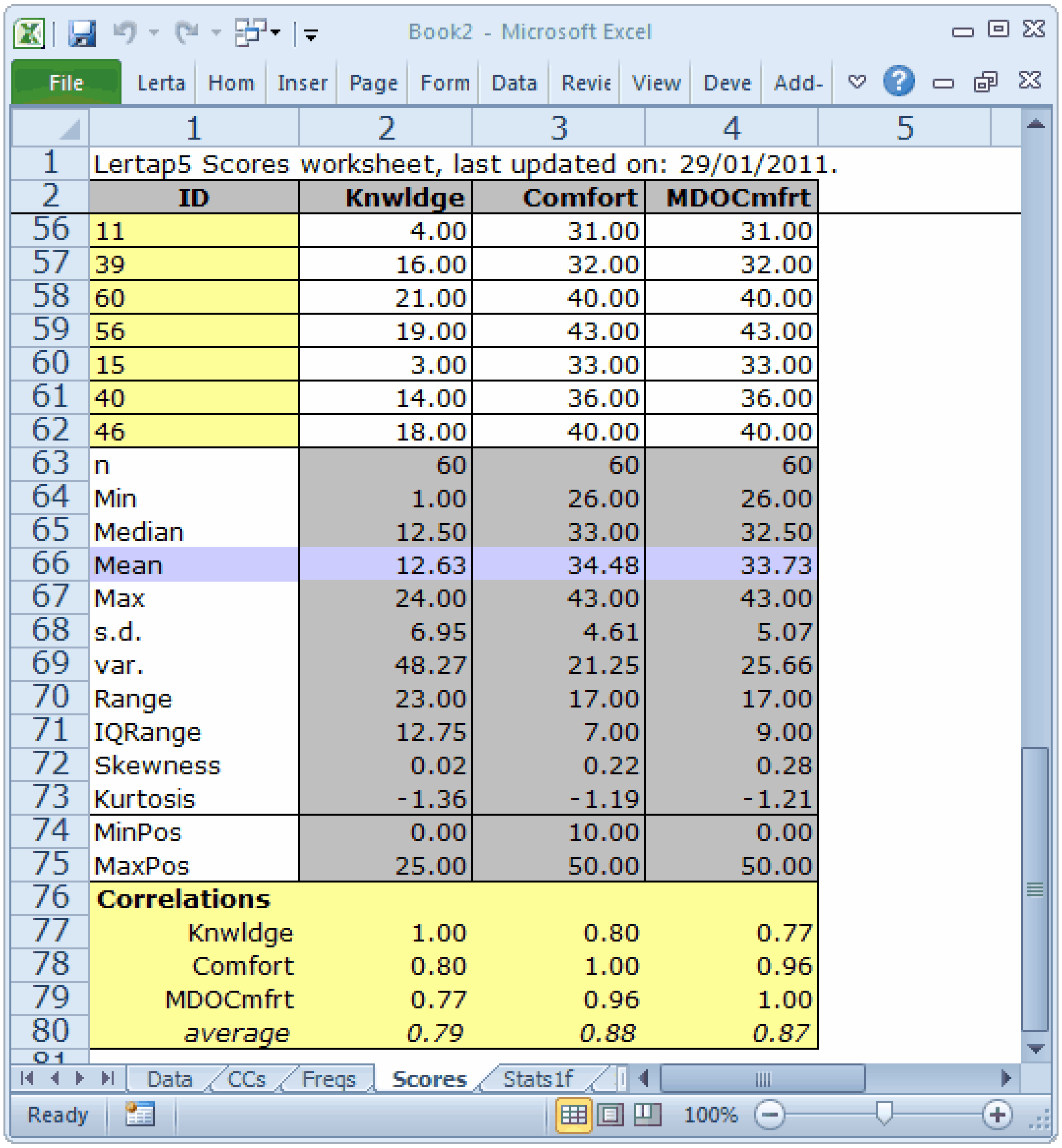
(To get even more *pol-arised, you could take in another affective example.)

What about missing data, such as the situation which arises when a student does not answer an item? Read all about Lertap's answer to this question by transplanting yourself to here; to quickly whisk away to MDO scoring, as used here for the third subtest, get whisk it.

Okay. Let’s have a squiz at these scores, Knwldge, Comfort, and MDOCmfrt.

Where are they?

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



Lertap5 Scores worksheet, last updated on: 29/01/2011.

Row	ID	Knwldge	Comfort	MDOCmfrt
56	11	4.00	31.00	31.00
57	39	16.00	32.00	32.00
58	60	21.00	40.00	40.00
59	56	19.00	43.00	43.00
60	15	3.00	33.00	33.00
61	40	14.00	36.00	36.00
62	46	18.00	40.00	40.00
63	n	60	60	60
64	Min	1.00	26.00	26.00
65	Median	12.50	33.00	32.50
66	Mean	12.63	34.48	33.73
67	Max	24.00	43.00	43.00
68	s.d.	6.95	4.61	5.07
69	var.	48.27	21.25	25.66
70	Range	23.00	17.00	17.00
71	IQRange	12.75	7.00	9.00
72	Skewness	0.02	0.22	0.28
73	Kurtosis	-1.36	-1.19	-1.21
74	MinPos	0.00	10.00	0.00
75	MaxPos	25.00	50.00	50.00
76	**Correlations**			
77	Knwldge	1.00	0.80	0.77
78	Comfort	0.80	1.00	0.96
79	MDOCmfrt	0.77	0.96	1.00
80	*average*	*0.79*	*0.88*	*0.87*

What do you make of it, this Scores sheet? It uses 80 rows, and 4 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 55 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 for each respondent.

Are the summary statistics self-explanatory? Good. You can find out how they're calculated in Chapter 10 of the manual, "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. Scale scores are requested by using the "Scale" control word on a *sub line.

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 very different subtests, with cognitive items in one, and affective items in the others.

Keep scoring?

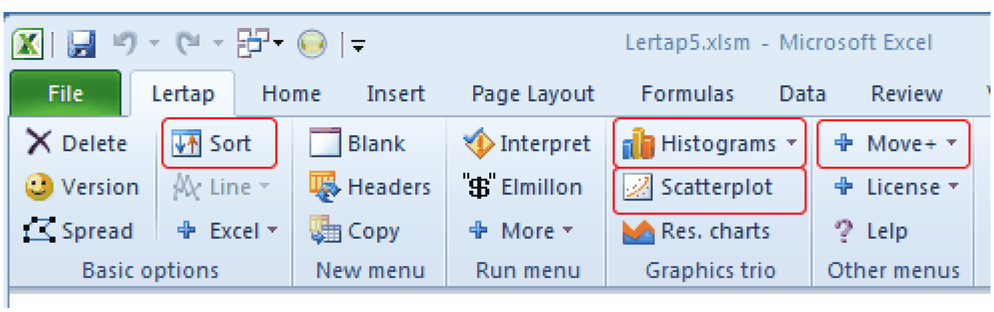
Tidbits:

Don't forget this all-important reference on CCs worksheets and their control lines: the on-line help system.

It is possible to score items which are **open-ended**, essay, "short answer", "free response", or "constructed response". See this paper, please, but note: the *alt line used towards the end of the paper is the old style; the *alt line's format was changed after this paper was written (a factor which should not limit your enjoyment of the paper).

3.6.1 Doing more with Scores

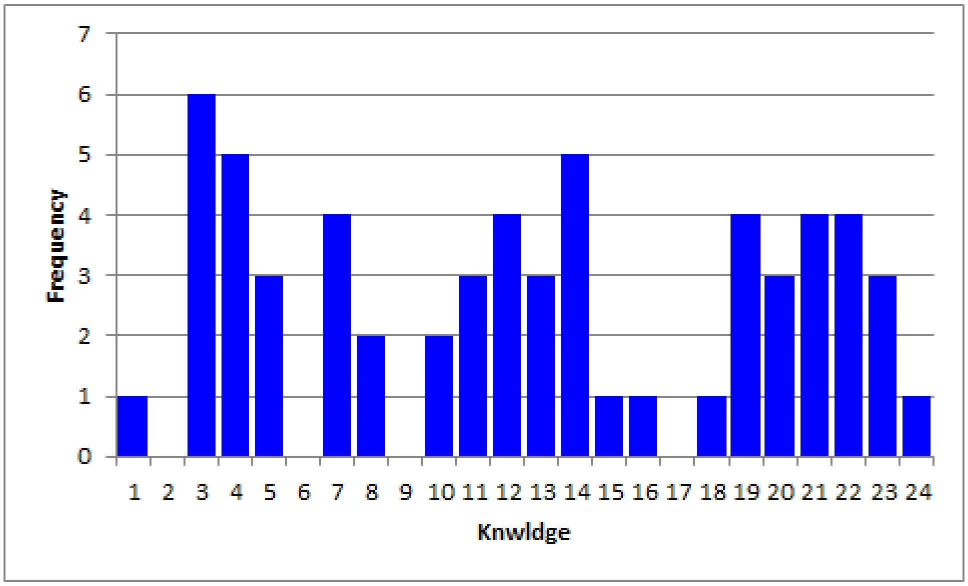
Once a Scores worksheet has been created, there are a few icons on the Lertap tab which let you do more with them:

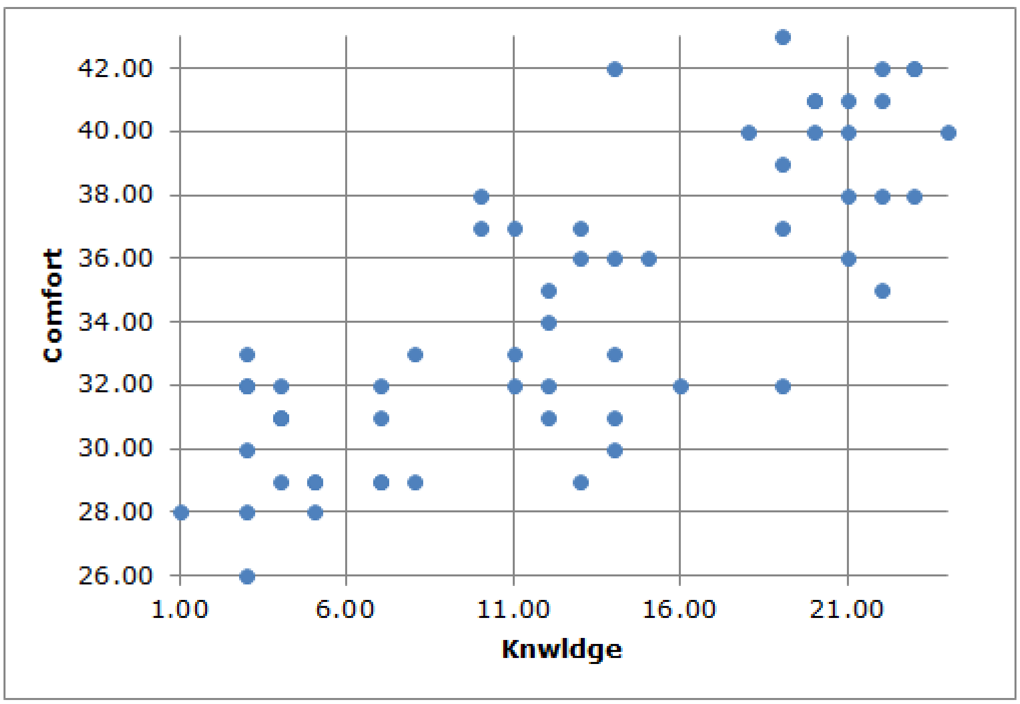


Click on the red boxes to branch out to corresponding topics.

(If the red boxes are not "hot" with your browser, try 'em here: Sort Histograms Scatterplot Move+)

Samples of the graphs produced by the histogram and scatterplot options are shown below:



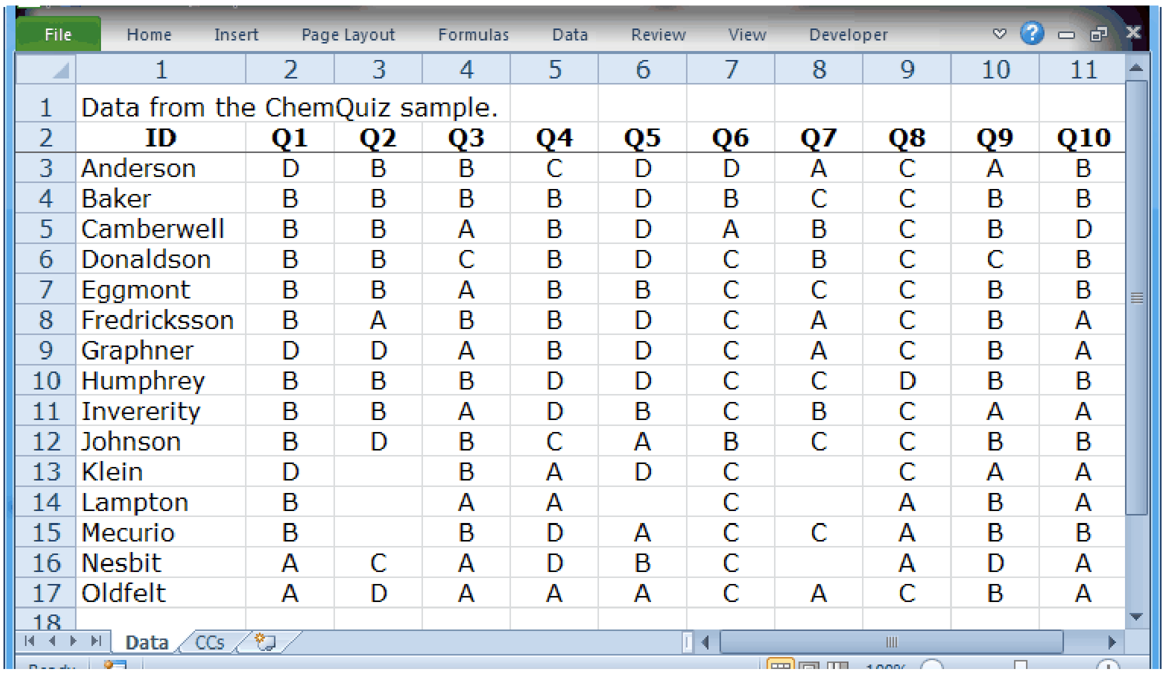


The next topic gets into the matter of missing data, and might be too technical for those on the Cook's Tour. We suggest you give missing data a miss, and jump to the first of Lertap's reports.

3.6.2 Missing data

In Lertap, "missing data" is a term which usually means that a person has not answered, or has omitted, an item.

If you look at a Lertap Data worksheet and see empty cells, such cells usually indicate that there was no answer to the item corresponding to the cell. In the screen snapshot below, Klien has not answered Q2, nor Q7. Lampton has left three items unanswered (2, 5, and 7); Mecurio did not answer Q2; Nesbit has not answered Q7.



	1	2	3	4	5	6	7	8	9	10	11
1	Data from the ChemQuiz sample.										
2	**ID**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
3	Anderson	D	B	B	C	D	D	A	C	A	B
4	Baker	B	B	B	B	D	B	C	C	B	B
5	Camberwell	B	B	A	B	D	A	B	C	B	D
6	Donaldson	B	B	C	B	D	C	B	C	C	B
7	Eggmont	B	B	A	B	B	C	C	C	B	B
8	Fredricksson	B	A	B	B	D	C	A	C	B	A
9	Graphner	D	D	A	B	D	C	A	C	B	A
10	Humphrey	B	B	B	D	D	C	C	D	B	B
11	Invererity	B	B	A	D	B	C	B	C	A	A
12	Johnson	B	D	B	C	A	B	C	C	B	B
13	Klein	D		B	A	D	C		C	A	A
14	Lampton	B		A	A		C		A	B	A
15	Mecurio	B		B	D	A	C	C	A	B	B
16	Nesbit	A	C	A	D	B	C		A	D	A
17	Oldfelt	A	D	A	A	A	C	A	C	B	A
18											

There are quite a number of users who prefer to use a special code for the case of unanswered items. Instead of leaving a cell empty, they might use a 9 to indicate missing data (there are historical reasons for this; some of the most popular data analysis programs have traditionally used 9s to represent missing data). When using scanners to process mark-sense answer sheets, the software driving the scanner may have its own missing data code, such as an asterisk (*); some scanners will code missing data and invalid responses with a Z. (A common invalid response for a scanner arises when a student has shaded in more than one response on the answer sheet.)

To fully understand how Lertap processes missing data, it helps to have a good understanding of what Lertap calls "response codes".

The letters (or digits) seen under a Lertap Data worksheet's item columns represent response codes. In the example above, it seems that items have used response codes of A, B, C, and D; this set, {A,B,C,D}, is, in fact, Lertap's default response code set for cognitive items -- for affective items, the default response code set is {1,2,3,4,5}.

It is common for users to have items which use other response codes. Whenever a test does not use the default response codes, the response codes used by that test's (or survey's) items are specified by using an Res= declaration on a *sub card.

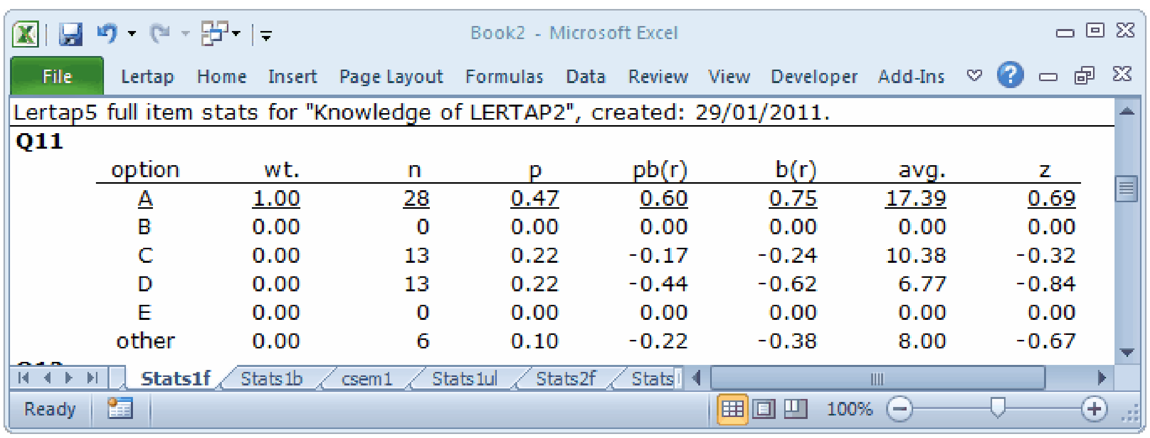
For example, Res=(A,B,C,D,E,F), Res=(1,2,3,4,5,6,7), and Res=(t,f) are all valid Res= declarations. In Lertap, a test may use up to 26 response codes.

Now, with this understanding of response codes in hand, the definition of missing data can be made a bit more precise: in Lertap, an item response is said to be missing whenever a cell in the Data worksheet has an entry which does not match the response codes used by the item. When this happens, Lertap says it has encountered an "other" response.

As Lertap goes about tallying item responses, it keeps track of the number of "other" responses in a special bin.

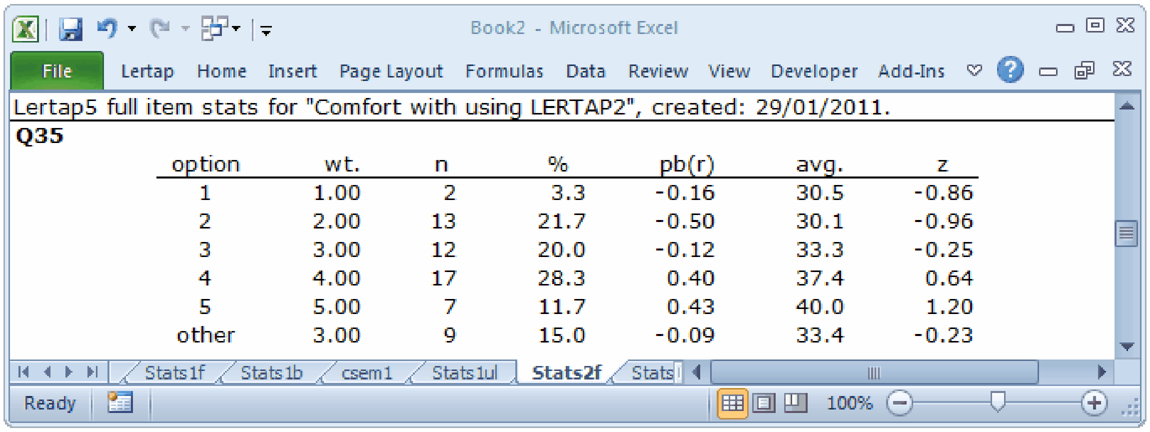
What's this bin called? The "other bin", naturally. Deep inside Lertap, each and every item is assigned a storage bin for "other" responses. The contents of an item's "other" bin are displayed in many of Lertap's reports.

As an example, look at these snapshots:


Lertap5 full item stats for "Knowledge of LERTAP2", created: 29/01/2011.

Q11

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	1.00	28	0.47	0.60	0.75	17.39	0.69
B	0.00	0	0.00	0.00	0.00	0.00	0.00
C	0.00	13	0.22	-0.17	-0.24	10.38	-0.32
D	0.00	13	0.22	-0.44	-0.62	6.77	-0.84
E	0.00	0	0.00	0.00	0.00	0.00	0.00
other	0.00	6	0.10	-0.22	-0.38	8.00	-0.67



Lertap5 full item stats for "Comfort with using LERTAP2", created: 29/01/2011.

Q35

option	wt.	n	%	pb(r)	avg.	z
1	1.00	2	3.3	-0.16	30.5	-0.86
2	2.00	13	21.7	-0.50	30.1	-0.96
3	3.00	12	20.0	-0.12	33.3	-0.25
4	4.00	17	28.3	0.40	37.4	0.64
5	5.00	7	11.7	0.43	40.0	1.20
other	3.00	9	15.0	-0.09	33.4	-0.23

In these two examples, the "other" bin for Q11 has n=6, meaning that there were six people missing data on the item. For Q35, there were n=9 people missing data.

So. Are you now full bottle on Lertap's definition of missing data? And an expert on Lertap's "other" bin?

Goodonyamate. But wait, there's more ....

What this stuff under the "wt." column? The answer to this question is also very relevant to understanding how Lertap processes missing data, especially for affective tests.

The "wt.", for "weight", indicates how many points are associated with each item option.

On Q11, the 28 people who chose response code A (usually referred to as option A, or alternative A) will get 1.00 points for their answer. Choosing any of Q11's other options gets no points. You guessed it: Q11 is a cognitive item whose right answer is A.

Not answering Q10 gets no points. Not answering Q10 gets no points. We have repeated this as it's important: an item's other bin can have scoring points attached to it.

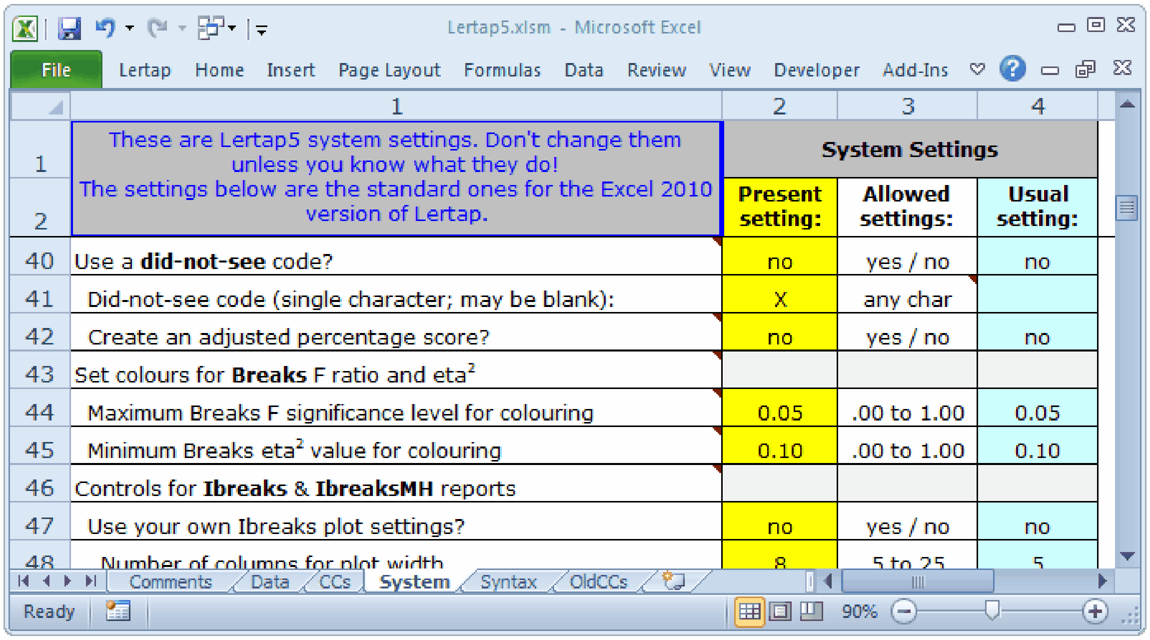
Look at Q35, an affective item. Each and every response code has associated scoring points, and, should someone not answer Q35, they will still get 3.00 points (!).

What we've displayed in these two examples represents Lertap's default handling for cognitive and affective missing data. A cognitive item's default missing data scoring action is 0.00 points; an affective item's default missing data score is equal to the mean of the other scoring wts.

With this in mind, take a break. When you return, have a look at the following topics to read more about Lertap and missing data.

3.6.2.1 Did-not-see option

Options in the System worksheet allow users to adjust cognitive item statistics so that they are based only on those cases (students) that had a chance to answer the item.



The did-not-see option is activated by placing "yes" in column 2 of the appropriate row in the System worksheet; in the example above, this is row 40.

Use of this option requires a corresponding did-not-see code, a single character recorded in the Data worksheet whenever a student did not have the chance to answer an item (for whatever reason). In the example above, the code is a blank. The code may be any single character, including a blank, an upper-case letter from the Latin alphabet (ABC...XYZ), a lower-case letter (abc...xyz), or a single Arabic numeral (0123456789).

How does Lertap process item responses with this new option? Well, for any item, Lertap first looks to see if a student's response corresponds to one of the response codes used by the item's options. If it does not, then Lertap looks to see if the "response" matches the did-not-see code, assuming the did-not-see option has been activated. If it does not, then the response is classed as "other", a response category often referred to in Lertap documentation as "missing data".

Lertap will automatically adjust all of its item statistics so that they exclude did-not-see cases (that is, of course, assuming the did-not-see option has been activated—if the option has not been activated, then what would have been a did-not-see response will be included in the "other" response category, and processed as missing data).

Use the Freqs worksheet to check on missing data and did-not-see cases, and note: if the did-not-see code is a blank, Freqs will include did-not-see cases in its "other" line.

To see what the various "Stats" reports look like when the did-not-see option is on, just page forward to the following topics.

Scoring

Whenever the did-not-see option is in effect, users have the option to have Lertap calculate two scores for every person on each subtest defined in the CCs worksheet: the usual subtest score, and a percentage score which is based on the maximum possible score which could have been obtained on the items actually presented.

The first score, the "usual" subtest score, will be the number of points earned on the number of items the person had the opportunity to answer. In the cognitive case, the most common scoring method is to give one point for each correct answer—in this case, the "first score" will simply be the number of items right.

The second score is only produced when "yes" is found in the System worksheet in the second column of the line which says "Create an adjusted percentage score?". This score is formed by dividing the first score, the "usual" subtest score, by the maximum possible score which the person could have had if s/he had correctly answered every presented item; this figure is then multiplied by 100 to get a percentage index.

Coming back to the common cognitive case, with one point for each correct answer, the second score is the number of answers right divided by the number of items presented, multiplied by 100.

Note that both of these scores exclude items which the person did not see, that is, did not have the opportunity to answer.

It is not necessary to use the PER control word on the *sub card in order to get the second score: it's computed automatically, providing "yes" is found in the second column of the line which says "Create an adjusted percentage score?".

Adjusted?

When would you want to have the percentage scores "adjusted"? What does "adjusted" mean, anyway?

Well, it turns out that certain online test generators can create tests of variable length; some students might see a 48-item test, others a 45-item test, and still others a 50-item test. A student's test score will depend on the number of items s/he had the chance to answer. Two students with the same "usual" subtest score, say 35, will not have the same percentage score if they have been presented with a different number of items.

For example, if Jorge got a score of 35 on a test with 50 items, his percentage score would be 70%, assuming one point for each correct item. Suppose it turned out that

Marisol also got a score of 35, but she was presented with only 45 items; her rounded percentage score will be 78%.

When you've asked Lertap to "Create an adjusted percentage score", the percentage test score will be adjusted according to the items a student actually saw.

If all students get the same number of items, there is no need to create an adjusted percentage score. Ah, but wait a minute! We should adjust this comment: if all students are presented with the same number of items, an adjustment is not required *if all items are scored the same way*. When items are scored in different ways, then the adjustment might again be useful. For example, if Item 26 is worth one point, but Item 27 is worth two points, then we might well want to make an appropriate adjustment to scores, depending not only on how many items a student saw, but also on the number of points the items are worth.

Lertap creates the adjusted score by dividing the "usual" subtest score by the actual maximum possible score a student could have achieved, based on the items presented to the student.

Sample Scores output

As you page forward to following topics, you'll come across an example from an actual online testing situation, one which used the Test Pilot system from McGraw-Hill.

Over 400 students at a large North American university took an online version of a test delivered by Test Pilot. Sampling from a pool of 80 cognitive multiple-choice items, Test Pilot served up tests whose lengths varied: some students were presented with 40 items, some with 43, some 45, and others 48. (Please read the second "tidbit" below -- the situation was actually a bit more complicated.)

The Lertap Data worksheet turned out to have 93 columns for item responses. Why 93 when there were only 80 items in the pool? Because a few items allowed for multiple responses (see second tidbit below).

In this example, blanks were used as the did-not-see code. Were you to look at the item responses for any student, scanning from left to right over the 93 columns allocated, you'd see actual responses, many blanks, and, for those items not seen by a student but not answered, a 9.

We could, and will, entertain a variety of Lertap scoring "methods" which will demonstrate how the did-not-see options affect test scores.

To begin, say we have the did-not-see option off, that is, we have "no" in the System worksheet row which says "Use a did-not-see code?". Also, assume we have the PER control word on the *sub card.

The Scores output will look as follows:

Record No.	EE101	EE101%
1	38.00	40.9
2	30.00	32.3
3	27.00	29.0
4	29.00	31.2
5	21.00	22.6

The first student had a test score of 38. Lertap says that the corresponding percentage score is 40.9, a value found by dividing the score, 38, by 93, the maximum possible test score if each item is scored giving one point for a correct answer.

But this is not correct in this case; no student had the opportunity to answer 93 items -- Test Pilot gave each student a random sample of approximately 40 items.

So, we activate the did-not-see option by putting "yes" in the second column of the System worksheet row which says "Use a did-not-see code?". We have "no" in the "Create an adjusted percentage score?" row. And, we still have PER on the *sub card. Our Scores now look as follows:

Record No.	EE101	EE101%
1	38.00	79.2
2	30.00	62.5
3	27.00	56.3
4	29.00	60.4
5	21.00	43.8

See how the percentage scores have changed? Lertap is now basing its percentage score on the maximum possible score which could have been earned on the actual items presented to the student. For the first student, the maximum was 48.

Dividing 38 by 48 and multiplying by 100 gives the 79.2 seen as the EE101% score for the first student.

Now, for the second student, does 30 divided by 48 equal 62.5? Yes. The problem is that the second student was presented with 40 items, not 48. The percentage score is wrong.

Whenever students are presented with a different number of items, or whenever items in the pool have different scoring patterns, "yes" should be used in the "Create an adjusted percentage score?" row.

So, let's see what happens. We put in that "yes". Do we still have PER on the *sub card? No, now it's not needed -- there would be no harm in having it there, but it is not required. Here are the scores:

Record No.	EE101	EE101%
1	38.00	79.2
2	30.00	75.0
3	27.00	67.5
4	29.00	72.5
5	21.00	52.5

The percentage scores above have each been adjusted, according to the maximum possible score a student could have obtained on the set of items s/he was presented with.

Yes, Miss? (A young woman in the fortieth row of the third balcony has her hand up.)

Why wouldn't I just use the adjusted percentage score option all of the time? Why not put that yes where it's supposed to be, and just leave it there for always?

A good question, thank you. The answer: Lertap pinches extra memory from the computer whenever it has to keep track of the maximum possible score each student could have achieved given the items presented. You save memory space, and also a tiny bit of processing time, by not adjusting the percentages. If you know that each student was presented with the same number of items, and all items were scored the same way, then say "no" to the "Create an adjusted percentage score?" option. Now you know, no?

Related tidbits:

Several learning management systems support the development and delivery of online cognitive and affective tests. Some, such as Angel Learning, can be made to randomly sample items from a database, presenting different students with different versions of a test, each with the same number of items. In classical test theory, when certain conditions have been met, these versions might be termed parallel forms, or equivalent forms. However, as of March 2006, the data file of item responses created by Angel does not indicate which test "form" a student took. Fortunately, the Angel output is padded with did-not-see codes so that each data record has the same length, a length equal to the total number of items in the database.

In the Test Pilot example discussed above, the test included a few items which used a "Check all of the following options which would be correct" format, effectively turning a single item into multiple true-false items. Test Pilot actually served up the same number of items to each student (40), but those items which used the "check all of the following" format did not consistently offer the same number of options. The practical end result was as described above: students received a variable number of test items.

Did-not-see data will affect the calculation of alpha, Lertap's reliability coefficient -- alpha is NOT corrected for did-not-see cases.

3.6.2.2 MDO cognitive, Statsf

There's a control word, "MDO", which may be used on the *sub card to control how Lertap processes missing data. MDO may be used with both types of test, cognitive and affective. The letters stand for "missing data out", meaning that records with missing data are to be excluded from Lertap's various calculations.

This topic discusses the effect of the MDO and did-not-see options on the cognitive test reports produced by Lertap; a following topic does likewise for affective tests.

Look at these CCs lines for a cognitive test:

1	*col (c3-c27)
2	*sub Res=(A,B,C,D,E,F), Name=(Knowledge of LERTAP2), Title=(Knwldge), Wt=0
3	*key AECAB BEBBD ADBAB BCCCB BABDC
4	*alt CEDBC CEDFD EDCBD DCCDD DEEDF
5	*col (c3-c27)
6	*sub MDO, Res=(A,B,C,D,E,F), Name=(Knowledge with MDO), Title=(MDOKnwl), Wt=0
7	*key AECAB BEBBD ADBAB BCCCB BABDC
8	*alt CEDBC CEDFD EDCBD DCCDD DEEDF

These eight "cards" involve the same 25 items. Being the avid reader you are, you know you've seen these items before -- they're from the Lertap quiz.

The cards define two subtests. The only difference between the two is that the second one has the MDO option on; you can see it on the 6th line.

Statsf reports

A squiz of the Lertap's Stats1f and Stats2f reports will serve to highlight the effect of using MDO:

Lertap5 full item stats for "Knowledge of LERTAP2", created: 20/03/2006.

Q15

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	1	0.02	0.03	0.08	14.00	0.20
B	1.00	30	0.50	0.53	0.66	16.67	0.58
C	0.00	13	0.22	-0.36	-0.50	7.92	-0.68
D	0.00	16	0.27	-0.33	-0.45	8.81	-0.55

Q16

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	12	0.20	-0.14	-0.19	10.75	-0.27
B	1.00	36	0.60	0.46	0.58	15.53	0.42
C	0.00	3	0.05	-0.26	-0.56	4.67	-1.15
D	0.00	6	0.10	-0.29	-0.50	6.50	-0.88
other	0.00	3	0.05	-0.23	-0.49	5.67	-1.00

The report above comes from Stats1f. It summarizes the performance of two items, Q15 and Q16, using a variety of statistics.

Both items use four options, employing response codes {A,B,C,D}.

A student selecting option B will get 1.00 points towards her/his test score, as seen under the wt. column -- we'd conclude that both items have just one correct answer as all the other wt. values are zilch (zero).

Thirty (30) students got Q15 right. Sixty (60) students were involved, so p for Q15's option B is 0.50, which is, of course, 30 divided by 60.

The pb(r) and b(r) columns are, respectively, point-biserial and biserial correlation coefficients, indexing the relationship between option selection and the criterion score. If the item forms part of the criterion, as it does in this case, Lertap applies a part-whole correction, taking out the inflation the correlation coefficients would otherwise have (see the manual for a more extensive discussion).

The avg. column shows the average criterion score for those students selecting each item option. On Q16, the 36 students who selected option B had average criterion scores of 15.53. As a z-score, 15.53 is 0.42 (the manual has more to say; you ought to read it some day, perhaps when next at the beach).

Note that Q16 has an "other" line. Three students had missing data for Q16. They were weak students; their avg. was low, as it was for those who chose distractors C and D.

Okay? Got it? Good; now have a peep at the stats for the same two items after the MDO option has been used:

Lertap5 full item stats for "Knowledge with MDO", created: 20/03/2006.

Q15

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	1	0.02	0.03	0.08	14.00	0.20
B	1.00	30	0.50	0.53	0.66	16.67	0.58
C	0.00	13	0.22	-0.36	-0.50	7.92	-0.68
D	0.00	16	0.27	-0.33	-0.45	8.81	-0.55

Q16

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	12	0.21	-0.17	-0.24	10.75	-0.32
B	1.00	36	0.63	0.42	0.54	15.53	0.36
C	0.00	3	0.05	-0.28	-0.59	4.67	-1.20
D	0.00	6	0.11	-0.32	-0.54	6.50	-0.94
other	0.00	3	0.05	-0.25	-0.52	5.67	-1.06

There's no change in the stats for Q15; all 60 students answered that item. But Q16 has changed; many of Q16's stats above, from p through z, differ from the Q16 stats seen earlier. Why? Because the three students missing an answer to Q16 have been excluded from the calculation of the stats.

Look at the p column, for example. For Q16's option B, p is now 0.63, corresponding to 36 divided by 57, not 36 divided by 60.

The avg. values are now computed using just the criterion scores for the 57 students who answered Q16. This applies to the other line too: 5.67, as a z-score in the distribution of 57 scores, would be -1.06.

Now, take a few seconds and sum down the p column for Q16. In the first report above the sum is 1.00 (100%). But in the second report the sum comes to 1.05 (105%). What's up?

The 0.05 p value for Q16's "other" row is just indicating the proportion of respondents who did not answer the question. The other four p values for Q16 sum to 1.00, which is what is expected when the MDO option is in use.

Statsf reports and the did-not-see option

When you've got the did-not-see option going, the Statsf report will be quite similar to what you've just seen immediately above. Look:

Lertap5 full item stats for "Knowledge with DNSI", created: 24/03/2006.

Q15

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	1	0.02	0.03	0.08	14.00	0.20
B	1.00	30	0.50	0.53	0.66	16.67	0.58
C	0.00	13	0.22	-0.36	-0.50	7.92	-0.68
D	0.00	16	0.27	-0.33	-0.45	8.81	-0.55

Q16

option	wt.	n	p	pb(r)	b(r)	avg.	z
A	0.00	12	0.21	-0.17	-0.24	10.75	-0.32
B	1.00	36	0.63	0.42	0.54	15.53	0.36
C	0.00	3	0.05	-0.28	-0.59	4.67	-1.20
D	0.00	6	0.11	-0.32	-0.54	6.50	-0.94
other			0.05				

There's just a wee difference in these results. Can you spot it? The other line is now missing most of its stats.

What about the joint operation of MDO and did-not-see. What happens then? Got time to look at results for two items delivered to 421 students over the internet, using the Test Pilot system from McGraw-Hill?

(c10) Q50

Option	n	/421
1	114	27.1%
2	94	22.3%
3	62	14.7%
4	42	10.0%
9	3	0.7%
?	106	25.2%

(c11) Q80

Option	n	/421
1	124	29.5%
2	196	46.6%
?	101	24.0%

In this example, a blank was used as the did-not-see code, and 9 as the code for missing data.

On Q50, three students failed to provide an answer, while 106 students did not see it (Test Pilot presented Q50 to 315 of the 421 students). Q80 was presented to 320 students, and they all answered it. Look at the corresponding Statsf summaries:

Lertap5 full item stats for "EE 101 T3 exam", created: 20/03/2006.

Q50

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	114	0.37	-0.15	-0.20	27.37	-0.20
2	1.00	94	0.30	0.25	0.33	30.74	0.52
3	0.00	62	0.20	-0.15	-0.22	26.87	-0.31
4	0.00	42	0.13	-0.06	-0.10	27.57	-0.16
other	0.00	3	0.26	-0.10	-0.37	23.67	-0.99

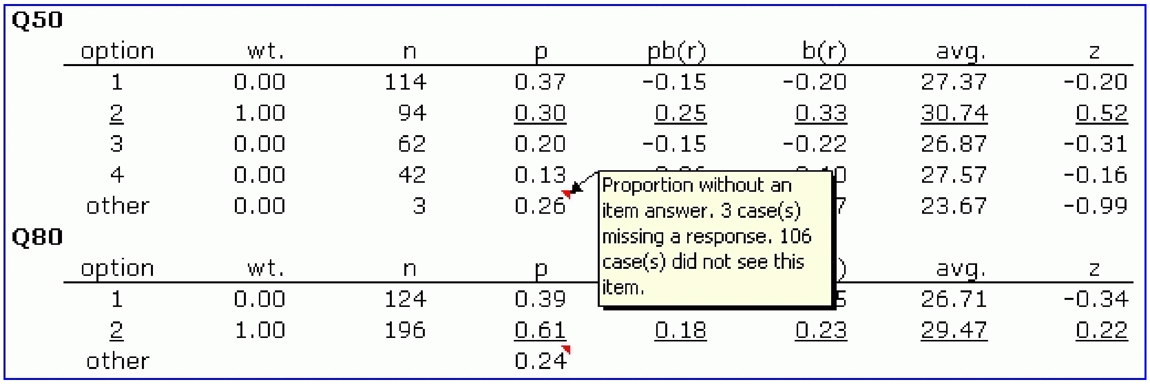
Q80

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	124	0.39	-0.27	-0.35	26.71	-0.34
2	1.00	196	0.61	0.18	0.23	29.47	0.22
other			0.24				

The statistics for both Q50 and Q80 have been computed by excluding the students who did not see the items, and by also excluding those with missing data.

Were you to sum the p values for Q50's four options, you'd have 0.37 + 0.30 + 0.20 + 0.13, or 1.00 (100%). The 0.26 on Q50's other line is the proportion of students who did not have an answer for the item, either because they did not see the item, or saw the item but did not answer it.

Let your mouse hover over the 0.26 value, and behold:

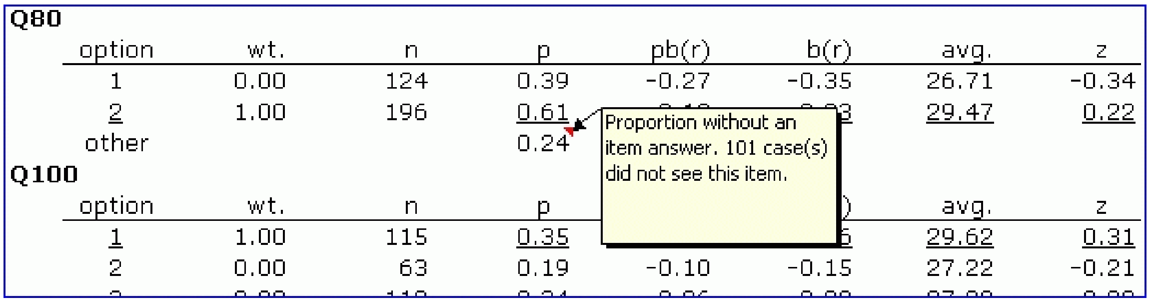


Q50

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	114	0.37	-0.15	-0.20	27.37	-0.20
2	1.00	94	0.30	0.25	0.33	30.74	0.52
3	0.00	62	0.20	-0.15	-0.22	26.87	-0.31
4	0.00	42	0.13	[illegible]	[illegible]	27.57	-0.16
other	0.00	3	0.26	[illegible]	[illegible]	23.67	-0.99

Q80

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	124	0.39	[illegible]	[illegible]	26.71	-0.34
2	1.00	196	0.61	0.18	0.23	29.47	0.22
other			0.24				



Q80

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	0.00	124	0.39	-0.27	-0.35	26.71	-0.34
2	1.00	196	0.61	[illegible]	[illegible]	29.47	0.22
other			0.24				

Q100

option	wt.	n	p	pb(r)	b(r)	avg.	z
1	1.00	115	0.35	[illegible]	[illegible]	29.62	0.31
2	0.00	63	0.19	-0.10	-0.15	27.22	-0.21

See how it works? When Lertap is running with both options, did-not-see and MDO, then the other line will have stats only when there were some students who did not answer the item. For Q50, the avg. criterion score for the three students who were missing data was 23.67; when this avg. score is inserted into the distribution of criterion scores for those 312 students who did answer the item, the corresponding z-score is -0.99.

Related tidbit:

As discussed in the manual, when more than one option to a cognitive item has a non-zero "wt." value, the pb(r) and b(r) statistics are corrected for part-whole inflation only for the option having the greatest wt.

3.6.2.3 MDO cognitive, Statsb

The previous topic discussed how the MDO and did-not-see options affect the information reported in Statsf reports.

Now you're set to see the corresponding Statsb reports:

Lertap5 brief item stats for "Knowledge of LERTAP2", created: 20/03/2006.										
Res =	A	B	C	D	E	F	other	diff.	disc.	?
Q15	2%	50%	22%	27%				0.50	0.53	A
Q16	20%	60%	5%	10%			5%	0.60	0.46	
Q17	12%	25%	57%				7%	0.57	0.56	
Q18	5%	43%	47%				5%	0.47	0.63	A
Q19	27%	10%	48%	8%			7%	0.48	0.76	
Q20	35%	40%	8%	7%			10%	0.40	0.70	D

Lertap5 brief item stats for "Knowledge with MDO", created: 20/03/2006.										
Res =	A	B	C	D	E	F	n	diff.	disc.	?
Q15	2%	50%	22%	27%			60	0.50	0.53	A
Q16	21%	63%	5%	11%			57	0.63	0.42	
Q17	13%	27%	61%				56	0.61	0.52	
Q18	5%	46%	49%				57	0.49	0.64	
Q19	29%	11%	52%	9%			56	0.52	0.74	
Q20	39%	44%	9%	7%			54	0.44	0.67	

The first report has an "other" column which indicates the number of people with missing data on each item, expressed as a percentage figure. The statistics in the diff. and disc. columns are based on calculations which include the people with missing data.

In the second report, the "other" column has been replaced by the "n" column. The entries in this column indicate how many people answered each item, and the various percentage figures, plus the diff. and disc. values, are based on n, that is, they exclude missing data. Lertap uses what is termed a "pairwise" exclusion rule to calculate the diff. and disc. values: if a student is missing data for the item, or did not see it, s/he is excluded from the calcs. for that item.

Lertap5 brief item stats for "EE 101 T3", created: 20/03/2006.								
Res =	1	2	3	4	n	diff.	disc.	?
Q50	37%	30%	20%	13%	312	0.30	0.25	
Q80	39%	61%			320	0.61	0.18	
Q100	35%	19%	34%	11%	326	0.35	0.13	
Q110	24%	39%	33%	5%	323	0.33	- 0.07	24
Q120	7%	43%	3%	47%	321	0.43	- 0.12	4
Q130	3%	6%	3%	88%	320	0.88	0.12	

The report above corresponds to the Test Pilot results mentioned in the previous topic. The numbers seen in the "n" column exclude the number of cases with missing data (if any), as well as the number of cases who did not see the item (if any). To see how many cases were missing data, or did not see the item, refer to the Statsf report which corresponds, or to the Freqs report.

Related tidbit:

A reminder from Chapter 10 of the manual: the Statsb reports are computed on an item level. The disc value they display is a conventional product-moment correlation between the item and the criterion, corrected for part-whole inflation.

3.6.2.4 MDO cognitive, Statsul

The two preceding topics have discussed how the MDO and did-not-see options affect Statsf and Statsb reports. Now: Lertap's third report for cognitive items, Statsul -- what happens when the MDO and did-not-see options are used?

Lertap5 U-L stats for "EE 101 T3 Fall 2005", created: 27/03/2006.							
Res =	1	2	3	4	other	U-L diff.	U-L disc.
Q50 upper	0.23	0.38	0.11	0.01	0.27	**0.23**	**0.30**
2nd	0.19	0.29	0.12	0.11	0.30		
3rd	0.29	0.31	0.12	0.13	0.15		
4th	0.33	0.06	0.17	0.18	0.26		
lower	0.31	0.08	0.23	0.07	0.31		

First up, above, the standard format for Statsul, what's seen before the MDO and did-not-see options are put to use.

Item Q50 is from the Test Pilot system mentioned in the previous topics. Over 400 students took the "EE 101 T3" exam over the internet; 25.2% of the students did not see Q50 as Test Pilot did not present it to them, while just under 1% of the students who saw Q50 didn't answer it.

Lertap5 U-L stats for "EE 101 T3 Fall 2005 with MDO", created: 27/03/2006.							
Res =	1	2	3	4	other	U-L diff.	U-L disc.
Q50 upper	0.31	0.52	0.15	0.02	0.00	**0.33**	**0.40**
2nd	0.27	0.41	0.17	0.15	0.00		
3rd	0.35	0.36	0.14	0.15	0.00		
4th	0.45	0.08	0.23	0.24	0.00		
lower	0.45	0.12	0.33	0.10	0.00		

The table above gives Q50 results after the MDO option has been turned on. Its statistics are based only on those students who answered the item. This means that all the did-not-see people, plus the we-saw-it-but-did-not-answer-it people, have been excluded from the proportions and from the calculations underlying U-L diff. and disc. The did-not-sees are excluded at this point as the did-not-see option has not yet been used; the did-not-see code has been processed as missing data.

Lertap5 U-L stats for "EE 101 T3 Fall 2005" (DNSI on), created: 27/03/2006.							
Res =	1	2	3	4	other	U-L diff.	U-L disc.
Q50 upper	0.31	0.52	0.15	0.02	0.01	**0.32**	**0.40**
2nd	0.27	0.41	0.17	0.15	0.00		
3rd	0.35	0.36	0.14	0.15	0.00		
4th	0.44	0.08	0.22	0.24	0.01		
lower	0.44	0.12	0.32	0.10	0.01		

In this table (above), the did-not-see option has been activated, but MDO has not. The "other" column is now indicating the proportion of students in each group, from upper down to lower, who were presented with Q50 by Test Pilot, but did not answer it. Those not presented with Q50, the did-not-sees, have been excluded.

For a refresher on how the U-L diff. and U-L disc. values are calculated, pay a visit to your local ice cream shoppe, then have a look at Chapter 10 of the manual.

3.6.2.5 MDO affective

The "MDO" control word is used on the *sub card to get Lertap to exclude cases with missing data from its calculations. MDO may be used with both types of test, cognitive and affective.

The discussion found in this topic assumes some familiarity with material found in the topics immediately preceding. If you haven't been through them, take a few minutes to read the "Missing data" topic, followed by the topic dealing with the "Did-not-see option". Then report back here.

Ready, set, *go?* Have a look at the following CCs lines:

```
*col (c28-c37)
*sub Aff, Name=(Comfort with using LERTAP2), Title=(Comfort), Wt=0
*pol +---- ++--+
*col (c28-c37)
*sub Aff, MDO, Name=(Comfort items with MDO), Title=(MDOCmfrt), Wt=0
*pol +---- ++--+
```

The lines above set out two subtests. Both are affective as the "Aff" control word has been used on each of the *sub cards.

Both subtests involve the same ten items; no doubt you recognize the subtest? Right -- it's the set of Likert-style "Comfort" questions found on the Lertap quiz. And, no doubt you also recall that the items themselves may be see in Appendix A of that best-seller, the Lertap manual? Very good.

The only differences between the two subtests are found in the *sub lines. The second subtest uses the "MDO" control word.

Okay; rig yourself up with a refreshment of some sort, polish your glasses, and have a gander at Lertap's reports for these two subtests.

The Statsf reports

Lertap5 full item stats for "Comfort with using LERTAP2", created: 2

	option	wt.	n	%	pb(r)	avg.	z
Q27							
	1	5.00	3	5.0	0.36	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	-0.24	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

Lertap5 full item stats for "Comfort items with MDO", created: 29/03

	option	wt.	n	%	pb(r)	avg.	z
Q27							
	1	5.00	3	5.0	0.36	41.7	1.57
	2	4.00	14	23.3	0.33	36.8	0.60
	3	3.00	22	36.7	0.05	34.1	0.07
	4	2.00	21	35.0	-0.51	30.2	-0.70
	5	1.00	0	0.0	0.00	0.0	0.00
Q28	option	wt.	n	%	pb(r)	avg.	z
	1	5.00	13	22.4	-0.20	32.2	-0.37
	2	4.00	27	46.6	0.43	36.3	0.46
	3	3.00	10	17.2	-0.22	31.6	-0.49
	4	2.00	8	13.8	-0.13	32.4	-0.33
	5	1.00	0	0.0	0.00	0.0	0.00
	other	0.00	2	3.3	-0.31	26.0	-1.62

The first table above shows item stats for Q27 and Q28 without MDO, while the second table reflects the results of using MDO.

Q27's stats are the same in both tables, are they not? No-one omitted this item, so the statistics are unchanged, aren't they?

No. In fact, they're not unchanged (fooled you, eh?). Everything's the same until we get to the pb(r), avg., and z columns, wherein some changes enter.

To understand why Q27's results differ, look at the "other" row for Q28.

In the first subtest, without MDO, Lertap has wt.=3.00, giving 3.00 points to the two (2) people who did not answer Q28. Not so in the second subtest, where those two people have been stripped of scoring points. There are two different scoring methods in operation here: without MDO, people missing an answer to an item are given points equal to the average value of the wt. figures for the item's options. When MDO is active, as in the second subtest, no points are given when someone omits an item.

The result? The subtest scores will differ. Scores on the first subtest will be higher as people who miss out items are still getting points. The mean (average) of the subtest scores on the first subtest will be higher than that for the second subtest; the point-biserial correlation values, pb(r), between an option and the criterion score, the subtest score, are likely to differ, as are the avg. and z values. The more missing data, the greater these differences are likely to be.

Even though everyone answered item Q27, the criterion measure used to calculate item option statistics, pb(r), avg., and z, differs from the first subtest to the second, generally resulting in different values for item Q27's output.

Now, about Q28. As noted, two people did not answer this question. Compare the values found in the % column for Q28: they're greater in the second subtest. The % figures for Q28 in the second subtest, the one using MDO, have been calculated with n=58, the number of people who actually answered the item. In the first subtest, the % values were calculated with n=60, the total number of people taking the test (survey).

In addition, the pb(r) values for the item options seen in the second subtest have been calculated on a pairwise basis -- they are based only on the people who actually answered the item.

To read a bit more about how Lertap computes the Statsf figures, go back for a look at the "MDO cognitive, Statsf" topic.

There are two main differences between the Statsf reports for cognitive and affective subtests: it is rare for people omitting a cognitive item to get scoring points, so the statistics for cognitive item options may not be noticeably different going from no MDO to MDO. Lertap will apply a correction for inflation to the pb(r) and b(r) values corresponding to the right answer to a cognitive question, but this correction is not applied in the Statsf report for affective items (but it is for Statsb: see below).

The Statsb reports

Lertap5 brief item stats for "Comfort with using LERTAP2", created: 29/03/2

Res =	1	2	3	4	5	other	pol.	mean	s.d.	cor.
Q27	5%	23%	37%	35%			-	2.98	0.88	0.55
Q28	22%	45%	17%	13%		3%	-	3.75	0.94	- 0.14

Lertap5 brief item stats for "Comfort items with MDO", created: 29/03/20										
Res =	1	2	3	4	5	n	pol.	mean	s.d.	cor.
Q27	5%	23%	37%	35%		60	-	2.98	0.88	0.48
Q28	22%	47%	17%	14%		58	-	3.78	0.95	- 0.10

In the normal case, without the MDO option, the Statsb report for cognitive items has an "other" column which indicates the percentage of non responses to an item. This column changes to "n" when MDO is in operation, as may be seen above.

Q27 has no missing data; its mean and s.d. values are the same in both tables. The Q27 cor. figures differ for the reason found earlier in the Statsf reports: the values of the criterion measure, the subtest score, change as we go from no MDO to MDO.

Q28's figures differ almost everywhere. The percentages, mean, s.d., and cor. statistics for the second table, where MDO is having its impact, are all computed using only the responses from the 58 folks who actually answered this item. In some other data analysis systems, such as SPSS, the correlation (cor.) between Q28 and the criterion would be said to done on a pairwise basis: only when a person has data for both variables are that person's results used in the calulations.

Another note about the cor. values found in the Statsb reports: they are corrected for part-whole inflation.

Using the did-not-see option

Suppose the did-not-see option has been turned on, with X used as the did-not-see code. The respective lines in the System worksheet would look like this:

These are Lertap5 system settings. Change them only if you understand them.	System Settings		
	Present setting:	**Allowed settings:**	**Usual setting:**
Use a **did-not-see** code?	yes	yes / no	no
Did-not-see code (single character; may be blank):	X	any char	

Comments / Data / CCs / **System** / Syntax / OldCCs / Problems

Next, have a look at a snippet of Freqs output:

(c37) Q35

Option	n	/60
X	6	10.0%
1	2	3.3%
2	13	21.7%
3	12	20.0%
4	17	28.3%
5	7	11.7%
?	3	5.0%

Six people did not see Q35; three did not answer it.

Okay? Now, suppose MDO is not operating. The Stats1f output for Q35 will look like this:

Lertap5 full item stats for "Comfort with using LERTAP2", created: 30/03/2006.

Q35

option	wt.	n	%	pb(r)	avg.	z
1	1.00	2	3.7	-0.05	30.5	-0.24
2	2.00	13	24.1	-0.16	30.1	-0.28
3	3.00	12	22.2	0.02	33.1	0.04
4	4.00	17	31.5	0.35	37.4	0.51
5	5.00	7	13.0	0.30	40.0	0.79
other	3.00	3	10.0	-0.02	32.0	-0.07

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

If the MDO option is turned on, the output will change:

Lertap5 full item stats for "Comfort items with MDO", created: 30/03/2006.

Q35

option	wt.	n	%	pb(r)	avg.	z
1	1.00	2	3.9	-0.15	30.5	-0.75
2	2.00	13	25.5	-0.51	29.8	-0.88
3	3.00	12	23.5	-0.19	32.6	-0.35
4	4.00	17	33.3	0.39	37.2	0.56
5	5.00	7	13.7	0.44	40.0	1.09
other	0.00	3	15.0	-0.26	29.0	-1.04

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

To grasp what Lertap has done, look down the % column for these two reports.

In the top report, the % values (and the columns to the right, from pb(r) to z) are based on n=54; the six people who did not see Q35 have been excluded from the calculations.

In the next report we've got MDO operating, and now we'll have n= what? Fifty-one (51). In this case, the report excludes the six did-not-sees, and the three did-not-answers.

Right. What about the corresponding Statsb reports? Thought you wouldn't ask. Here they be:

Lertap5 brief item stats for "Comfort with using LERTAP2", created: 30/0

Res =	1	2	3	4	5	n	pol.	mean	s.d.	cor.
Q35	4%	24%	22%	31%	13%	54	+	3.26	1.07	0.94

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

Lertap5 brief item stats for "Comfort items with MDO", created: 30/03/20

Res =	1	2	3	4	5	n	pol.	mean	s.d.	cor.
Q35	4%	25%	24%	33%	14%	51	+	3.27	1.10	0.56

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

Where do you stand now? You see what happens, or are you in the did-not-see group?

Questions? Crank up your email program, and point it at: support@lertap.com.

3.7 Cognitive subtest statistics

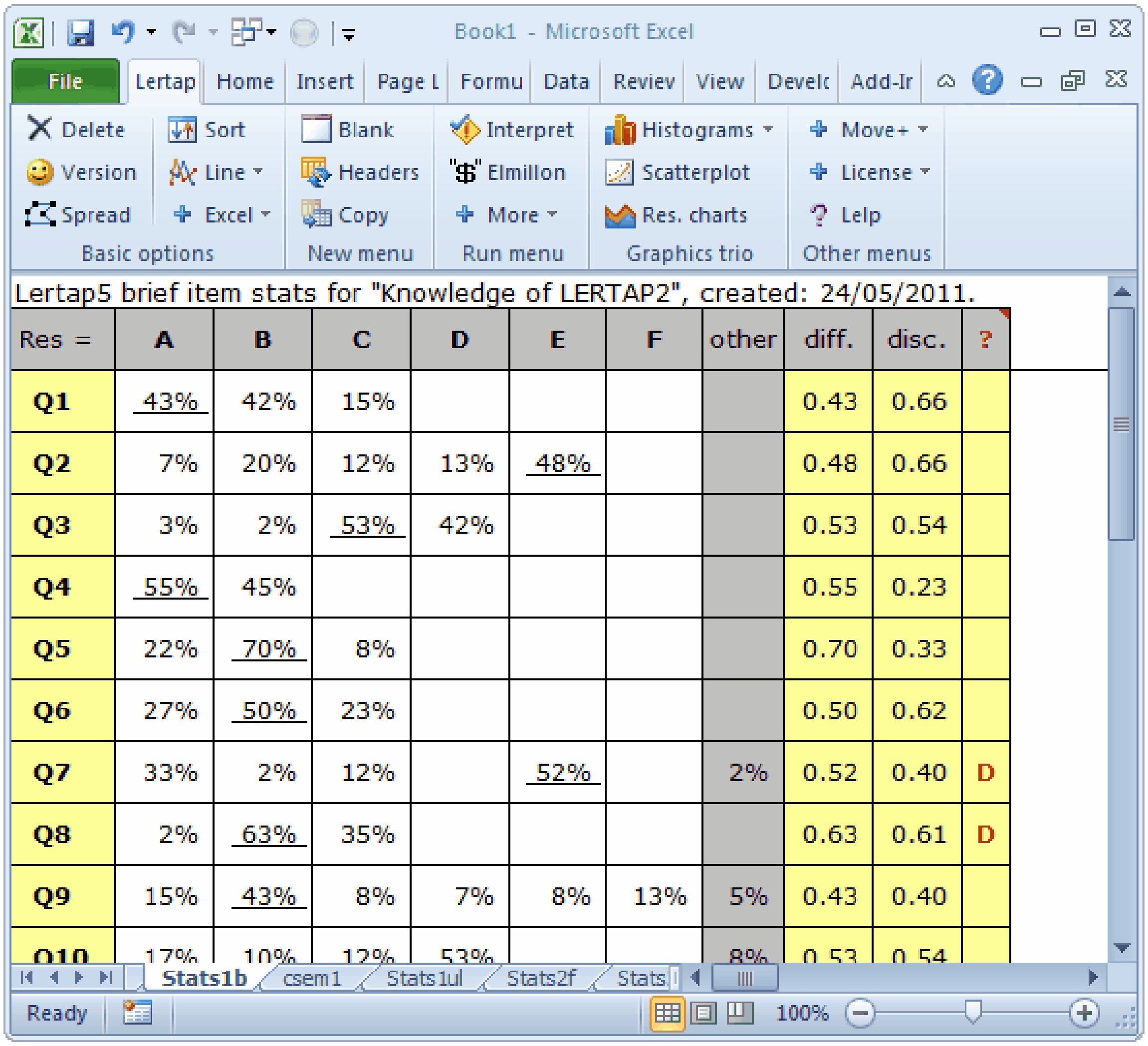
Three worksheets of statistics are usually produced for each cognitive subtest. These are referred to as the “full”, “brief”, and “upper-lower” sheets. The information provided by these sheets often overlaps to a considerable extent, as you’re about to see.

In the case of our running example, the Cook's tour, 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 other sheets of this ilk, named “Stats2f”, “Stats2b”, "Stats3f", and "Stats3b". These relate to the second subtest, the affective one, which has been processed twice (and scored differently each time).

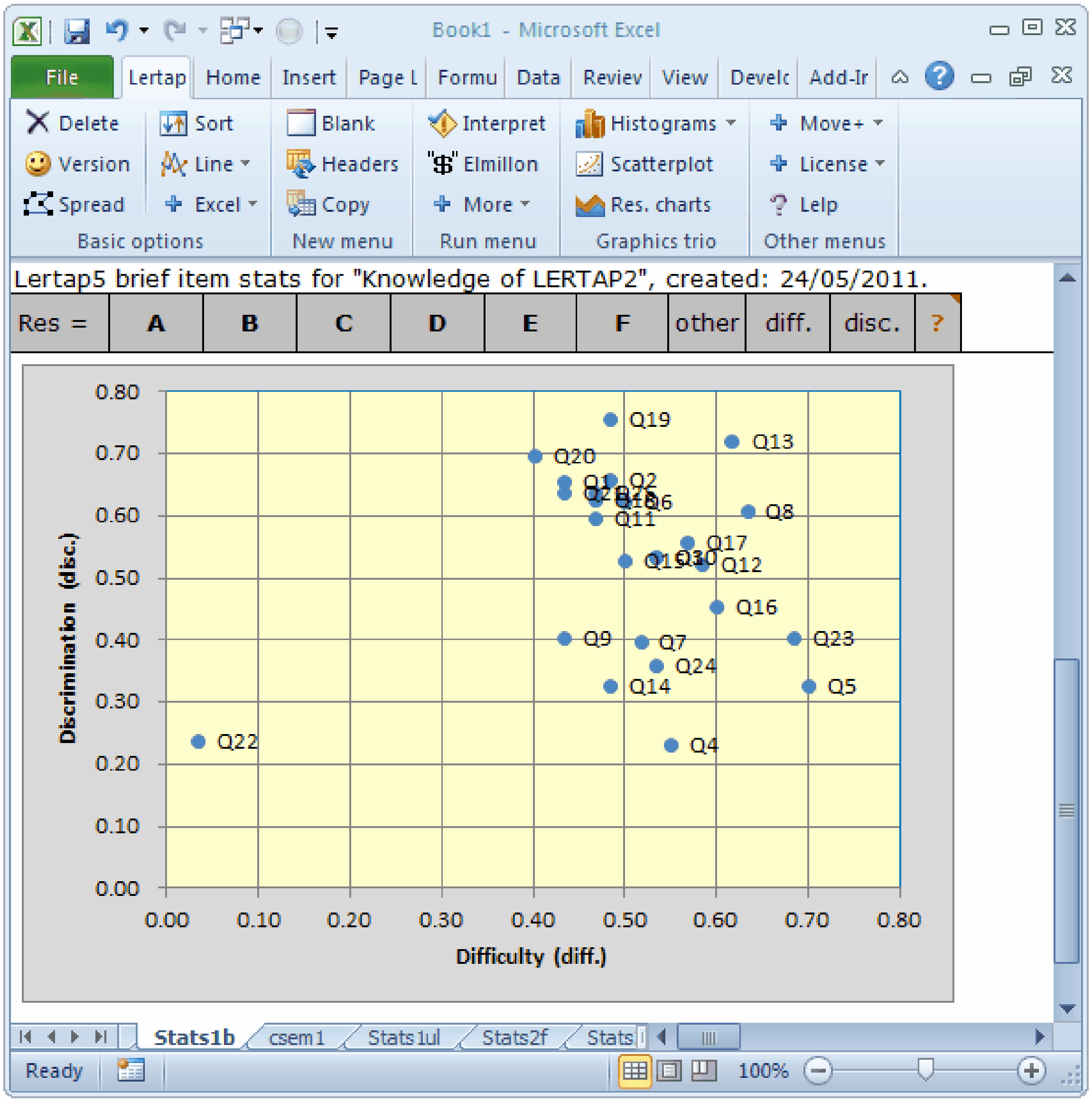
Page ahead, the coast is clear.

3.7.1 Brief statistics (cognitive)

Have a look at the Stats1b worksheet, top and bottom:



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	



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 (that is, "responses" which do not correspond to an item's options, including non responses and data processing errors). 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 options might be regarded as rendering questionable service. Read much more about the **? column** in an earlier topic, "MathsQuiz".

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. and disc. columns is discussed in following topics. The information in the Stats1f sheet is actually taken from lines found in another of the three statistical summaries, the "full stats" sheet, Stats1f.

The scatterplot found at the bottom of Stats1b is based on the Diff. and Disc. columns. The interpretation of such scatterplots is discussed in numerous other topics, particularly "M.Nursing". In this case, the scatterplot is quite a good one. Only two items, Q22 and Q4, have discrimination figures below 0.30. Q22 had a diff. value of just 0.03, meaning that it's a difficult item, correctly answered by only 3%. Otherwise, the diff. values are good.

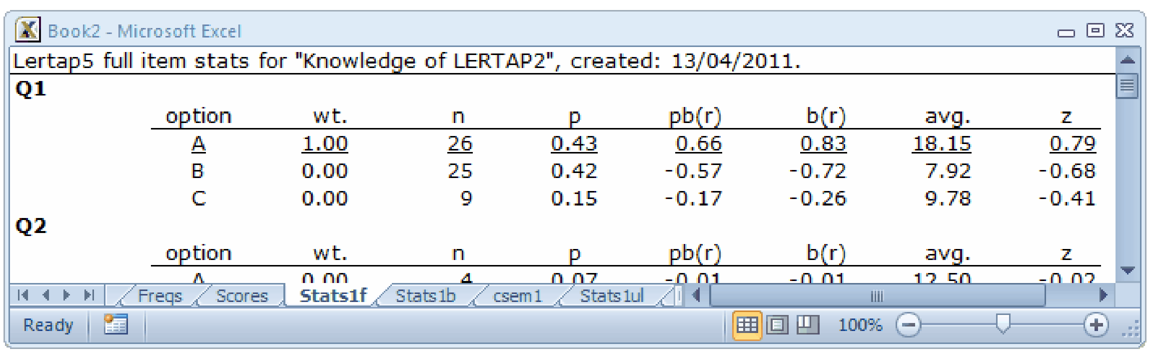
(Note that in making these comments about item disc. and diff., we're assuming that we want to have a discriminating test, one meant to identify the strong and the weak students. This isn't always the case at all. At times all we want to do is see how a class did on the items, looking for items which may have been particularly difficult, information which we could then use to review course content with the whole class.)

Full on!

3.7.2 Full statistics (cognitive)

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

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



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.

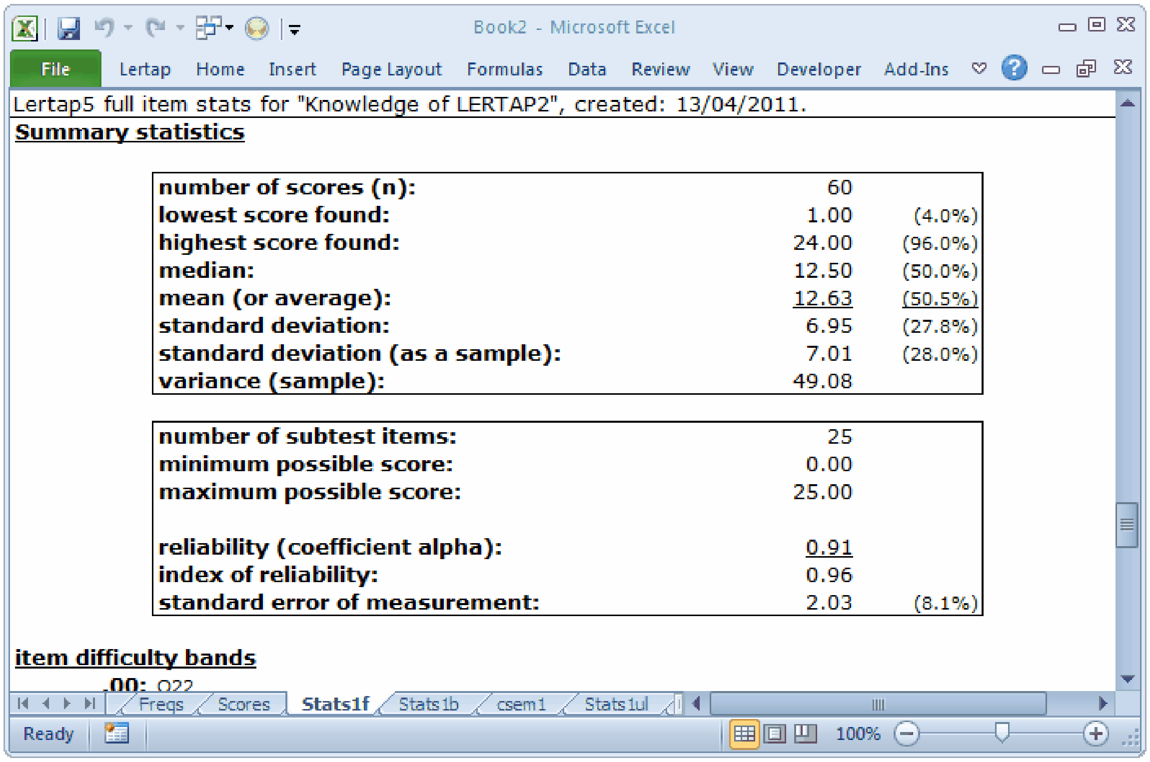
Lertap's default criterion score is an internal one, equal to the subtest score. It is possible to set up an external criterion analysis via one of the toolbar's Run options.

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—they should 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:



Lertap5 full item stats for "Knowledge of LERTAP2", created: 13/04/2011.

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	

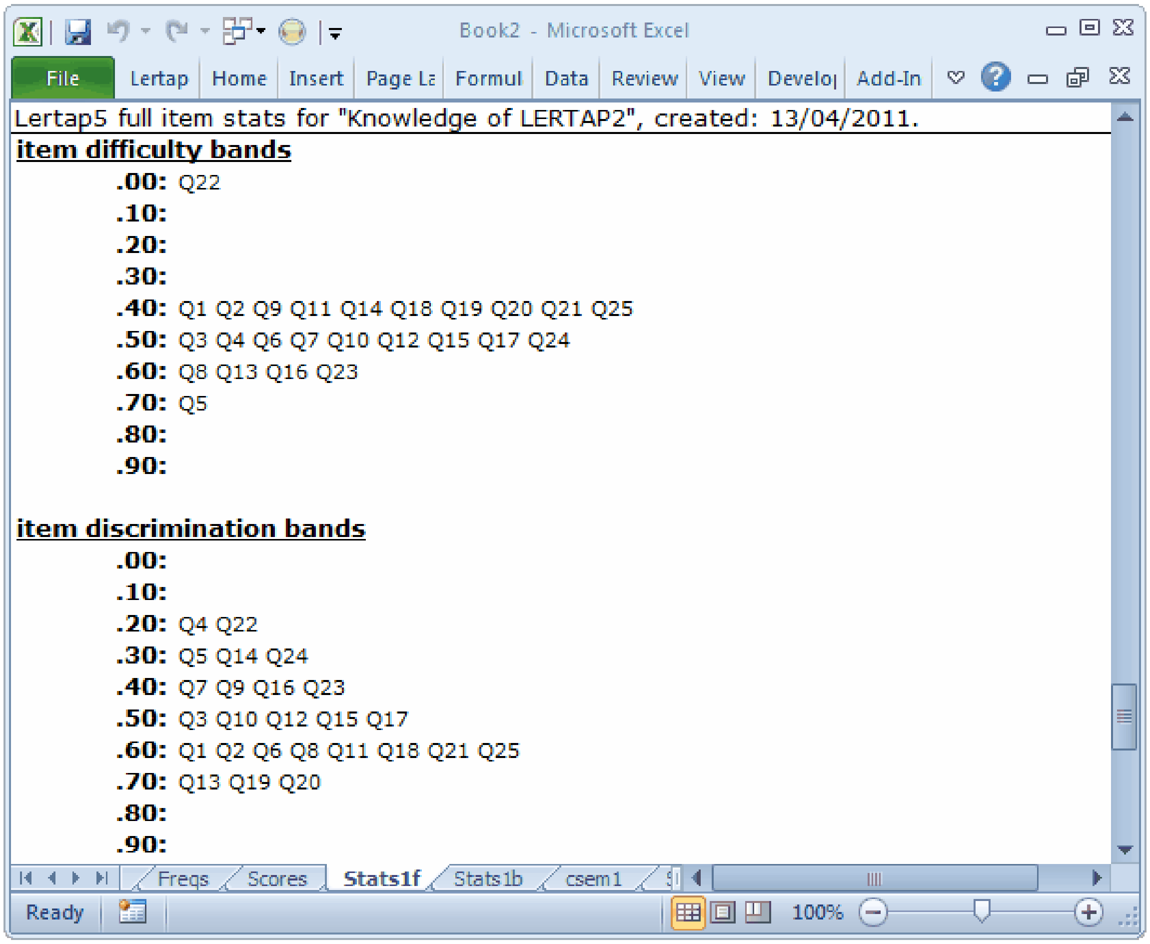
number of subtest items:	25	
minimum possible score:	0.00	
maximum possible score:	25.00	
reliability (coefficient alpha):	0.91	
index of reliability:	0.96	
standard error of measurement:	2.03	(8.1%)

item difficulty bands

.00: Q22

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”:



Lertap5 full item stats for "Knowledge of LERTAP2", created: 13/04/2011.

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

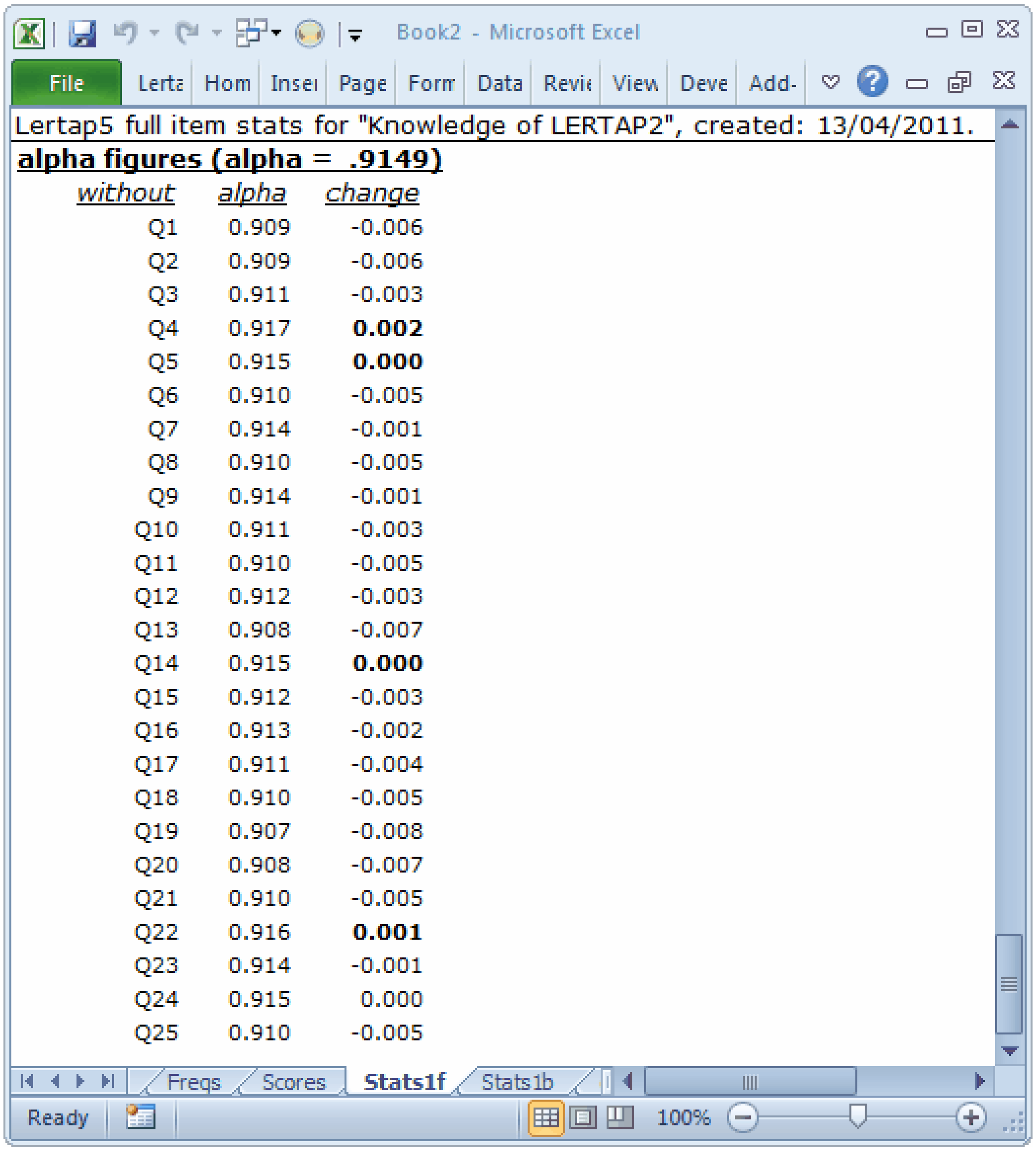
.00:
.10:
.20: Q4 Q22
.30: Q5 Q14 Q24
.40: Q7 Q9 Q16 Q23
.50: Q3 Q10 Q12 Q15 Q17
.60: Q1 Q2 Q6 Q8 Q11 Q18 Q21 Q25
.70: Q13 Q19 Q20
.80:
.90:

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:



Lertap5 full item stats for "Knowledge of LERTAP2", created: 13/04/2011.

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**
Q6	0.910	-0.005
Q7	0.914	-0.001
Q8	0.910	-0.005
Q9	0.914	-0.001
Q10	0.911	-0.003
Q11	0.910	-0.005
Q12	0.912	-0.003
Q13	0.908	-0.007
Q14	0.915	**0.000**
Q15	0.912	-0.003
Q16	0.913	-0.002
Q17	0.911	-0.004
Q18	0.910	-0.005
Q19	0.907	-0.008
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.

Next, uppers and ~~downers~~ lowers.

3.7.3 Upper-lower statistics

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 was 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—separate test results into groups, one with the strongest performers, one with the weakest, and, maybe, the group in-between—and then look at item-level results in each group.

Now, our Cook's Tour of Lertap has been around for a fair number of years. Lertap has changed since the Cook's Tour first emerged.

One of the largest changes has to do with the Stats1ul report. An example of the current Stats1ul report is seen back at the MathsQuiz topic, and we recommend it (would you expect us not to?). Once you're into the MathsQuiz example, you'll also get to learn about the csem1 report made by Lertap.

Affective item reports are next.

3.8 Affective subtest statistics

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".

If you've been faithfully following the Cook's Tour, you will have found that Lertap has made two reports for the second subtest, Stats2f plus Stats2b, and then two reports for the third subtest, Stats3f plus Stats3b. This is because the CCs worksheet for the example we've been following has three 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 two affective subtests came next; the first one of these, which is really the second subtest in the overall picture, has its results in the Stats2f and Stats2b reports. Why 2? Because it's the second in line.

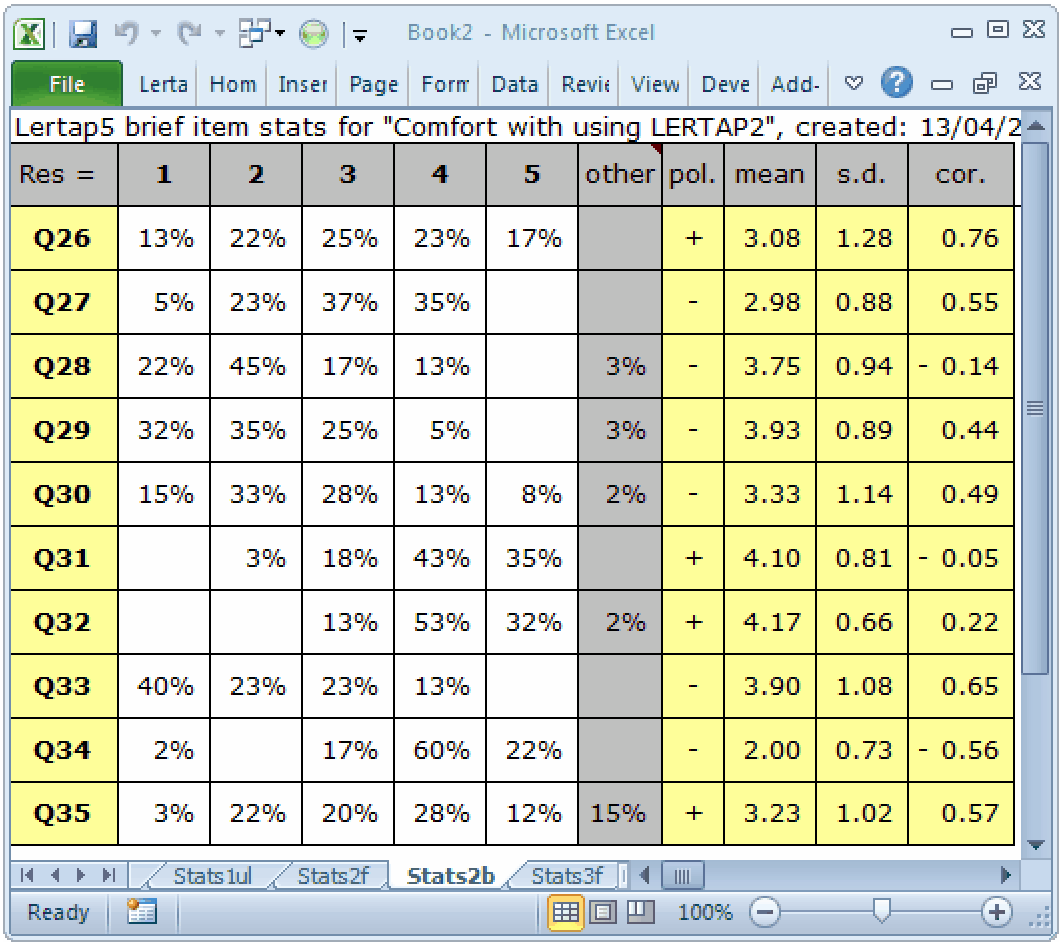
Bet you can't guess what the two reports for the third subtest are named? Hint: **not** Stats4f and Satas4b; if that was your guess you need a big break from this stuff.

The original Cook's Tour, the one mentioned in Chapter 2 of the manual, did not have the third subtest. For your information, what led us to introduce it, the third subtest, was a series of requests from users, particularly one in Australia, and a small smattering in the United States. They wanted a change in the way Lertap processed affective test data. We modified the **MDO** option in order to silence them. Read what we did here (you'll be gone for a while if you follow this link, but it'll be ever so worth your time).

Okay, ready to be affected by some affective reports? Click.

3.8.1 Brief statistics (affective)

Have a squiz—here are the brief results for the "Comfort" affective subtest (or "scale"—a collection of related affective items is often referred to as a **scale**):



Lertap5 brief item stats for "Comfort with using LERTAP2", created: 13/04/2

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

The Stats2b worksheet shown above has its row and column headers hidden at Lertap's request—you can turn them back on easily.

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.

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.

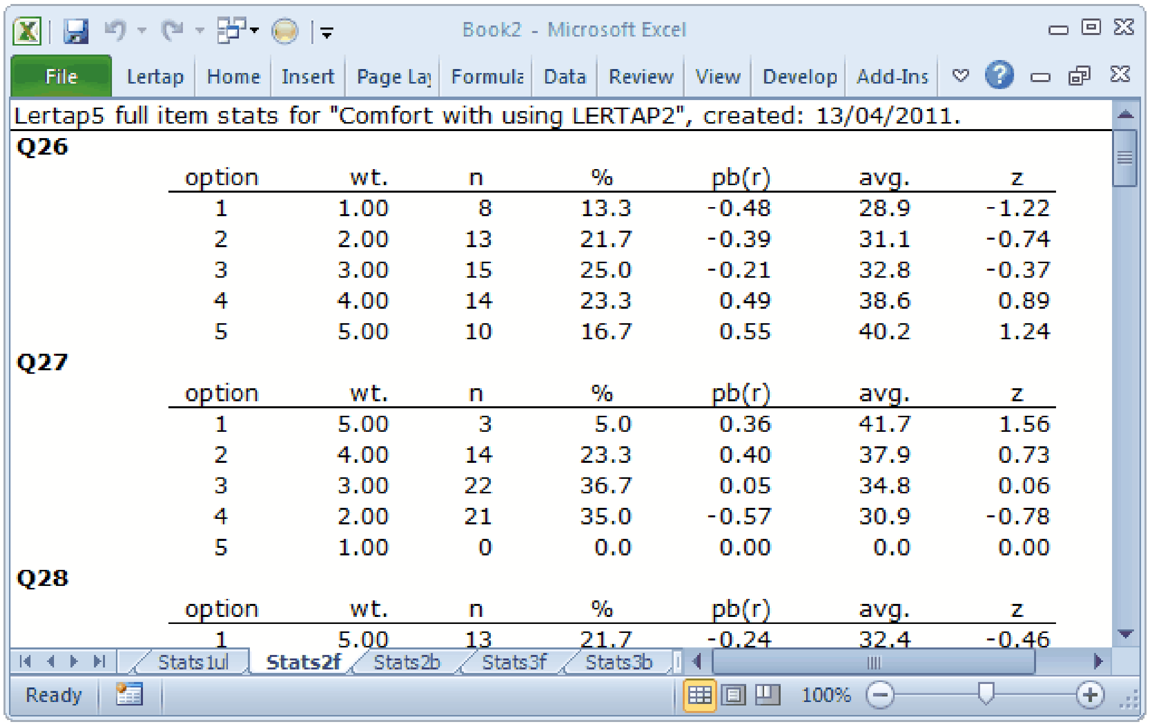
The average of the Q26 responses, and their standard deviation, are found under the "**mean**" and "**s.d.**" columns, respectively. The "**cor.**" column gives the value of the Pearson product-moment correlation coefficient between the item and the criterion score. At this point the criterion score 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. Next?

3.8.2 Full statistics (affective)

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:


Book2 - Microsoft Excel

File | Lertap | Home | Insert | Page Lay | Formula | Data | Review | View | Develop | Add-Ins

Lertap5 full item stats for "Comfort with using LERTAP2", created: 13/04/2011.

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	0.55	40.2	1.24

Q27

option	wt.	n	%	pb(r)	avg.	z
1	5.00	3	5.0	0.36	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	-0.24	32.4	-0.46

Stats1ul | Stats2f | Stats2b | Stats3f | Stats3b

Ready 100%

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.

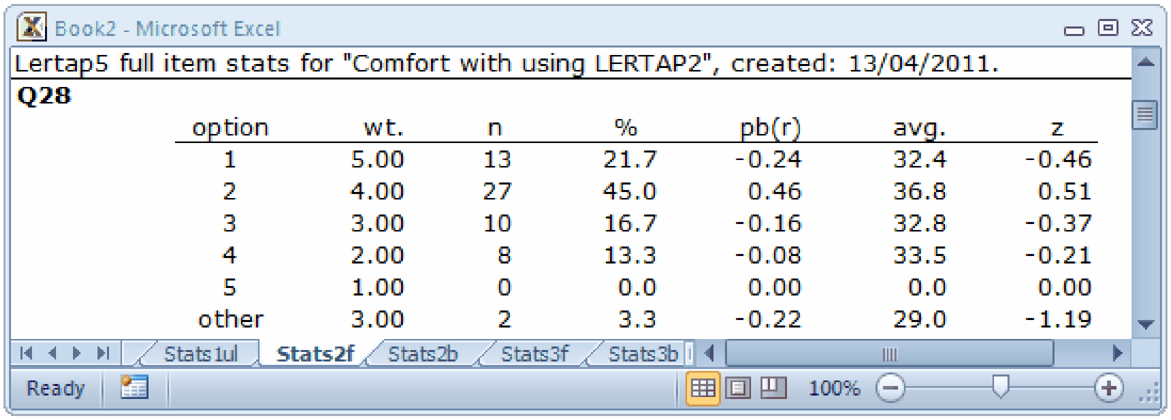
The pb(r) column gives the point-biserial correlation of each option with the criterion score, that is, the subtest score. At this level, the pb(r) figure is <u>not</u> corrected for part-whole inflation—it is for cognitive items, but not for affective ones, and this subtest, Comfort, is affective.

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 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:


Book2 - Microsoft Excel

Lertap5 full item stats for "Comfort with using LERTAP2", created: 13/04/2011.

Q28

option	wt.	n	%	pb(r)	avg.	z
1	5.00	13	21.7	-0.24	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

Stats1ul Stats2f Stats2b Stats3f Stats3b

Ready 100%

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.

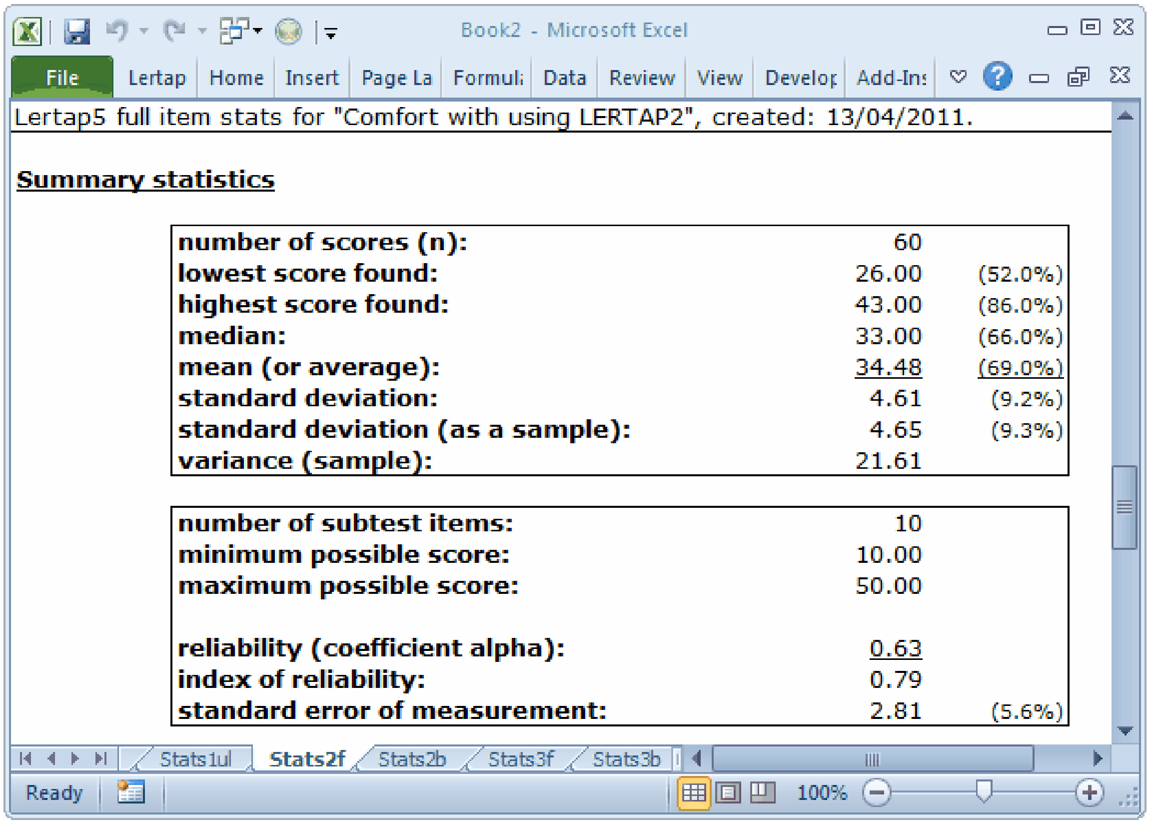
The reason Lertap assigns a weight, or score, for "others" is to try and keep the response summaries "honest". If "others" got no weight, as is the case for cognitive items, then the item mean would be lowered, and, if users just looked at results in the brief stats sheet, Stats2b, a false impression 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 always assumed to be "on", and the MDO control word on the *sub card extinguishes it. Read more MDO to be in the know.

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

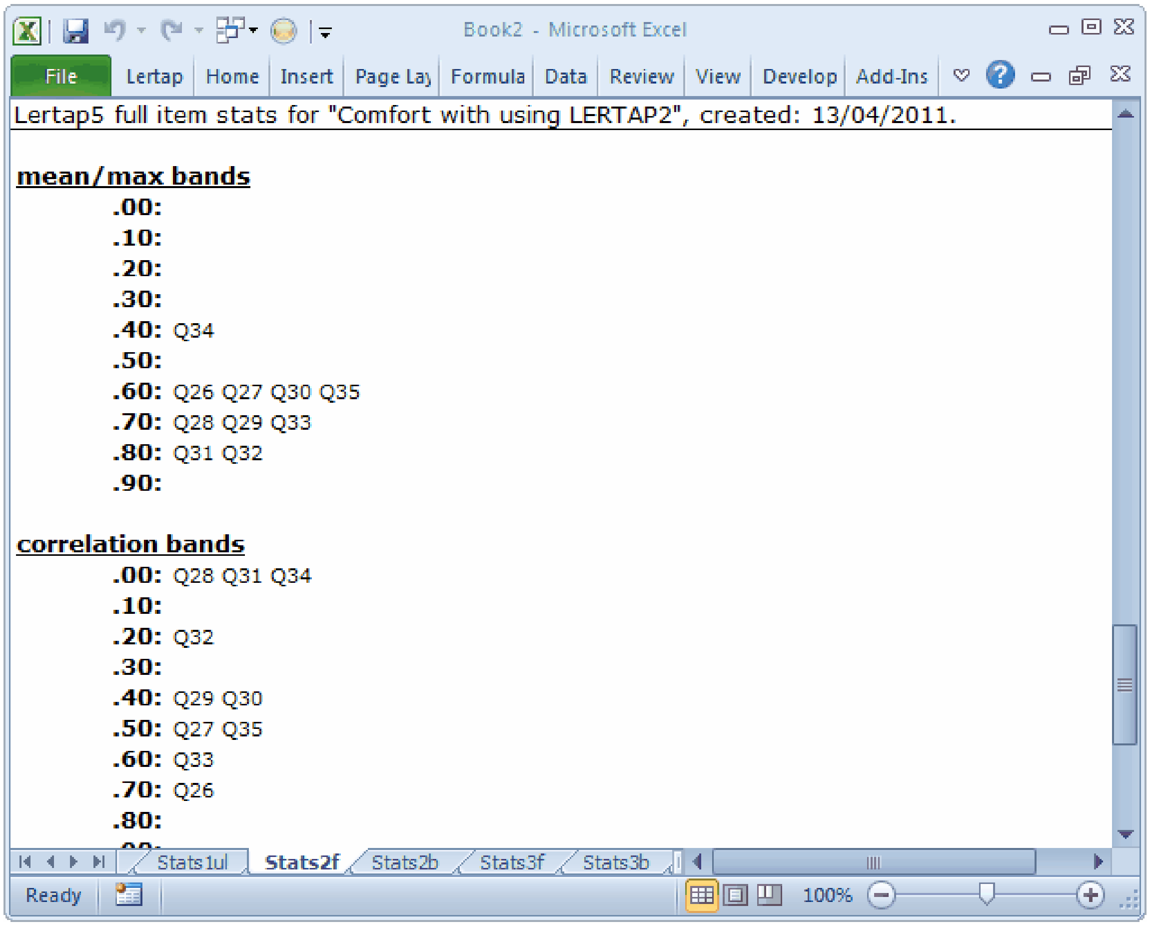


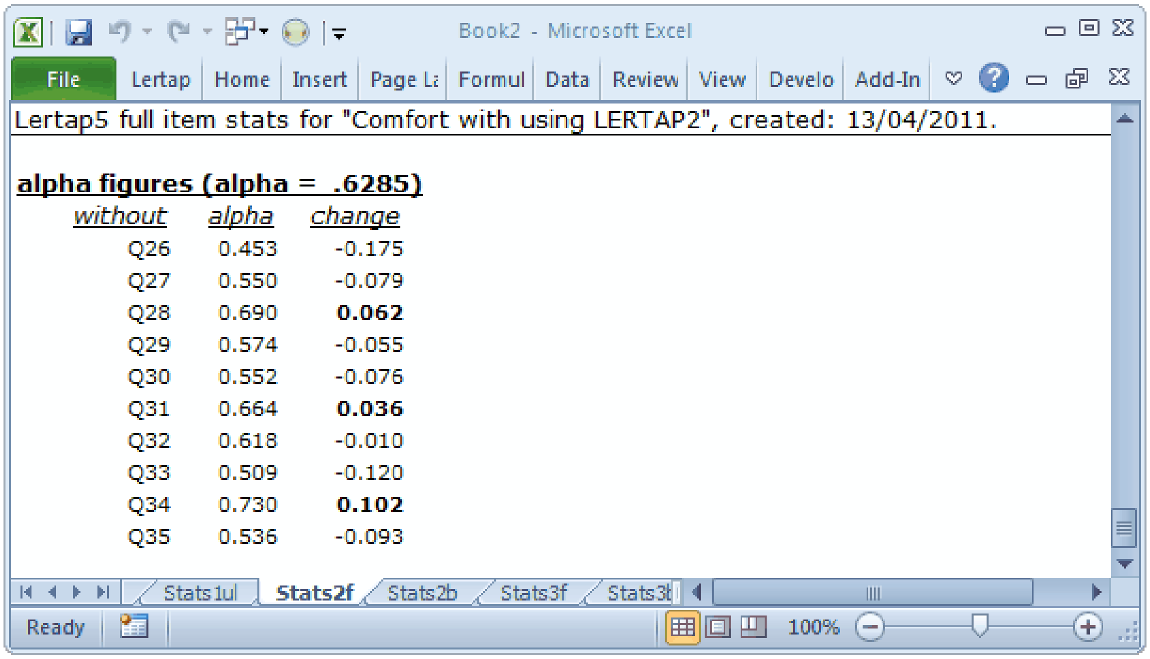
Lertap5 full item stats for "Comfort with using LERTAP2", created: 13/04/2011.

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%)


Book2 - Microsoft Excel
File
Lertap
Home
Insert
Page Lay
Formula
Data
Review
View
Develop
Add-Ins
Lertap5 full item stats for "Comfort with using LERTAP2", created: 13/04/2011.
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:
Stats1ul
Stats2f
Stats2b
Stats3f
Stats3b
Ready
100%



Lertap5 full item stats for "Comfort with using LERTAP2", created: 13/04/2011.

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. There are times when the information pertaining to an item is of great interest, and reliability, at times, is not a major focus. Another topic in this document, the CEQ survey, has a collection of survey items with low reliability. The CEQ researcher based her interpretation of survey results only on the item statistics, ignoring reliability.

Coming to the end of the Cook's Tour.

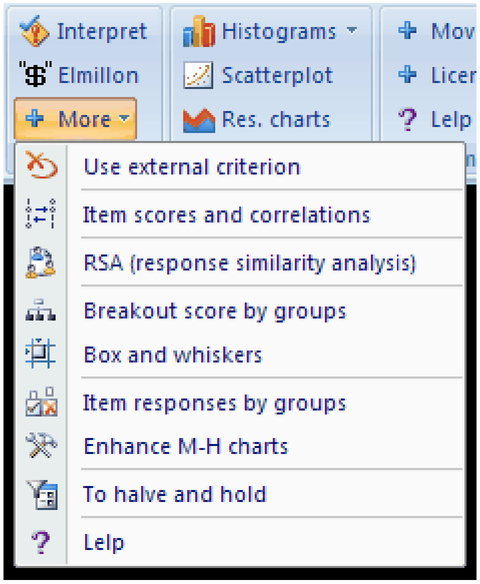
3.9 Our research questions

Back a few topics, 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.

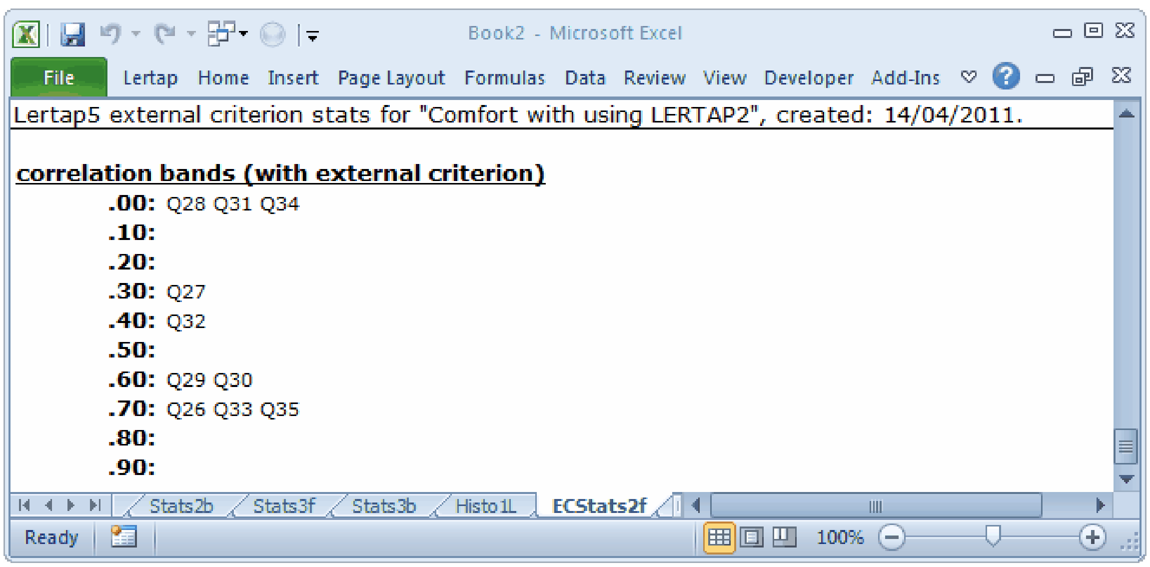
Did the subtests have adequate reliability? Well, the cognitive subtest, "Knwldge", came through with an alpha reliability of 0.91, standard error of measurement of 8.1% (or 2.03 in raw score terms), which is not bad. The affective subtest, "Comfort", fared less well, having an alpha figure of 0.63, 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 (often not a bad compromise at all).

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.

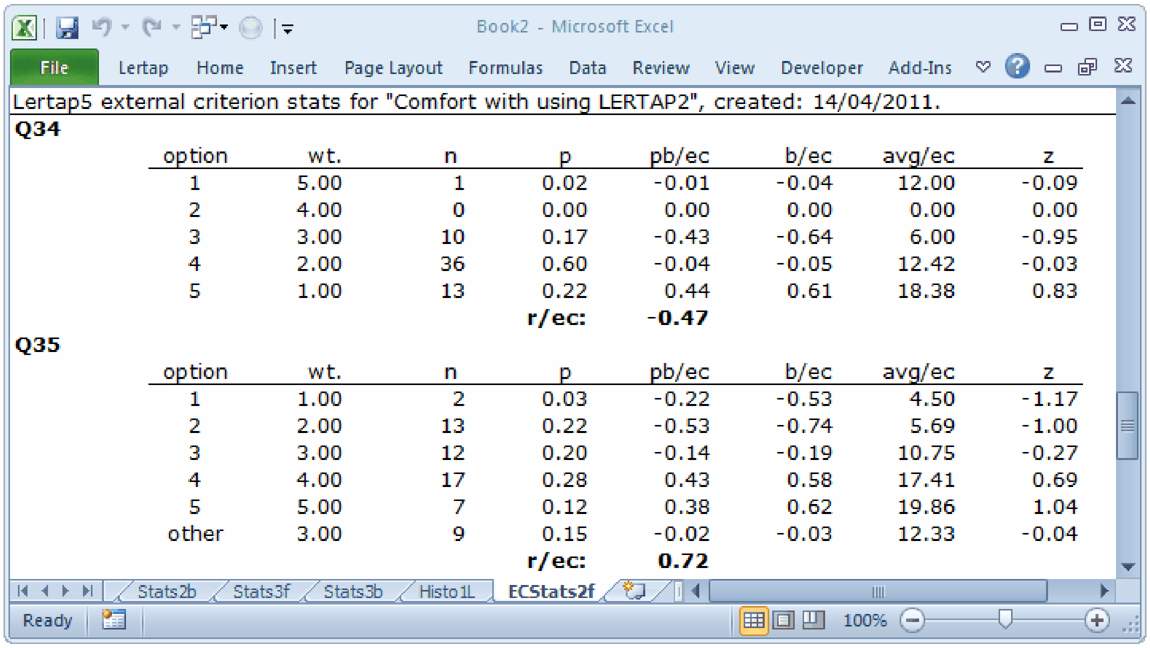
We could use Lertap's support for **external-criterion** analyses to dig a bit more here, asking for correlations between each of the Comfort items with the overall Knwldge score. We go up to the Lertap tab, to the **Run menu**, then to the **+ More** option, and, finally, a click on "Use external criterion". 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 a new report called ECStats2f, with item-level summaries of correlations, and the “correlation bands (with external criterion)”:



Items Q28, Q31 and Q34 actually had negative correlations with the Knwldge score, something readily visible by just scrolling up to the item-level data in the ECStats2f report. For example, here's a snapshot of Q34 (and Q35), with r/ec for Q34 at -0.47:



Lertap5 external criterion stats for "Comfort with using LERTAP2", created: 14/04/2011.

Q34

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
1	5.00	1	0.02	-0.01	-0.04	12.00	-0.09
2	4.00	0	0.00	0.00	0.00	0.00	0.00
3	3.00	10	0.17	-0.43	-0.64	6.00	-0.95
4	2.00	36	0.60	-0.04	-0.05	12.42	-0.03
5	1.00	13	0.22	0.44	0.61	18.38	0.83
			r/ec:	**-0.47**			

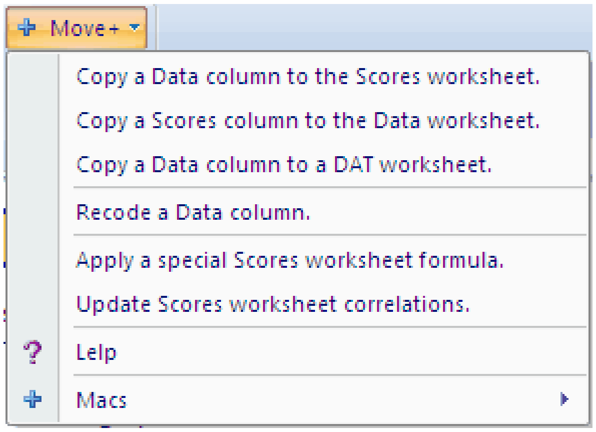
Q35

option	wt.	n	p	pb/ec	b/ec	avg/ec	z
1	1.00	2	0.03	-0.22	-0.53	4.50	-1.17
2	2.00	13	0.22	-0.53	-0.74	5.69	-1.00
3	3.00	12	0.20	-0.14	-0.19	10.75	-0.27
4	4.00	17	0.28	0.43	0.58	17.41	0.69
5	5.00	7	0.12	0.38	0.62	19.86	1.04
other	3.00	9	0.15	-0.02	-0.03	12.33	-0.04
			r/ec:	**0.72**			

(r/ec stands for correlation with external criterion.)

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 tab on the Excel ribbon 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 "Use external criterion" 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. This might be done, for example, if a program such as SPSS were to be used for more complex data analyses.

Here end-ith the Cook's Tour. *Hope you enjoyed it.* Questions? Just write: support@lertap.com.

Go back to the opening page if you'd like: click.

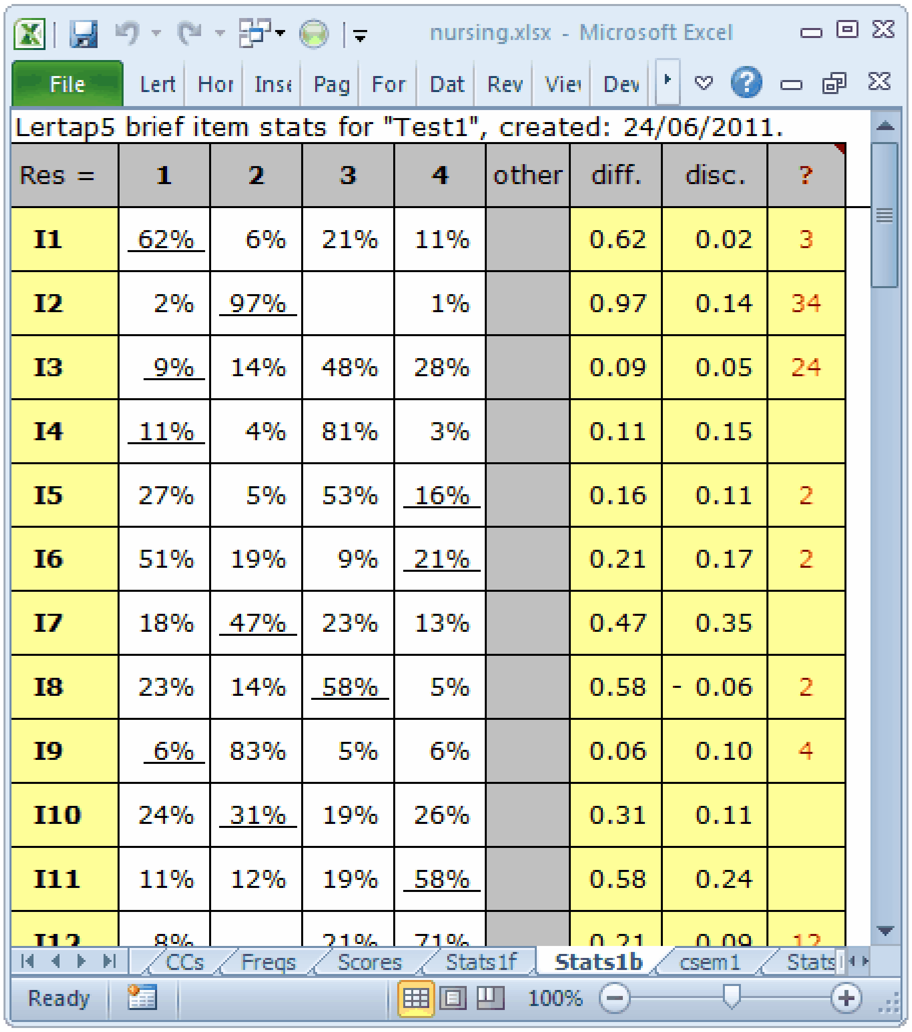
Tidbits:

For more about external criteria, see the "Using an external criterion" topic in Chapter 8 of the manual.

4 Printing results

Excel has good to very good printing capabilities, and these days it seems easier than ever to pick up something in Excel and paste it into Word or PowerPoint.

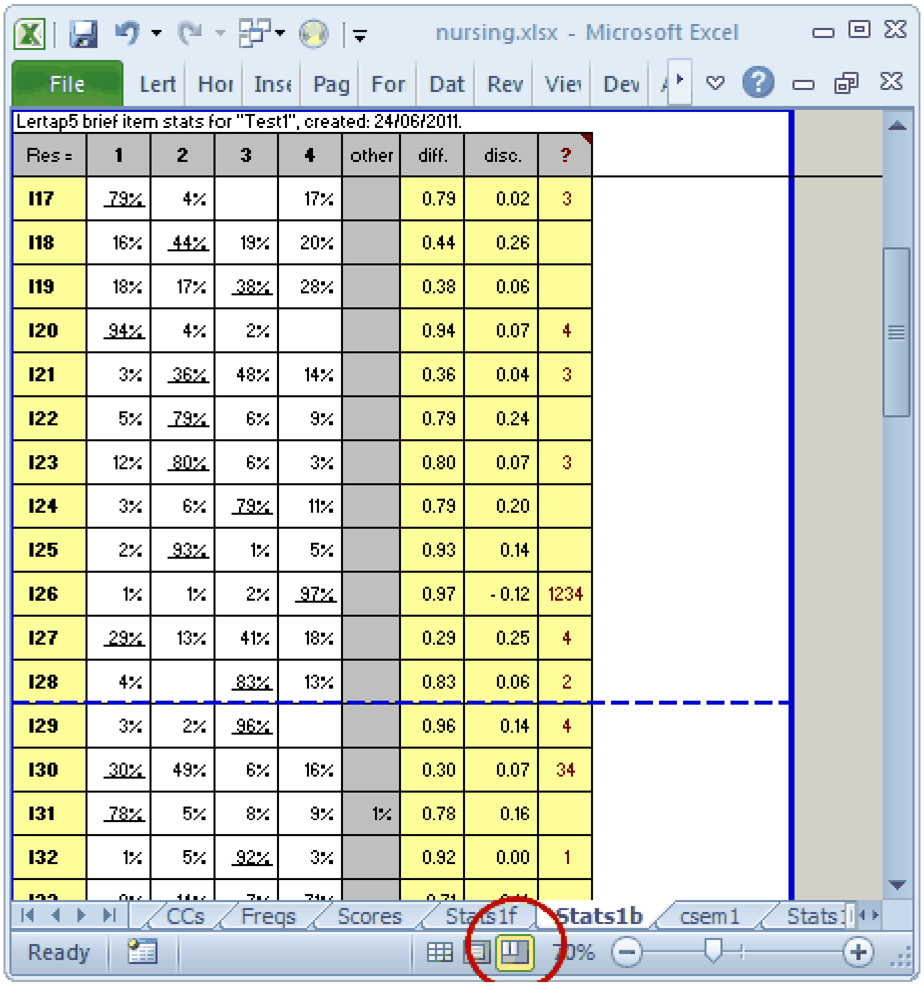
Let's start with Lertap's Stats1b report, with a thought to printing.



Res =	1	2	3	4	other	diff.	disc.	?
I1	62%	6%	21%	11%		0.62	0.02	3
I2	2%	97%		1%		0.97	0.14	34
I3	9%	14%	48%	28%		0.09	0.05	24
I4	11%	4%	81%	3%		0.11	0.15	
I5	27%	5%	53%	16%		0.16	0.11	2
I6	51%	19%	9%	21%		0.21	0.17	2
I7	18%	47%	23%	13%		0.47	0.35	
I8	23%	14%	58%	5%		0.58	- 0.06	2
I9	6%	83%	5%	6%		0.06	0.10	4
I10	24%	31%	19%	26%		0.31	0.11	
I11	11%	12%	19%	58%		0.58	0.24	
I12	8%		21%	71%		0.21	0.09	12

The Stats1b report is one of the easiest to print. How? Well, as per usual, right? Go to File, and select Print. Of course.

But Stats1b can go on for many lines, more than what will fit on a single printed page. What if you wanted to control where pages break?


Lertap5 brief item stats for "Test1", created: 24/06/2011.

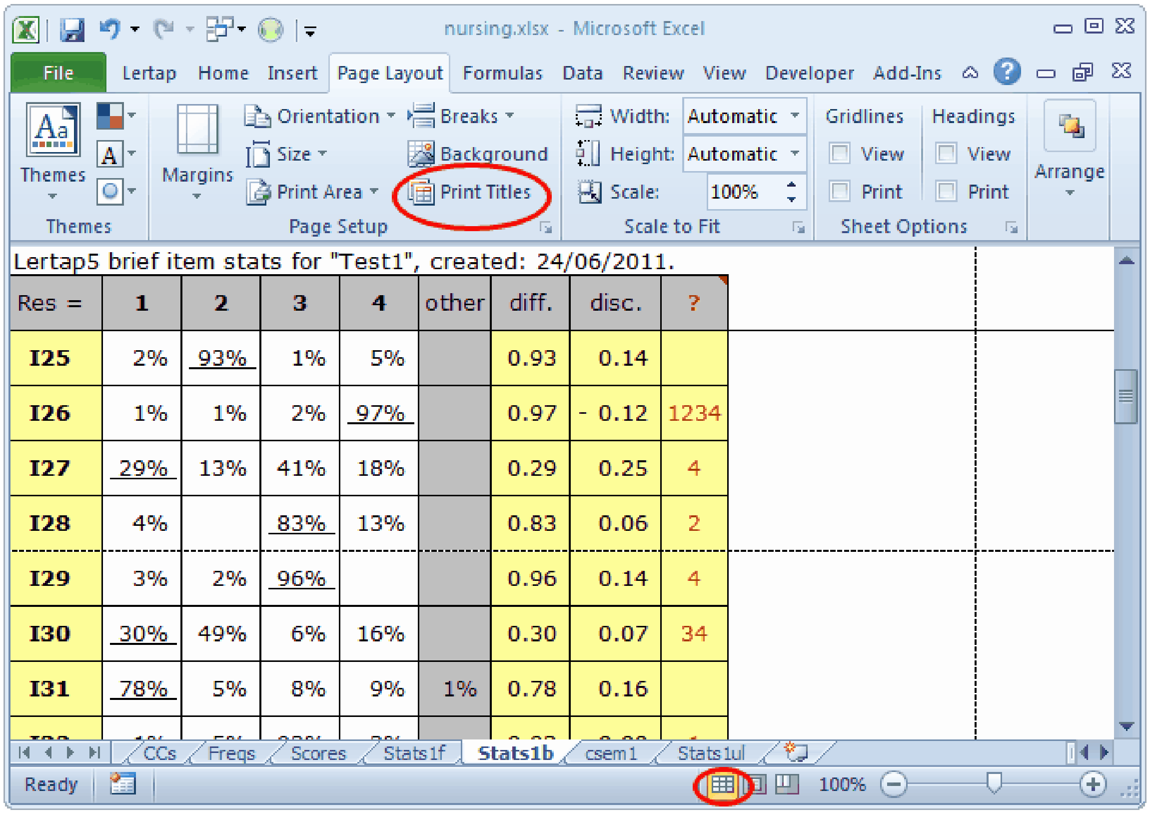
Res =	1	2	3	4	other	diff.	disc.	?
I17	79%	4%		17%		0.79	0.02	3
I18	16%	44%	19%	20%		0.44	0.26	
I19	18%	17%	38%	28%		0.38	0.06	
I20	94%	4%	2%			0.94	0.07	4
I21	3%	36%	48%	14%		0.36	0.04	3
I22	5%	79%	6%	9%		0.79	0.24	
I23	12%	80%	6%	3%		0.80	0.07	3
I24	3%	6%	79%	11%		0.79	0.20	
I25	2%	93%	1%	5%		0.93	0.14	
I26	1%	1%	2%	97%		0.97	-0.12	1234
I27	29%	13%	41%	18%		0.29	0.25	4
I28	4%		83%	13%		0.83	0.06	2
I29	3%	2%	96%			0.96	0.14	4
I30	30%	49%	6%	16%		0.30	0.07	34
I31	78%	5%	8%	9%	1%	0.78	0.16	
I32	1%	5%	92%	3%		0.92	0.00	1

Click on the page-break icon, circled in red. Then just drag the dotted line to where you want the break to be. Neat (Excel has worked this way for many years, but it hasn't always been so easy to get page breaks to show).

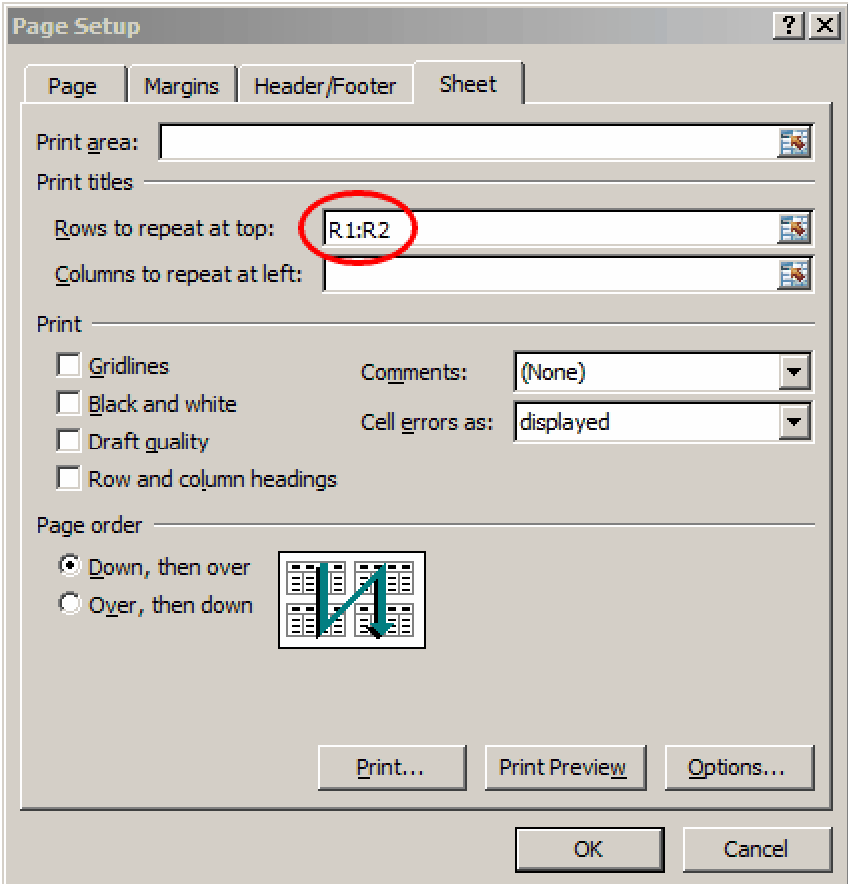
Now, as easy as this is, there's going to be a small problem. In this case, the page break is going to come after I28. The next page will begin with I29's results, but there won't be any column headers.

It would be nice to have Lertap's first two rows carry over to every printed page, wouldn't it? That way we'll have the column headers, and we won't have to dart back to the first page to remember which column has diff. values, and which has disc. figures.

Not hard to do.



Make sure that the view has switched back to normal (small red circle at bottom). Then, on the Page Layout tab, click on Print Titles.



And that'll do it. Every printed page will now have headers from Stats1b's first two rows, R1:R2.

The Stats1f report, and also the Stats1ul report, are similar to Stats1b in that they're rather amenable to printing, even though they are substantially more verbose than Stats1b.

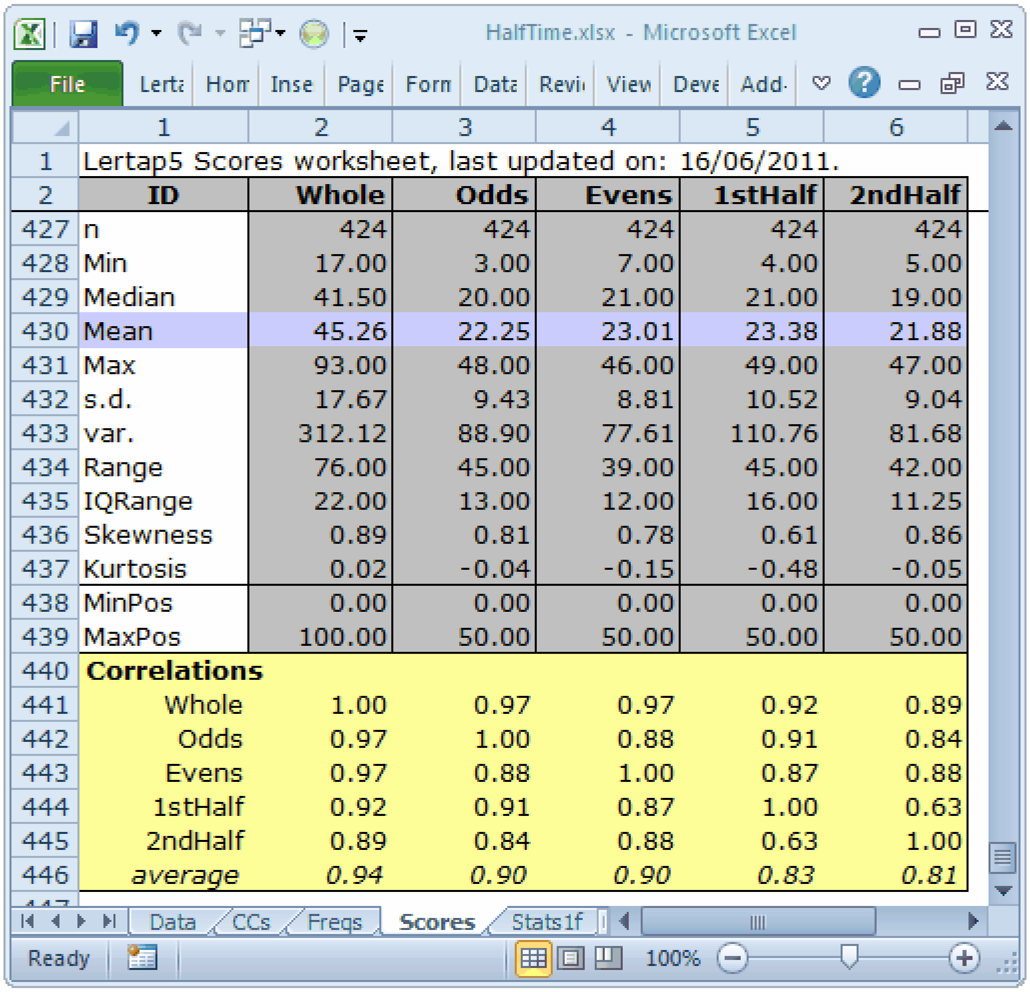
But there are many other Lertap reports which might have several rows of information above the bits which we really want to print.

Among these would be, for example, the Scores report, and Breaks1bw. See the next topic for comments on how to print these.

4.1 Print partial reports

It will often be the case that only parts of one of Lertap's reports will be considered useful for printing.

For example, here's the lower part of a typical Lertap Scores report:

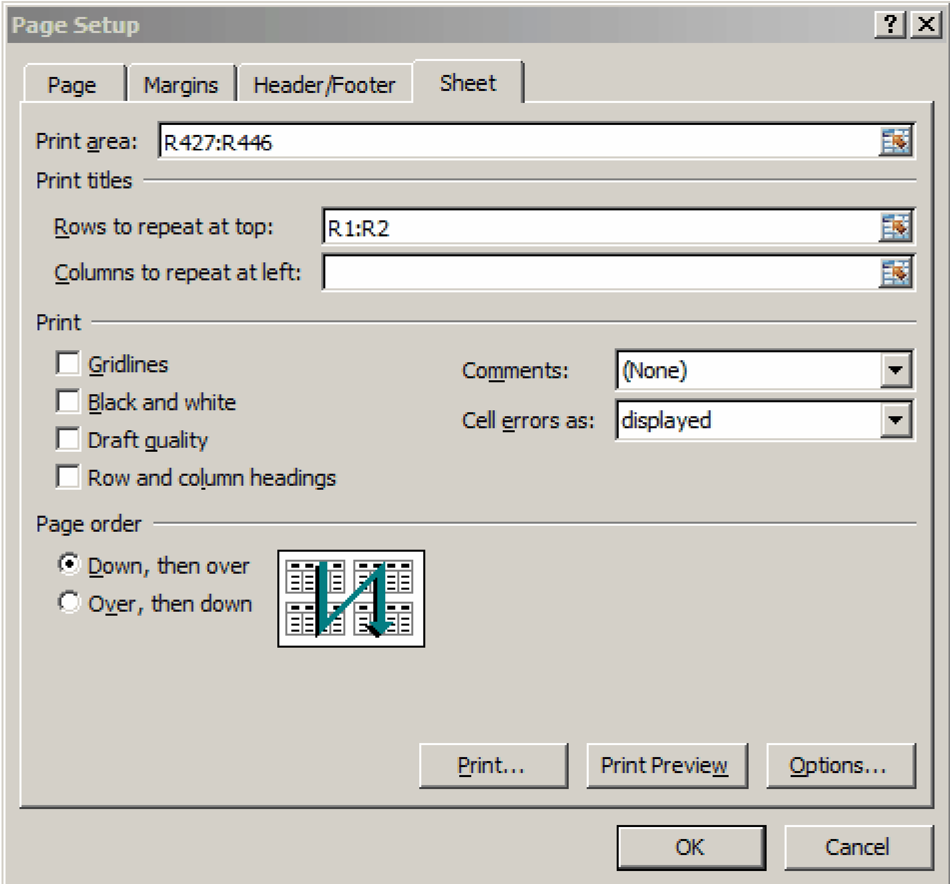


Lertap5 Scores worksheet, last updated on: 16/06/2011.

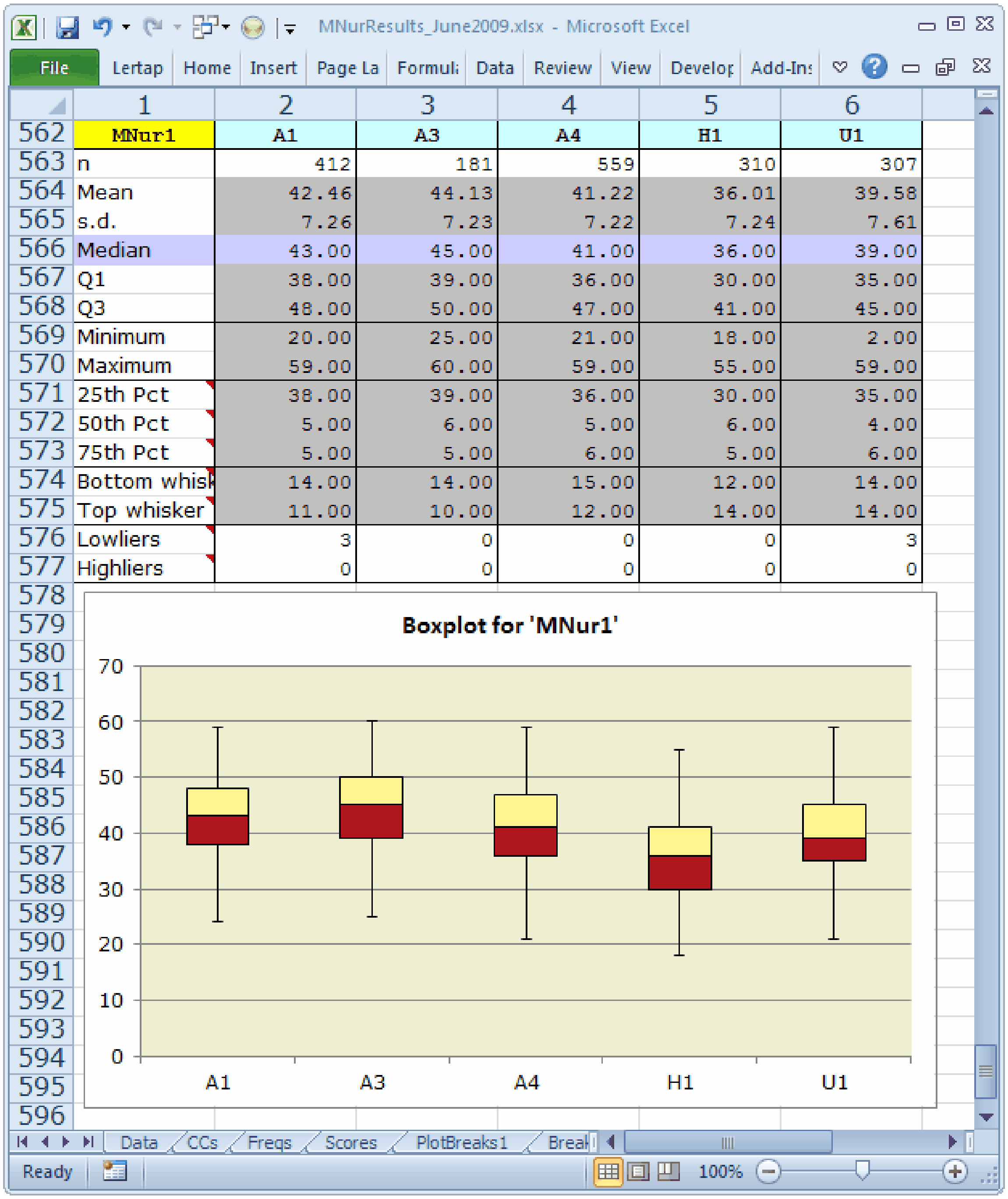
	ID	Whole	Odds	Evens	1stHalf	2ndHalf
427	n	424	424	424	424	424
428	Min	17.00	3.00	7.00	4.00	5.00
429	Median	41.50	20.00	21.00	21.00	19.00
430	Mean	45.26	22.25	23.01	23.38	21.88
431	Max	93.00	48.00	46.00	49.00	47.00
432	s.d.	17.67	9.43	8.81	10.52	9.04
433	var.	312.12	88.90	77.61	110.76	81.68
434	Range	76.00	45.00	39.00	45.00	42.00
435	IQRange	22.00	13.00	12.00	16.00	11.25
436	Skewness	0.89	0.81	0.78	0.61	0.86
437	Kurtosis	0.02	-0.04	-0.15	-0.48	-0.05
438	MinPos	0.00	0.00	0.00	0.00	0.00
439	MaxPos	100.00	50.00	50.00	50.00	50.00
440	**Correlations**					
441	Whole	1.00	0.97	0.97	0.92	0.89
442	Odds	0.97	1.00	0.88	0.91	0.84
443	Evens	0.97	0.88	1.00	0.87	0.88
444	1stHalf	0.92	0.91	0.87	1.00	0.63
445	2ndHalf	0.89	0.84	0.88	0.63	1.00
446	*average*	*0.94*	*0.90*	*0.90*	*0.83*	*0.81*

Rows 3 through 426 of this report are not shown. They contain the test scores for each student, which we may not want to print. For the moment, all we want to print is comprised of what's seen in the snapshot above.

The following shows what to do. (How to get this Page Setup box to show is mentioned in the previous topic.)

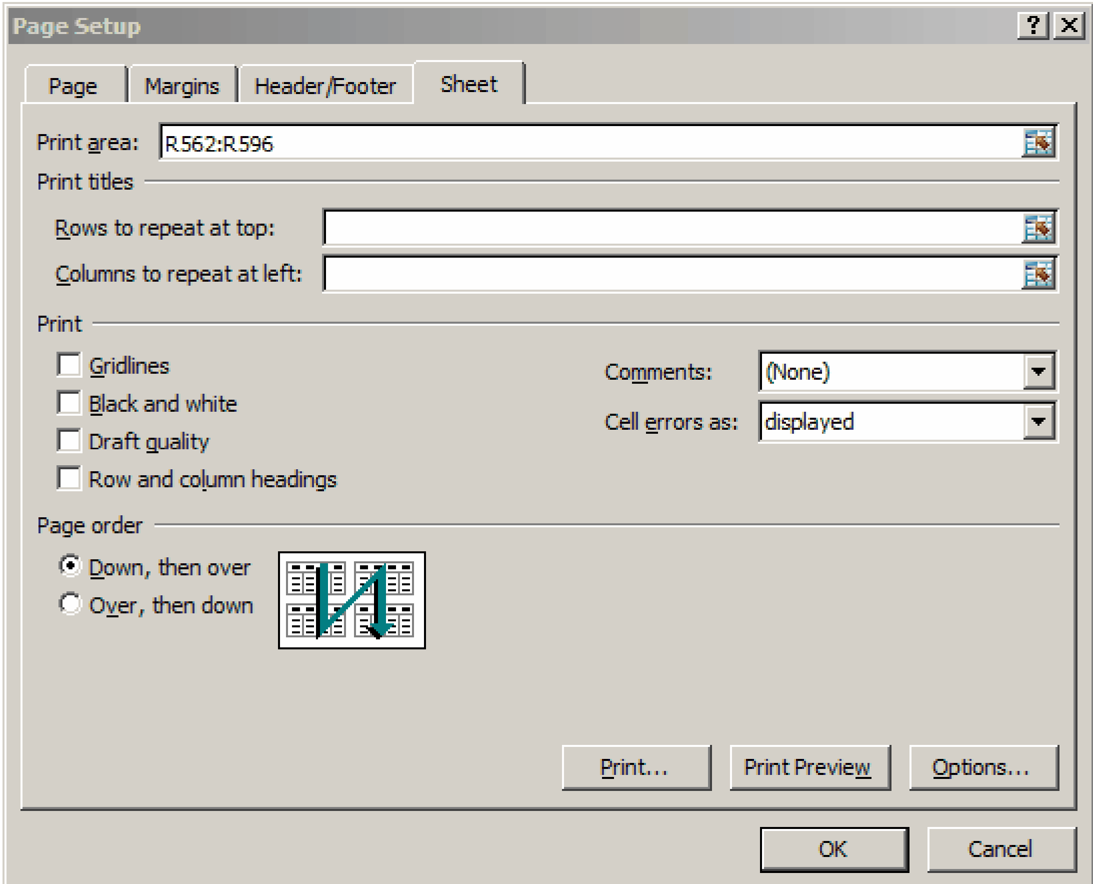


How about part of a Breaks1bw report?



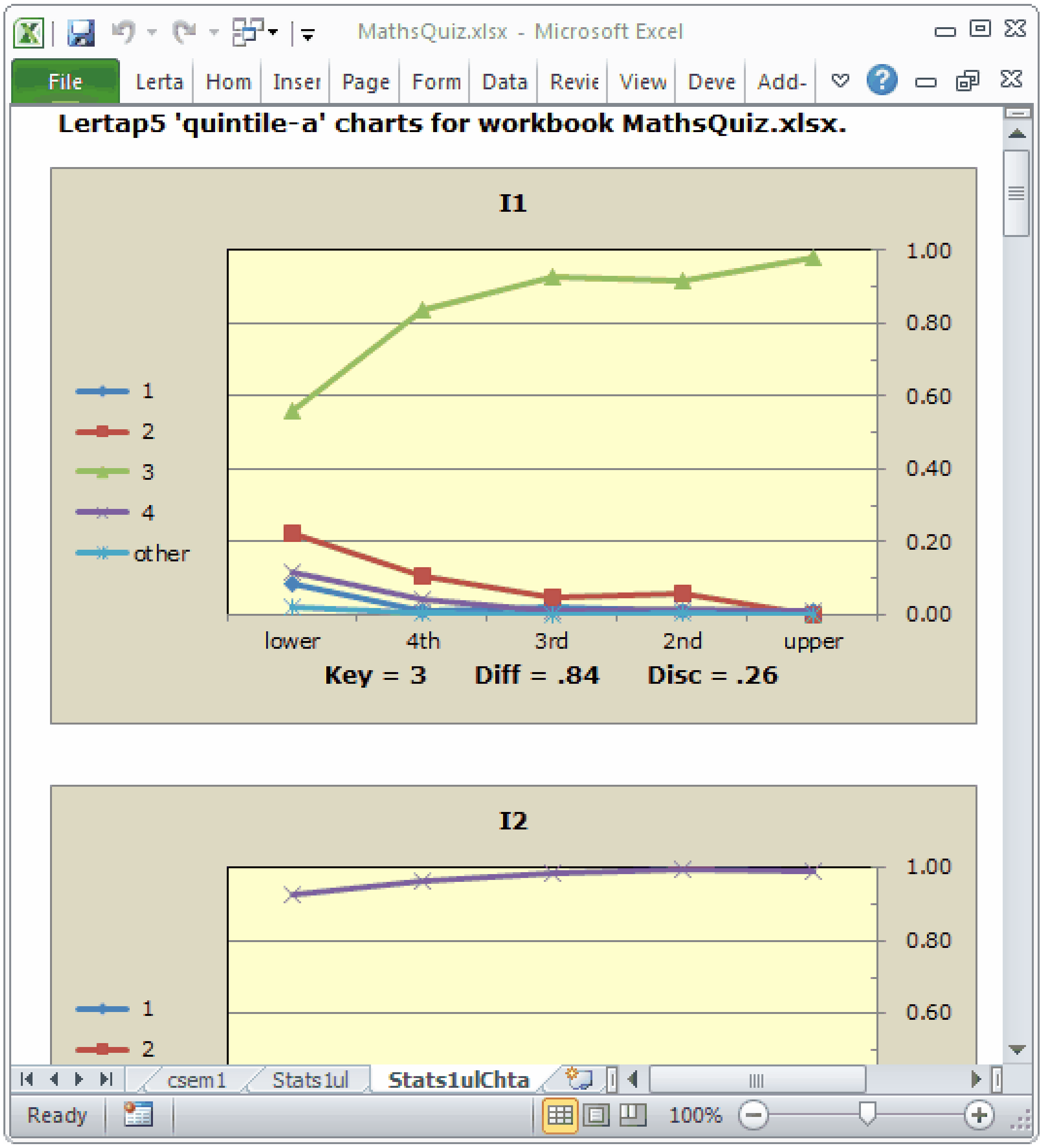
	1	2	3	4	5	6
562	MNur1	A1	A3	A4	H1	U1
563	n	412	181	559	310	307
564	Mean	42.46	44.13	41.22	36.01	39.58
565	s.d.	7.26	7.23	7.22	7.24	7.61
566	Median	43.00	45.00	41.00	36.00	39.00
567	Q1	38.00	39.00	36.00	30.00	35.00
568	Q3	48.00	50.00	47.00	41.00	45.00
569	Minimum	20.00	25.00	21.00	18.00	2.00
570	Maximum	59.00	60.00	59.00	55.00	59.00
571	25th Pct	38.00	39.00	36.00	30.00	35.00
572	50th Pct	5.00	6.00	5.00	6.00	4.00
573	75th Pct	5.00	5.00	6.00	5.00	6.00
574	Bottom whisk	14.00	14.00	15.00	12.00	14.00
575	Top whisker	11.00	10.00	12.00	14.00	14.00
576	Lowliers	3	0	0	0	3
577	Highliers	0	0	0	0	0

In this case there won't be any rows to repeat at the top, and the following settings will work:


Page Setup
Page
Margins
Header/Footer
Sheet
Print area:
R562:R596
Print titles
Rows to repeat at top:
Columns to repeat at left:
Print
Gridlines
Black and white
Draft quality
Row and column headings
Comments:
(None)
Cell errors as:
displayed
Page order
Down, then over
Over, then down
Print...
Print Preview
Options...
OK
Cancel

4.2 Print quintiles

An example of quintile plots (which are Excel charts) is shown below. Read more about such plots/charts here.



There will usually be one chart for each item. 200 items? 200 charts!

Inserted note, 10 January 2015: should you be lucky enough to have Excel 2010 or Excel 2013 on your computer, and have installed version 5.10.5 or better (see how to get it if you don't have it), then you're in "fat city" as far as printing goes: just marvel at this document (a pdf file) and see that Lertap 5 now has a print controller.

On the other hand, if your Lertap 5 is an older version, get a cuppa and read the following.

Printing these charts can require a bit of extra effort for a number of reasons. There are often very many of them. More often than not, they need to be resized so that they will fit on a printed page, and then some more fiddling has to be done to see that Excel does not print part of a chart on one page, with the rest of it at the top of next page.

These things can be done if you have a recent version of Lertap for Excel 2007, Excel 2010, or Excel 2011. (When this topic went to press, respective Lertap version numbers were 5.8.2.1 (Excel 2007), 5.9.2.1 (Excel 2011 for Macintosh), and 5.10.1 (Excel 2010). To see the version number of your Lertap, click on the small yellow smiley face found on the left-hand side of the Lertap toolbar.)

There are two chart-changing macros which can be used to fix up the quintile plots so that they're easier to print. They're called **ChartChanger2** and **ChartChanger3**.

The ChartChanger2 special macro allows you to delete a quintile or two, and then reformat the display so as to recover the space used by the deleted quintiles. ChartChanger2 is available in all versions of Lertap.

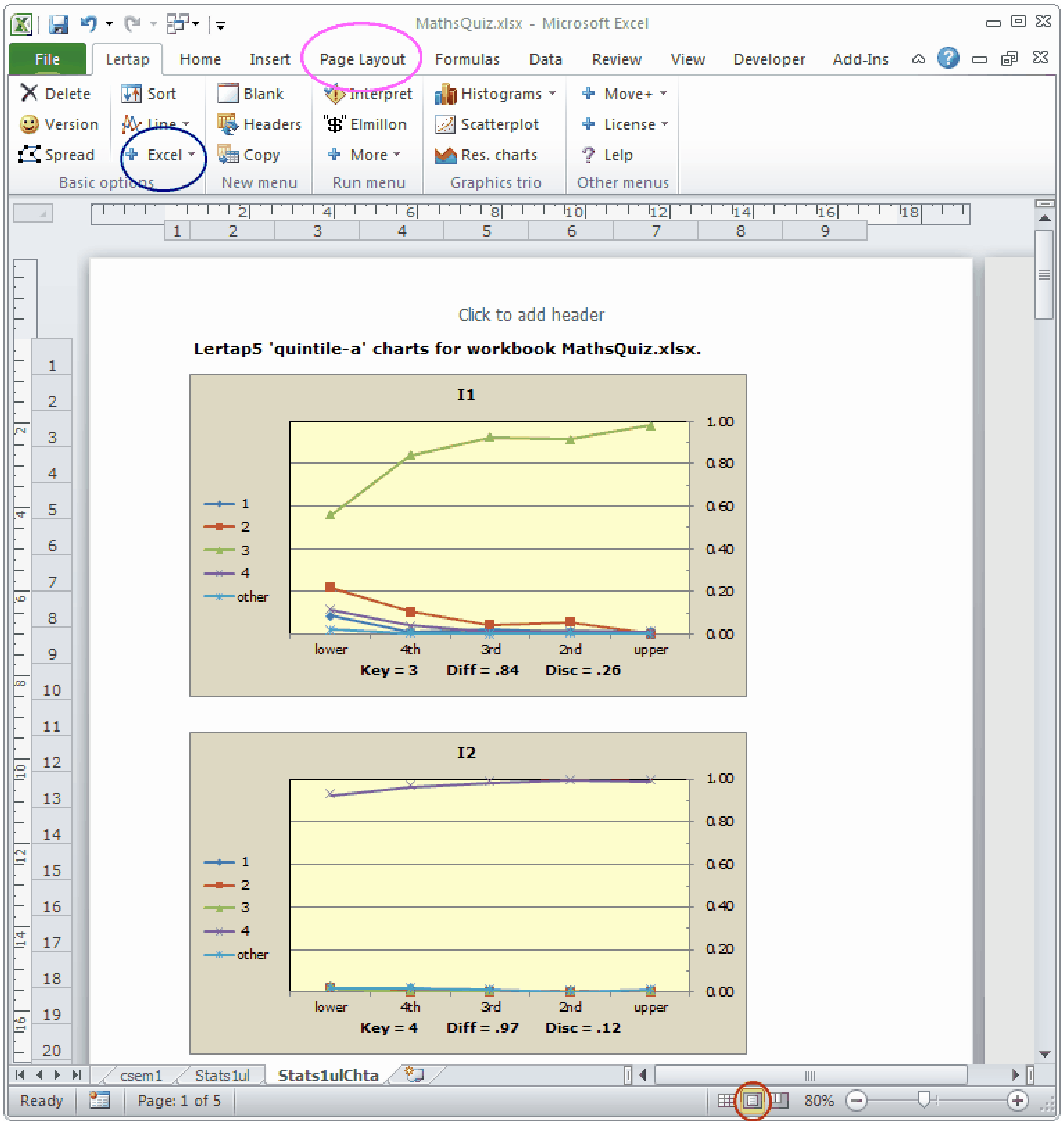
We'll use quintiles from the MathsQuiz dataset as an example of using ChartChanger2.

To begin, imagine that you're playing along with this discussion. Also, imagine that you've got a recent version of Lertap for Excel 2007 or Excel 2010 (version 5.8.2.1 or above).

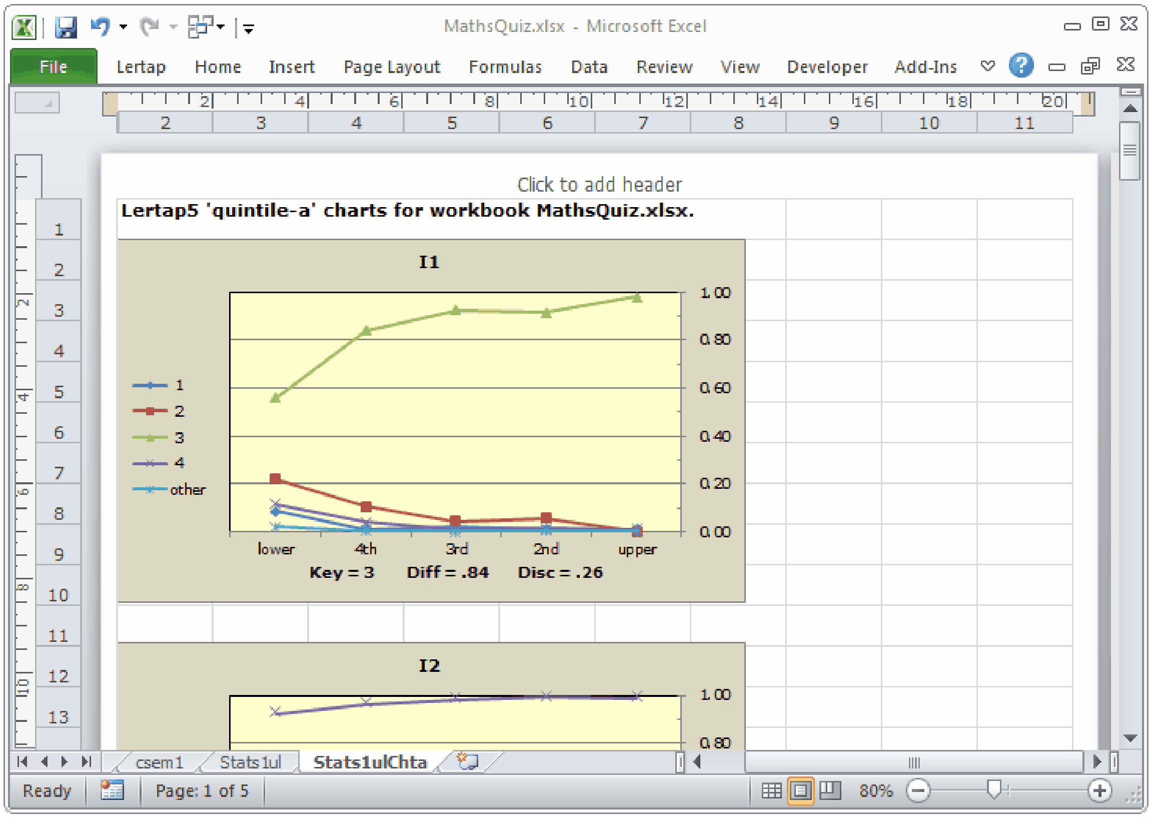
You download the MathsQuiz dataset, and generate the quintiles there at your own desk, or on your laptop as you float on an air mattress and navigate around the pool.

You get the quintiles for the 15 test items. They're interesting. They're colorful. Many are fascinating. But they're too big. You'd like a more concise display.

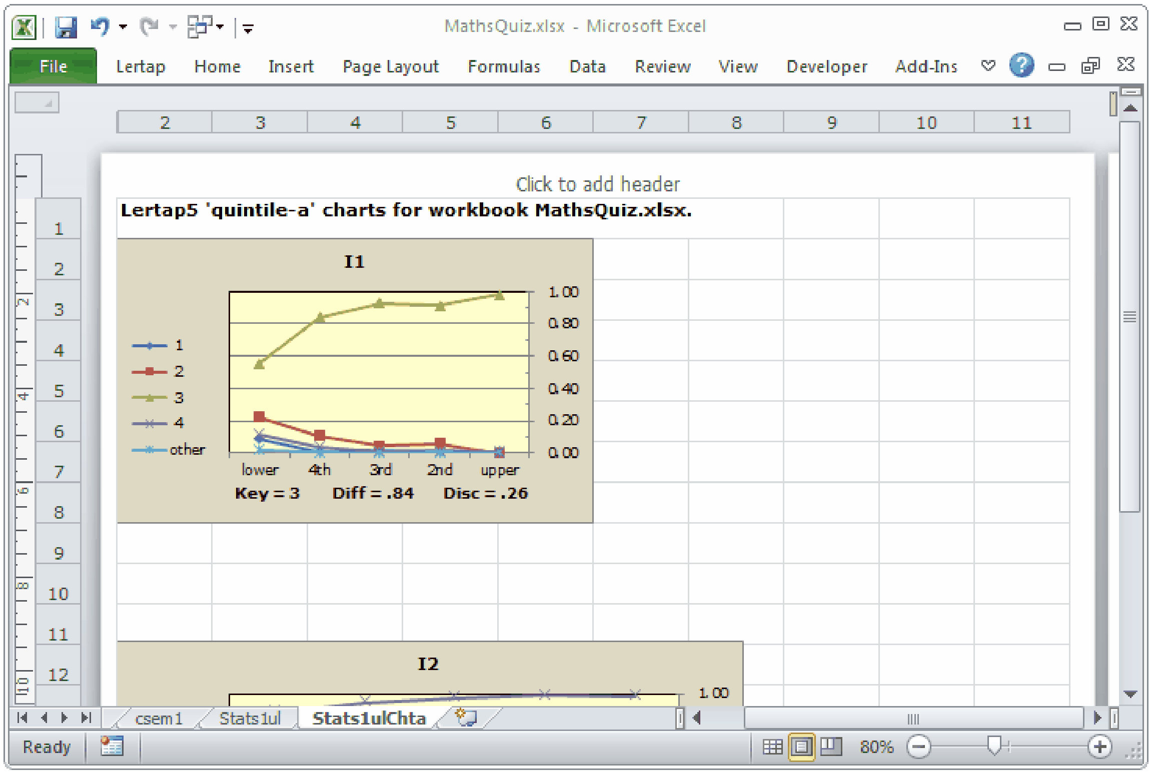
You do these things: switch to Page Layout view (small red circle, bottom of screen snapshot below). Turn on gridlines (larger blue circle around the word Excel: click on it and select Gridlines). You use some of the many options under Excel's Page Layout menu to change page margins (purple circle). Lertap always leaves an empty first column on its quintile pages; you reduce the width of this column to zero.



If your printer is set to use A4 paper, you might then see something like the next screen snapshot. (If your printer has "Letter" paper, you'll have a wider but slightly shorter page to work with.)



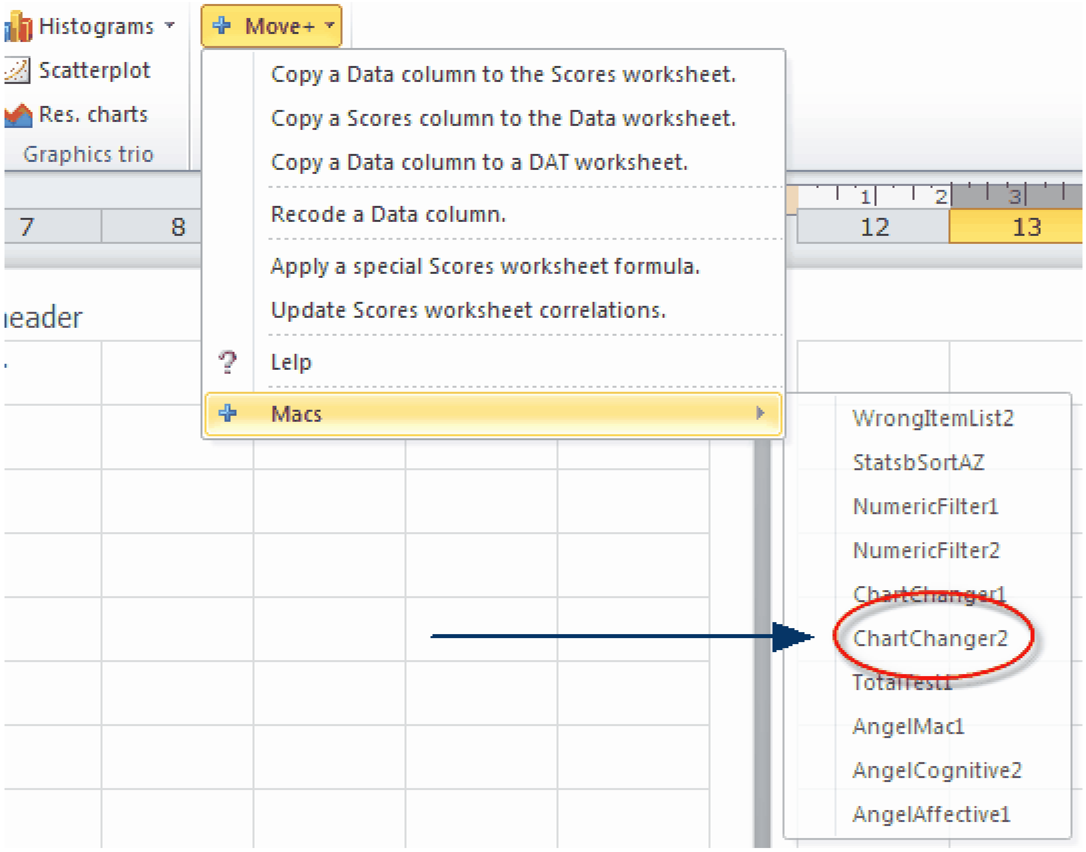
If we continue to work with portrait page orientation, we have 10 columns in the worksheet to play with. We'll grab the first quintile at its lower-right corner, and gently, lovingly, shrink it:



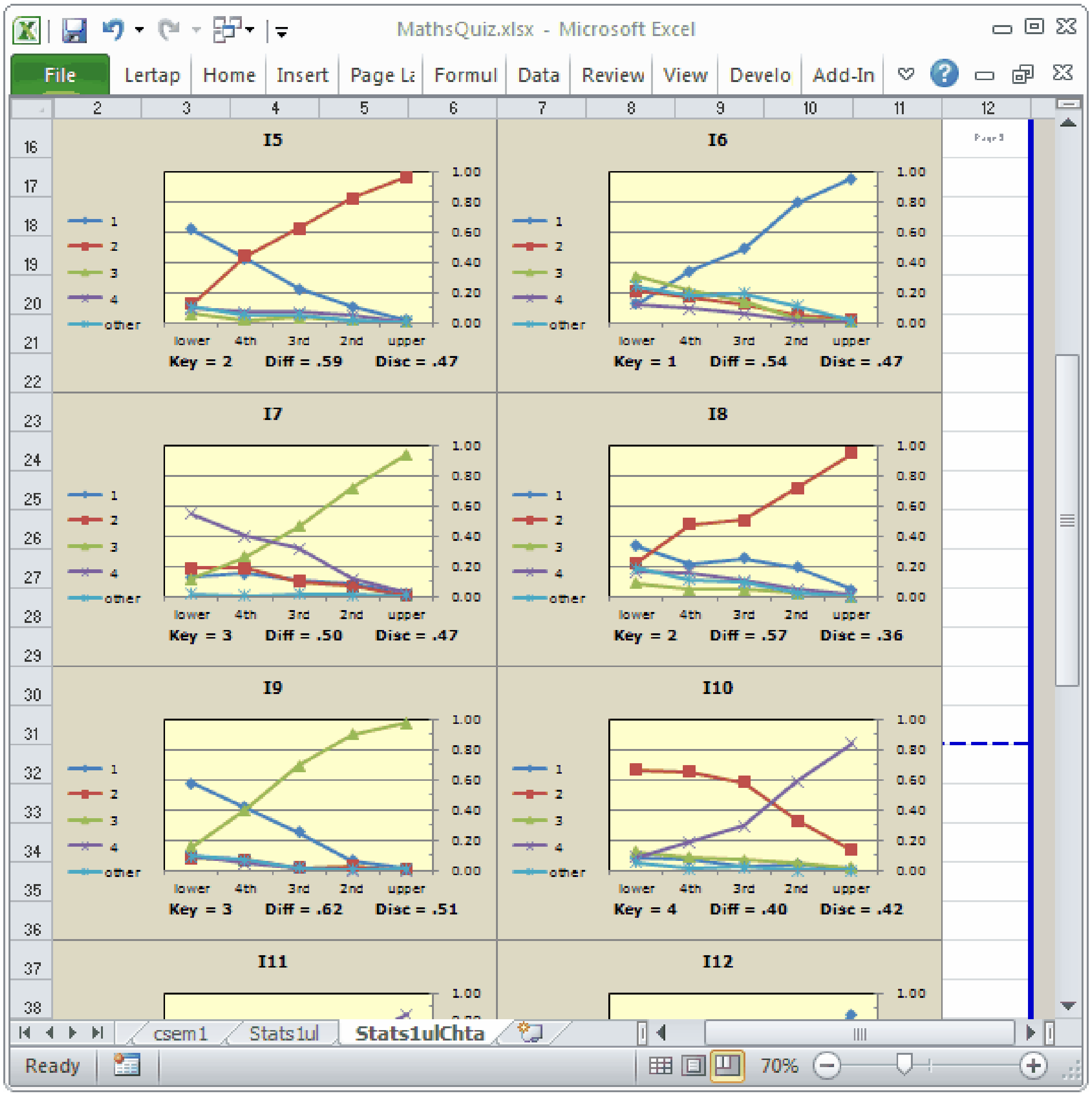
Our shrink has *very carefully* positioned the first chart so that its edges **align with gridlines** (is your shrink as good as ours?).

Note for Excel 2010 users: if your Lertap is up to date (version 5.10.x or better), you will have access to **ChartChanger3** as well as ChartChanger2. One of the main features of ChartChanger3 is its ability to re-size plots and maintain grid alignment at the same time. Read about it here (also see the Tidbits at the bottom of this topic).

Okay, **ChartChanger2**, where are you?



We ask ChartChanger2 to position two charts per row of the display, and opt to have no blank lines between the rows.



(Note that we have now switched on the Page Break Preview option; it's at the bottom of the screen snapshot above, to the left of where it says 70%.)

Okay, what have we? The first eight quintile charts, corresponding to items I1 through I8, will, according to the screen snapshot above, fit on the first printed page, as will the upper portions of the charts for I9 and I10.

We've got a bad page break. Worse than bad breath (to some). We don't want those "upper portions"; we want whole charts, not parts. Can we get all of charts I9 and I10 to appear on the next page?

Certainly. Easy peasy.

We position the mouse on the blue dashes to the right of I10's chart, and drag the dashed line up, ever so slowly, ever so carefully, stopping when it's level with the top of the I10 chart.

Ready, set, print. Turned out well. How you'd go at your place?

Tidbits:

Access to the **ChartChanger3** macro may be available to all users, even those without Excel 2010. A version of it is likely to be found in the **Lertap5MacroSetA.xlam** file. Write to larry@lertap.com for help.

5 Preparing Data

The samples and examples found in this website have, of course, all been prepared in advance. Almost all of the examples have real datasets ready for you to download. When they come down the internet "wire" to your computer, they do so (or are supposed to do so) as Excel workbooks, set up as **Lertap workbooks**. (Click here for a definition of what makes an Excel workbook a "Lertap workbook".)

Okay, now let's anticipate the day when you'll want to make your own Lertap workbook. How will you do it?

First of all, let's take a straightforward, common, example: you've got a pile of test or survey results sitting next to you on your desktop. They might be the actual questions, with marks where students or participants have indicated their selected answers (such as circles around item options). How to enter the results into Excel so that they meet the requirements of a Lertap workbook?

Not too hard to do at all. You'd go to the New menu, and there you'd probably select the "Blank" option. This action will set up a new Excel workbook and it'll have the two worksheets, Data and CCs, required to make it a "Lertap workbook". Easy-peasy, right? Then all you need do is type answers into the Data sheet, and put at least two lines in the CCs worksheet: a *col line, followed by a *key line if you have a test, or a *sub line if you have a survey.

No need to get overly anxious about these tasks -- we've got two pre-prepared samples on this website with **answer sheets** having actual item responses. These may be used for practice; you print the answer sheets, create a new Lertap workbook, and then enter the responses, by hand (fingers), into the Data worksheet (great fun, much more rewarding than vacuuming the carpets, if that's what you were supposed to be doing). One of these samples is ChemQuiz, the other is CEQ. These exercises are typical of those completed by instructors who want to process a class test or survey.

After you've done this a couple of times, you may come to regard one of the other options on the New menu as a shortcut. This could happen, for example, if you repeatedly use the same test or survey. In this case, the New menu's "Headers" option, or the "Copy" option, could save you a wee bit of time.

Someone asks: "*Must I use Lertap's New menu to set up a new workbook? Can't I use Excel?*"

No. Yes.

No, users don't have to use Lertap's New menu. Yes, Excel's File tab, and then its New option, will certainly create a new workbook. Its worksheets will have names such as "Sheet1", "Sheet2", and so on. You must then change the names of the worksheets; one of them should be named Data, the other CCs. If Excel gives you "Sheet3" as well, you can delete it, although this is not required. The advantage of using the "Blank" option on Lertap's New menu is that it does these things for you, saving time.

What about this situation: somehow you obtain an Excel workbook already set up with rows containing student data, such as their IDs, maybe a class code of some sort, and then columns with item responses. Great. Make sure the name of the worksheet with the data is called "Data". Make sure this Data worksheet has some sort of title its first row, and that its second row has column headers, such as, for example, ID, Class, I1, I2, and so on. Then add a CCs worksheet, and pop at least two lines into it. (Look here for an example of this process.)

Now, we've saved a very common situation for last: you've got a data file created by a **scanner**.

Scanners prepare their data files in a variety of ways. There are some which will create a "**csv**" file (comma-separated values), and these files can be opened straightaway in Excel without much ado at all. Should this happen to you, open the csv file with Excel, and then save it as an Excel workbook, using the options on the File tab. Then go about making sure that the name of the worksheet with data is named "Data", format the Data worksheet so that its first two rows are as expected by Lertap, add a CCs worksheet with at least two lines, and *Bob's your uncle!* (We've got a beaut example of this very process. Clickity click click here.)

Many scanners create data files which are text files, often called **ASCII** files. These files will usually have an extension of "**dat**" or "**txt**".

Excel has a "**Text Import Wizard**" which greatly eases the task of preparing text files so that Excel can work with them. It isn't too hard to use; the next topic shows how (*Ted's your aunty!*).

How'd you like to have a sample dat file to work with? Okay, can do, the next paragraph has an option to download such a file, which was originally from the University of Iowa. It's called "mondaty.dat".

How'd you like to have some comments on how to prepare this dat file for Lertap? Hmmm, once again, you're in luck, we've got a special document for you (a web page, in this case): click here, and NOTE: the download link in this document looks for the file at www.uiowa.edu. This link no longer works. Download the file from here instead.

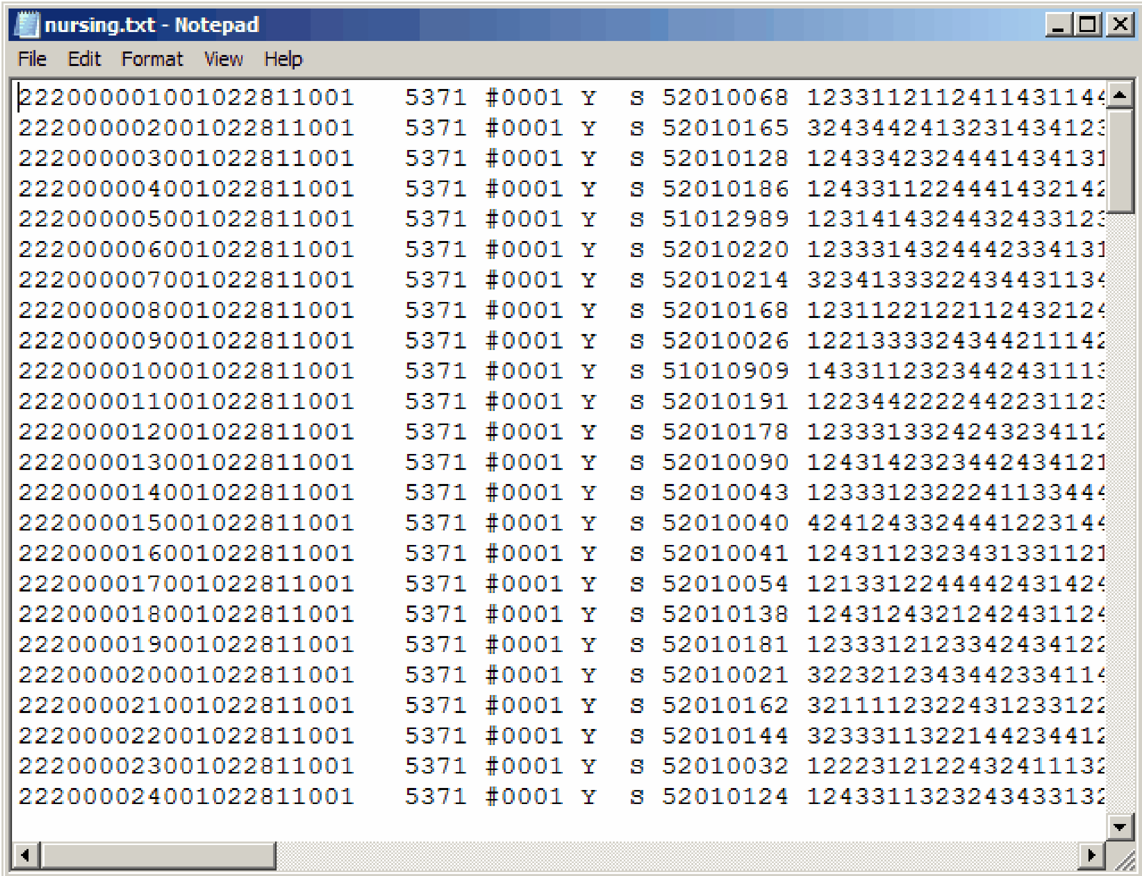
You'll note that the document does not use the Text Import Wizard. What you could do for a bit of excitement is to use the Text Import Wizard to accomplish what's done by copy and paste in the document. The copy-and-paste approach works in this document as the mondaty.dat file is very simple, with no ID fields of any sort before the actual item responses; this is not at all characteristic of what scanner output files usually look like. See why: page ahead.

Tidbit:

If *Bob* is not your uncle, and *Ted* is not your aunty, just pretend that they are when you come to Australia (where these terms are common).

5.1 Scanner data import (1)

The momentous matters mentioned in this topic have to do with importing an "**ASCII**" file. Such files are often created by scanners. To learn more about such devices, scanners, try an internet search for "mark-sense reader", or "optical mark reader".



```
222000001001022811001   5371 #0001 Y  S 52010068 1233112112411431144
222000002001022811001   5371 #0001 Y  S 52010165 3243442413231434123
222000003001022811001   5371 #0001 Y  S 52010128 1243342324441434131
222000004001022811001   5371 #0001 Y  S 52010186 1243311224441432142
222000005001022811001   5371 #0001 Y  S 51012989 1231414324432433123
222000006001022811001   5371 #0001 Y  S 52010220 1233314324442334131
222000007001022811001   5371 #0001 Y  S 52010214 3234133322434431134
222000008001022811001   5371 #0001 Y  S 52010168 1231122122112432124
222000009001022811001   5371 #0001 Y  S 52010026 1221333324344211142
222000010001022811001   5371 #0001 Y  S 51010909 1433112323442431113
222000011001022811001   5371 #0001 Y  S 52010191 1223442222442231123
222000012001022811001   5371 #0001 Y  S 52010178 1233313324243234112
222000013001022811001   5371 #0001 Y  S 52010090 1243142323442434121
222000014001022811001   5371 #0001 Y  S 52010043 1233312322241133444
222000015001022811001   5371 #0001 Y  S 52010040 4241243324441223144
222000016001022811001   5371 #0001 Y  S 52010041 1243112323431331121
222000017001022811001   5371 #0001 Y  S 52010054 1213312244442431424
222000018001022811001   5371 #0001 Y  S 52010138 1243124321242431124
222000019001022811001   5371 #0001 Y  S 52010181 1233312123342434122
222000020001022811001   5371 #0001 Y  S 52010021 3223212343442334114
222000021001022811001   5371 #0001 Y  S 52010162 3211112322431233122
222000022001022811001   5371 #0001 Y  S 52010144 3233311322144234412
222000023001022811001   5371 #0001 Y  S 52010032 1222312122432411132
222000024001022811001   5371 #0001 Y  S 52010124 1243311323243433132
```

The snapshot above shows the Windows **Notepad** program displaying the partial contents of a data file created by a scanner.

Each line of data represents information for one student.

Each line has a number of distinct "fields", or groups, of information. The first of these has 21 characters, starting with 222. After the 222 comes a 6-character sequential code, which may be read as a number (000001 to 000024 in the lines above), followed by something which looks like it could be a date code, ending with 0001.

The second field has just four characters, 5371.

In this example, we will be interested in the contents of the fifth field (a single character, "S"), the sixth field (eight characters, 52010068 going through 52010124, as seen above), and the seventh field.

The seventh field has item responses; 88 of them. (Obviously, most of these are not showing above.)

What we'd like to do is import the contents of these three columns so that we can use the information they contain with Excel.

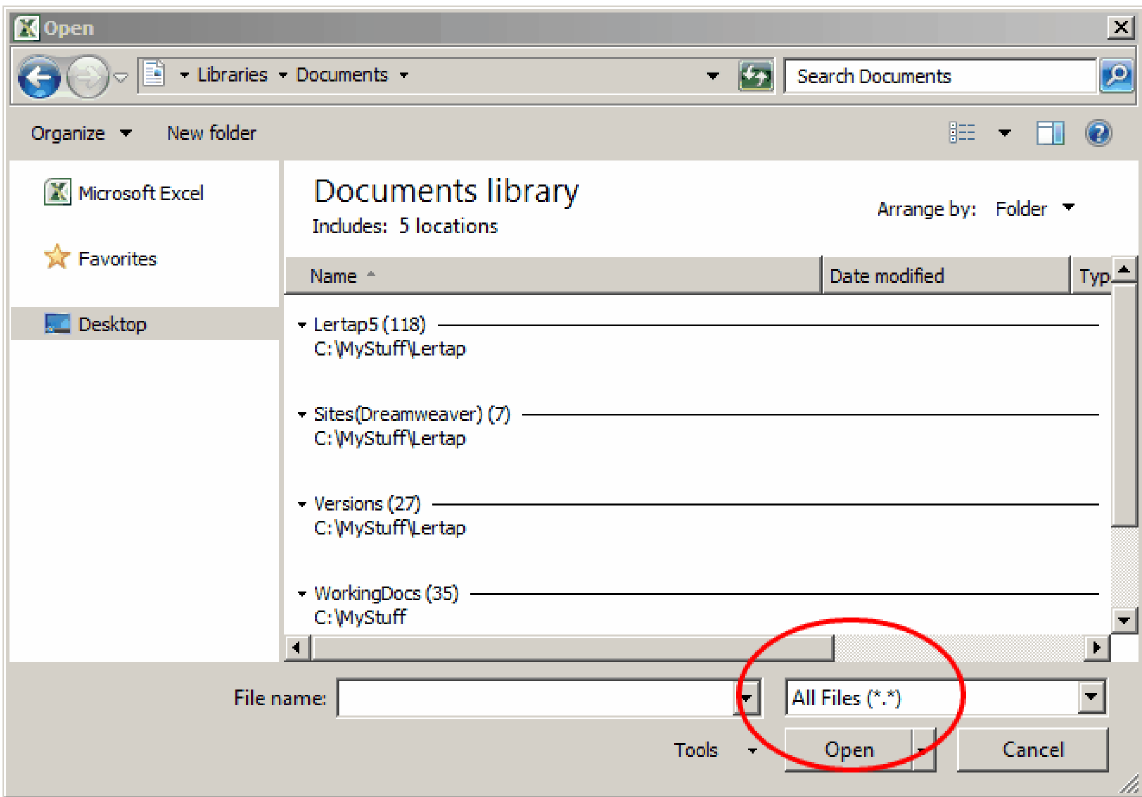
It's pretty easy to do.

Are you ready?

Make sure you've made a note of where the data file is on your computer. In our case, the file is named "nursing.txt", and we can't tell you the name of the folder where it's found as that would give away the source of the data, which is supposed to be confidential.

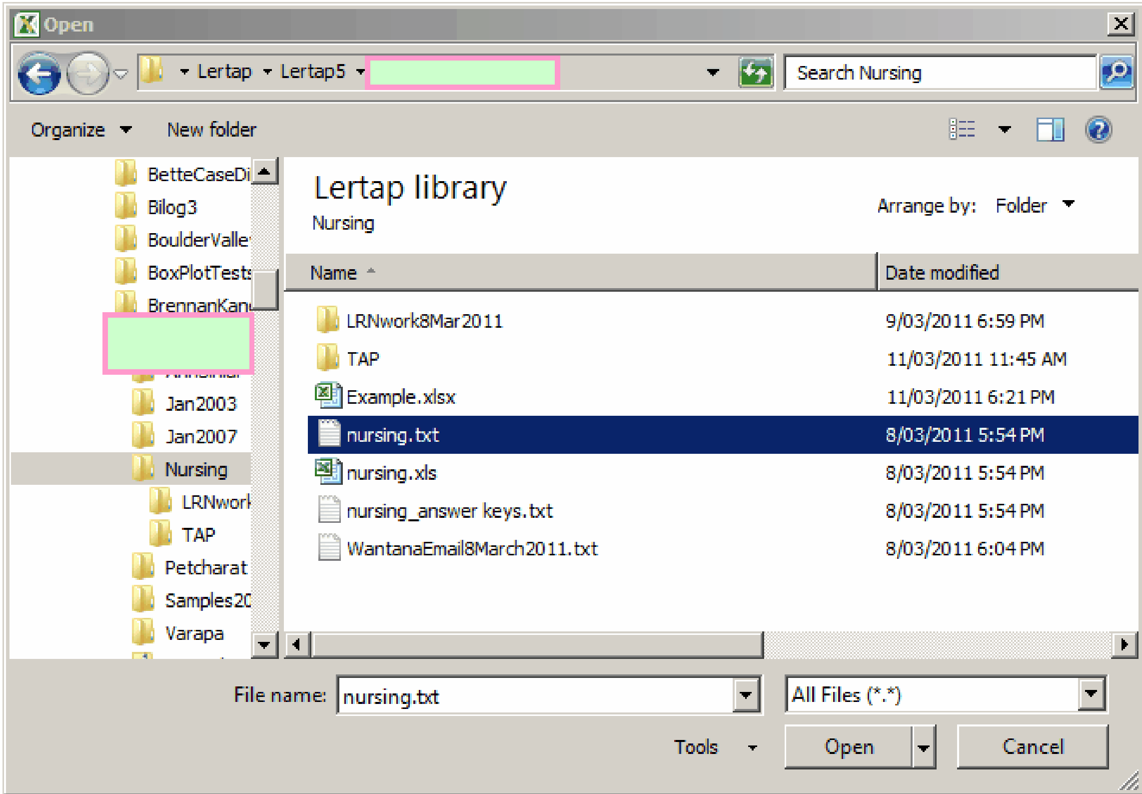
Start Excel.

File / Open



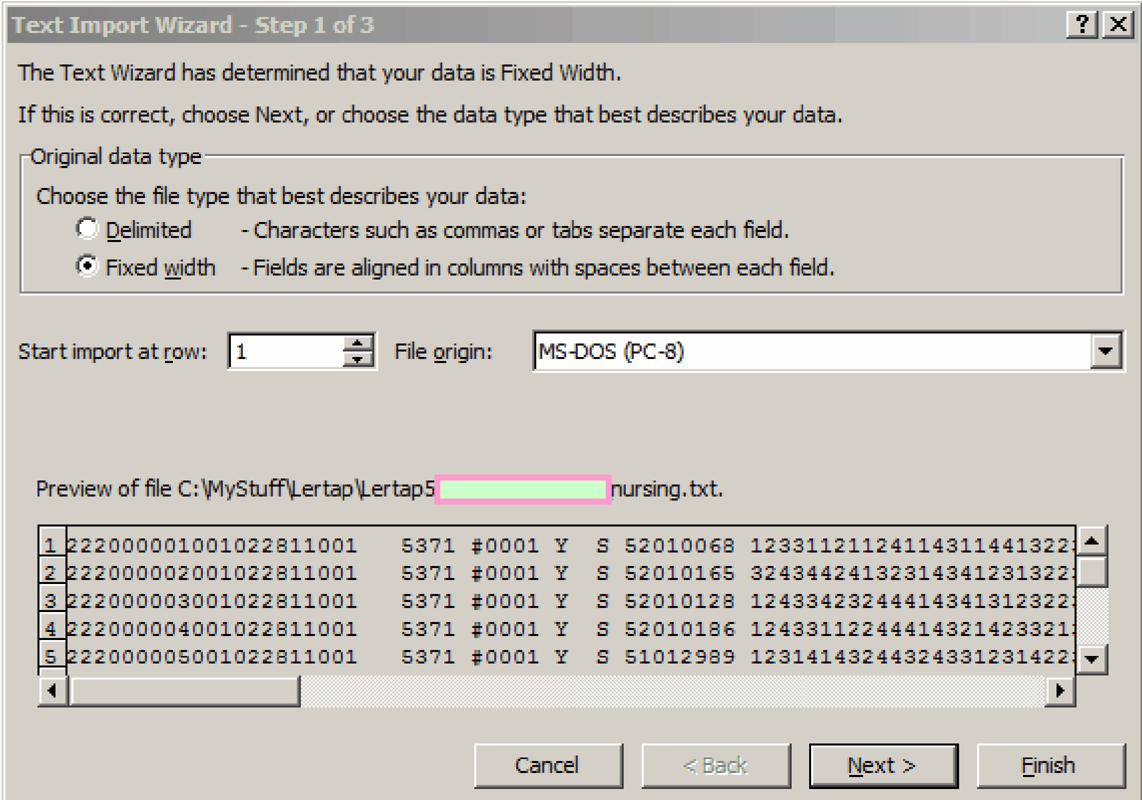
We're going to be looking for a txt file (many scanners will call their main output file a "dat" file, but today we've got a scanner tossing out "txt" files).

We tell Excel we're interested in "All Files (*.)" so that it doesn't limit our file search to just files of the type it usually favors (such as xlsx, xlsm, xlam, xls, and so on).

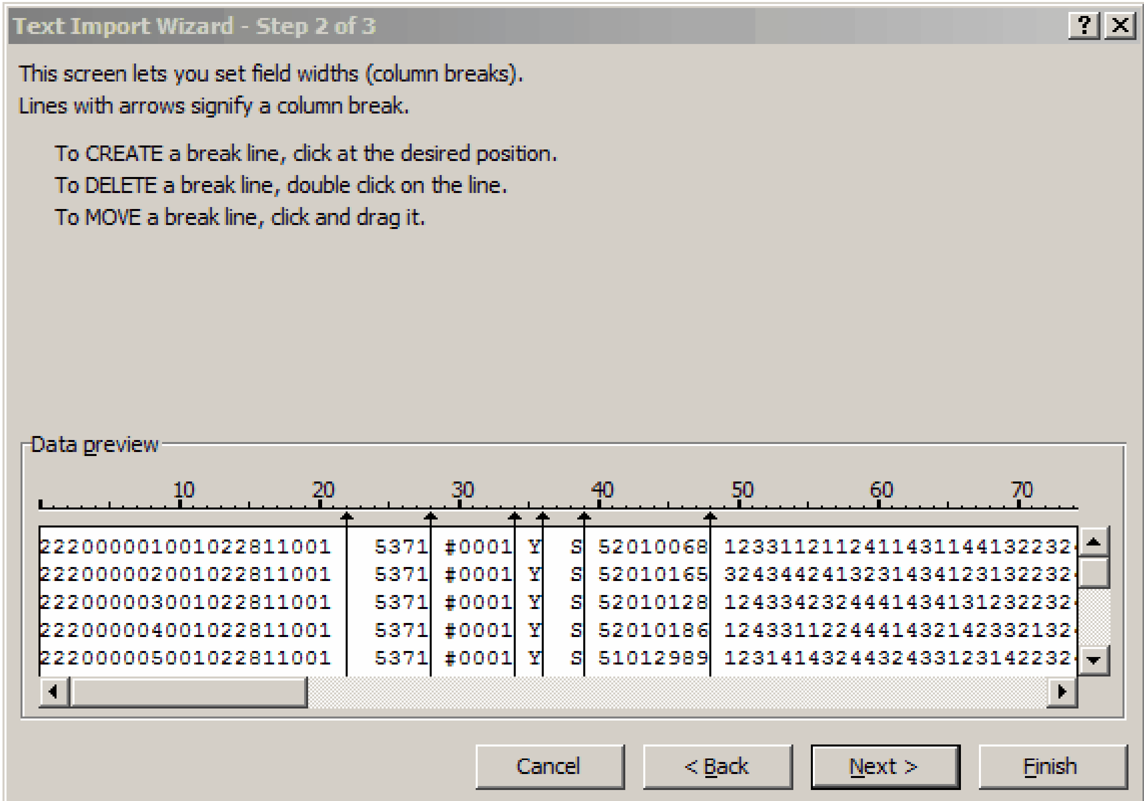


(The green boxes seen above are an attempt to hide the source of this example.)

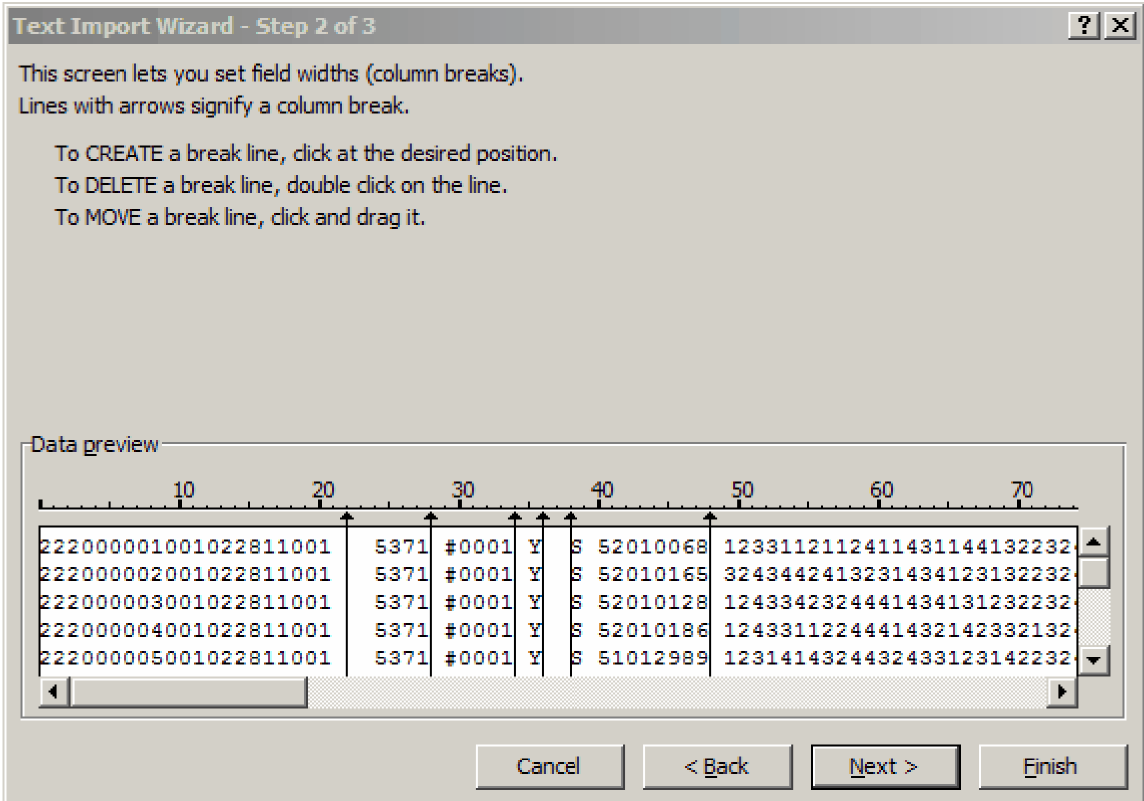
Open (The **Text Import Wizard** swings into action.)


Text Import Wizard - Step 1 of 3
The Text Wizard has determined that your data is Fixed Width.
If this is correct, choose Next, or choose the data type that best describes your data.
Original data type
Choose the file type that best describes your data:
Delimited - Characters such as commas or tabs separate each field.
Fixed width - Fields are aligned in columns with spaces between each field.
Start import at row: 1
File origin: MS-DOS (PC-8)
Preview of file C:\MyStuff\Lertap\Lertap5 nursing.txt.
Cancel
< Back
Next >
Finish

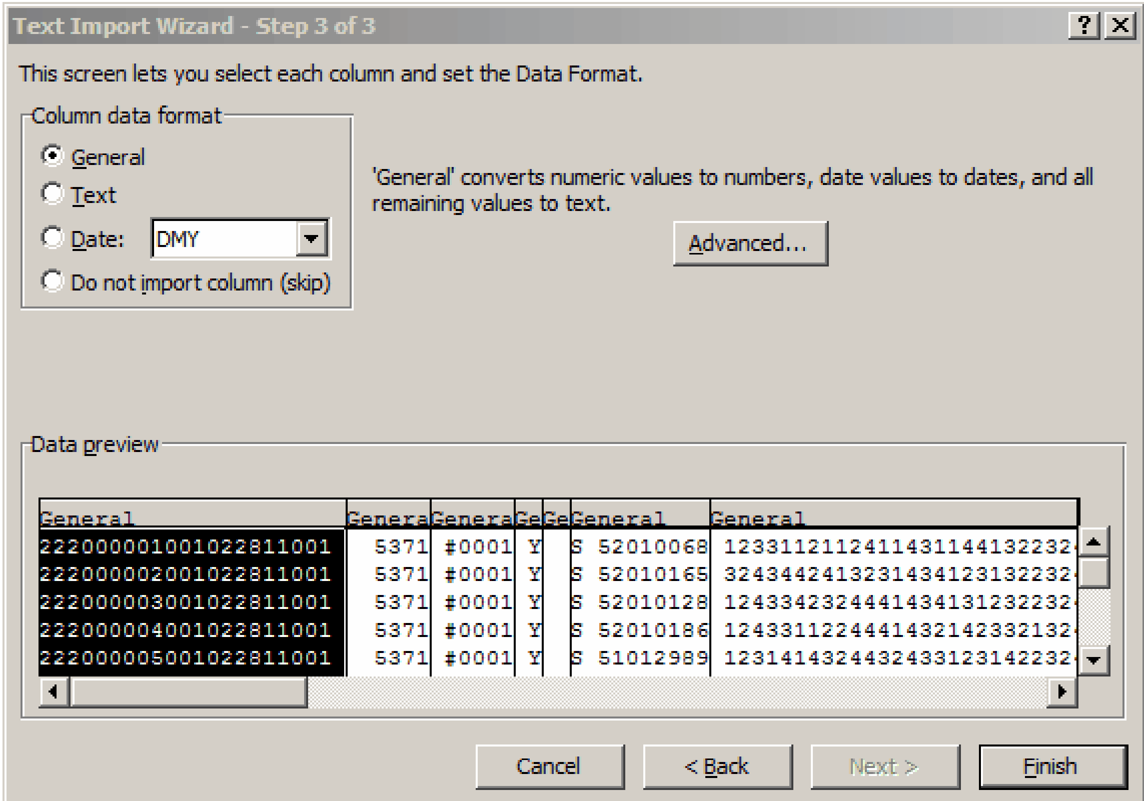
Next >



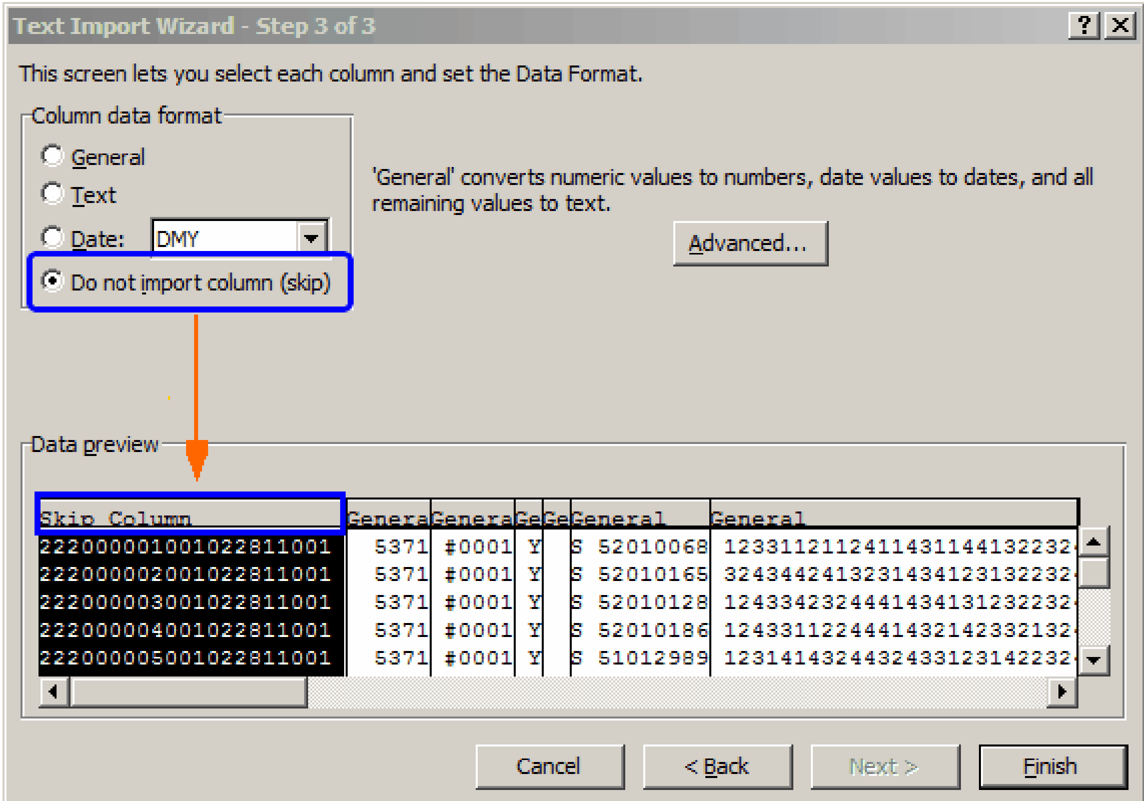
We wanted to join the field with the "S" to the next field to the right, so we used the mouse to move one of the arrows to the left:


Text Import Wizard - Step 2 of 3
This screen lets you set field widths (column breaks).
Lines with arrows signify a column break.
To CREATE a break line, click at the desired position.
To DELETE a break line, double click on the line.
To MOVE a break line, click and drag it.
Data preview
2220000001001022811001 5371 #0001 Y S 52010068 1233112112411431144132232
2220000002001022811001 5371 #0001 Y S 52010165 3243442413231434123132232
2220000003001022811001 5371 #0001 Y S 52010128 1243342324441434131232232
2220000004001022811001 5371 #0001 Y S 52010186 1243311224441432142332132
2220000005001022811001 5371 #0001 Y S 51012989 1231414324432433123142232
Cancel
< Back
Next >
Finish

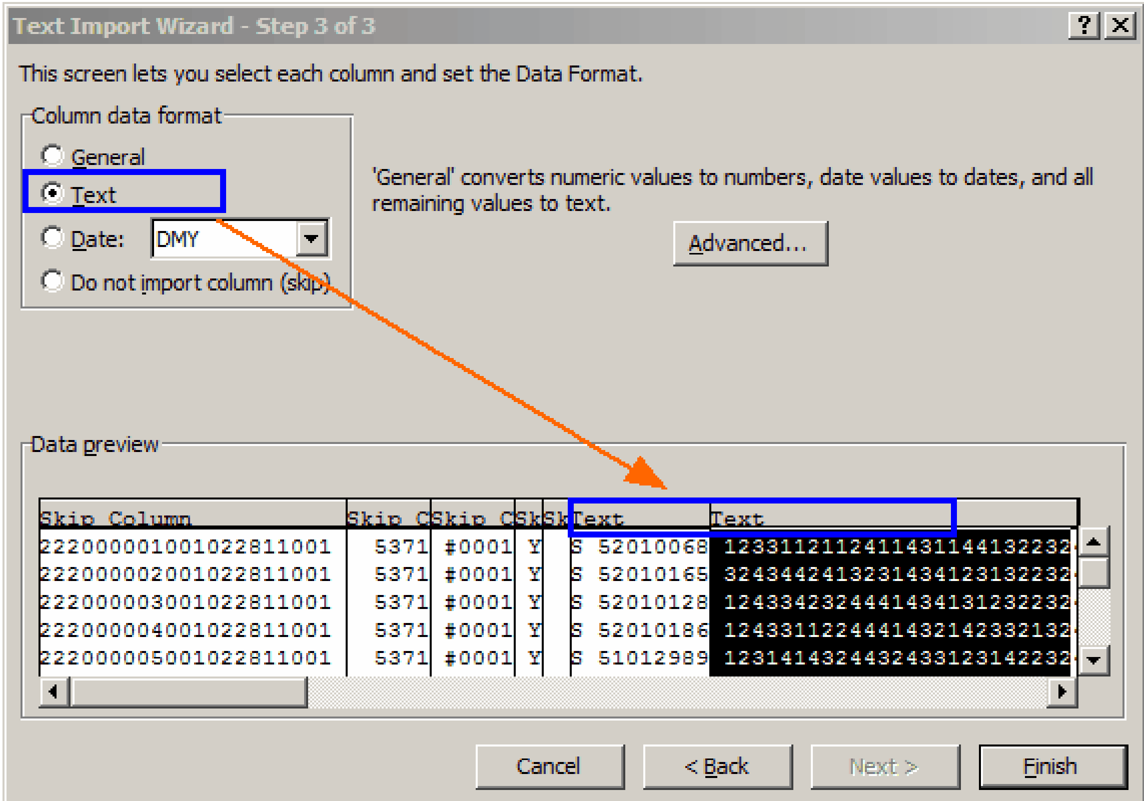
Next >



Now the Text Import Wizard has blocked off our fields, referring to each as a "column", highlighting the first one. We don't want this column, so we click on "Do not import (skip)".

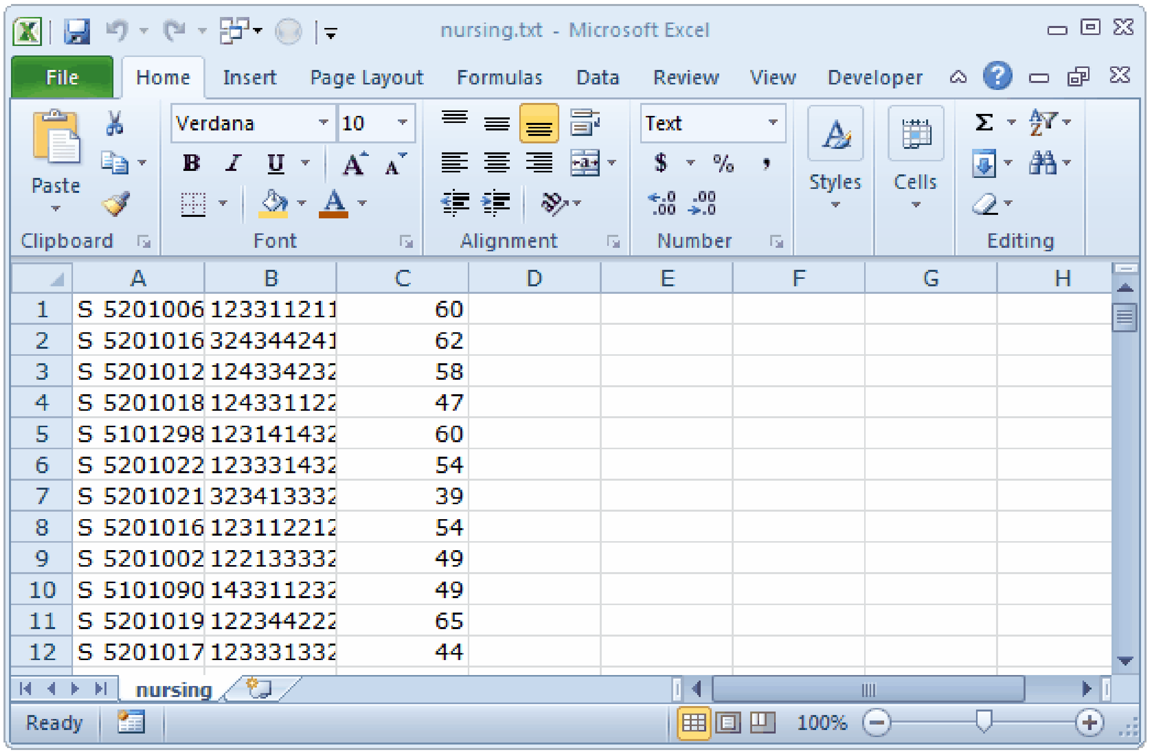


We skip the next four columns (5371, #0001, Y, and the empty one), and then tell the Wizard that the fifth and sixth columns are to come in as "Text".



Finish

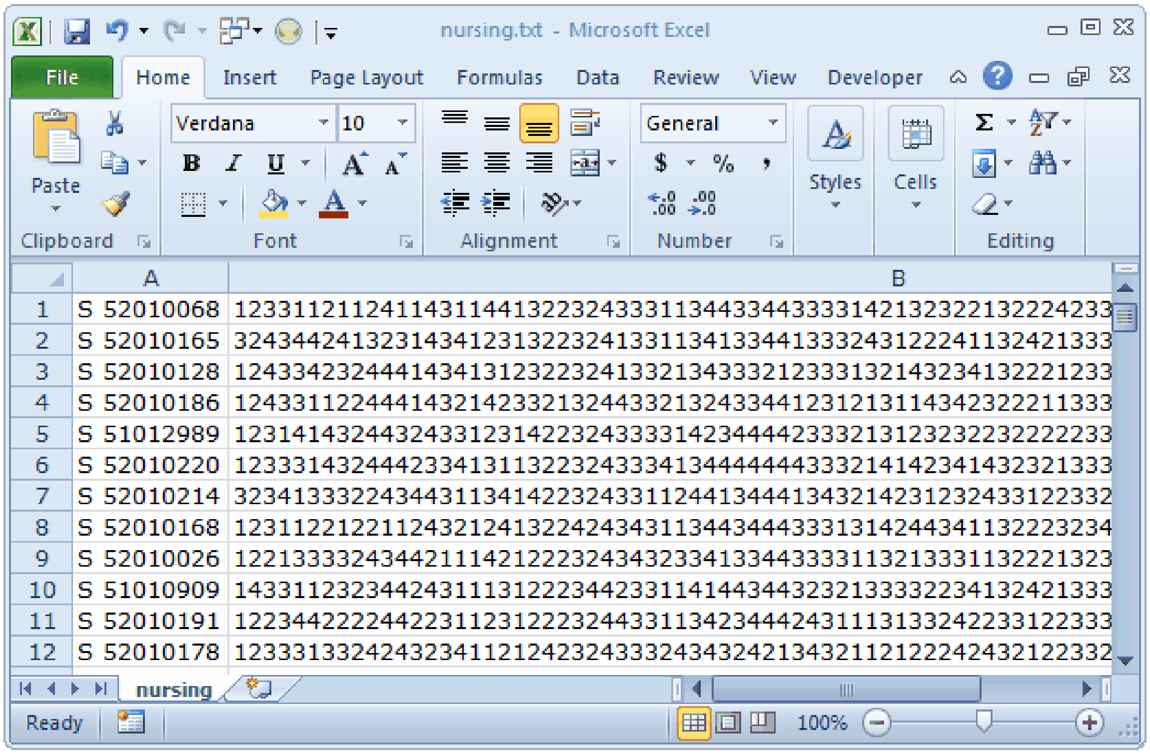
This will end the Text Import Wizard, and drop us back into Excel. Here's what we see now:



Uh-oh! There was an eighth field in the nursing.txt file which we had not seen. It's showing up here, in our new Excel worksheet, as column C. (In other words, we goofed.)

Is this a problem? Yes, in this case it is. Shortly we're going to be using "The Spreader" to dissect the long string of item responses so that each response gets tucked into its own column, and we know that The Spreader won't do it if any of the columns to the right are occupied with information. But not to worry!

We delete column C, and expand columns A and B so that we can see more contents:



Time to party, eh?

Maybe. Nothing wrong with a nice party from time to time, but actually all we've done thus far is extract two fields of data from the nursing.txt file, bringing them into Excel by using the Text Import Wizard. Although we've gone through a fair whack of screens to get to this point, it truly takes little time to do these things once you've been through them a couple of times.

So it is that mature readers will realize that we may not merit a party yet. We've done something which the 8-year old kids raucously playing kickball next door could do without more than a 10-second pause in their game. And, anyway, what's a party without Lertap?

We have more work to do; we're not yet ready for Lertap.

Page ahead to see what we did to earn that party entry.

Tidbits:

You ASCII for something? American Standard Code for Information Interchange. (Is that what you wanted to know?)

We goofed when working on this topic, as mentioned above. However, it is certainly possible to import and use fields which may come after the field with the string of item responses. The field we forgot about above, the one which rode into our Excel workbook as column C, surprising us in the process, was actually a useful one. It contained the test score for each student, as computed by the scanner's software.

We might have done well to intentionally import this field as well. In the Text Import Wizard, we'd leave it as a "General" column (not text) -- since this is the default action, the score popped into column C above as a number, which is good. But we'd want it to be left of the column with the string of item responses. Can do? Sure; easy peasy. Insert a column before column B. The test score now moves right, becoming column D. Select it. Cut it. Paste it into column B.

(If we don't clear the columns to the right of the one with the string of item responses, The Spreader will ring alarm bells, and decline to function.)

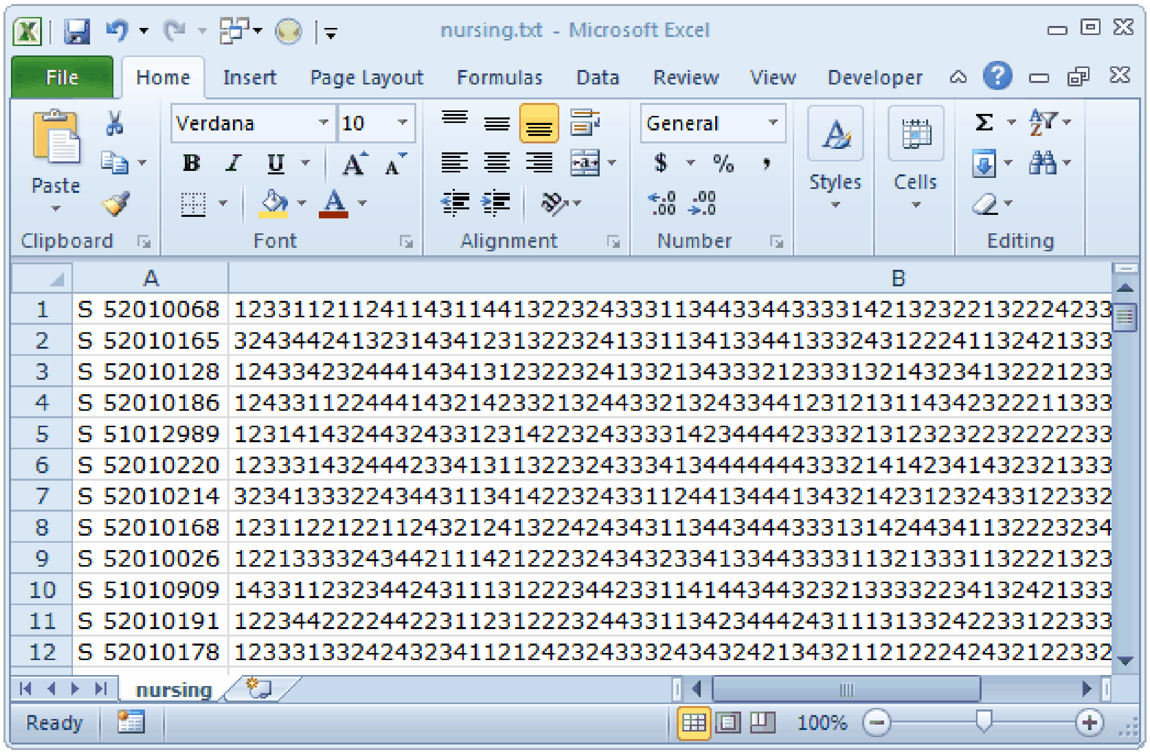
We can then get Lertap to correlate the scanner's score with the one created by Lertap. How? Easy. Dead easy. Use an option on the Move menu, "Copy a Data column to the Scores worksheet". Easy easy easy. Hopefully the correlation will turn out to be 1.00. (The two scores should be identical; if they're not, get a new scanner -- well, maybe check the *key line in the CCs worksheet first to make sure you didn't perchance make a wee goof of your own -- endeavor to keep the scanner operators on your side; don't suggest they need a new scanner unless your department has the funds to buy it.)

5.1.1 Over to Lertap

Alright, what we have, on entry to this topic, is an Excel workbook with just one worksheet. The worksheet's name is "nursing".

The worksheet is making use of just two columns, A and B (if we were using the R1C1 reference style, the columns would be 1 and 2 instead of A and B).

The first column has a student ID. The second has a string of 88 characters. Each character in this string represents an item response. In the screen snapshot seen below, student S 52010068 chose option 1 on the first item, 2 on the second item, and 3 on items three and four (and so on).



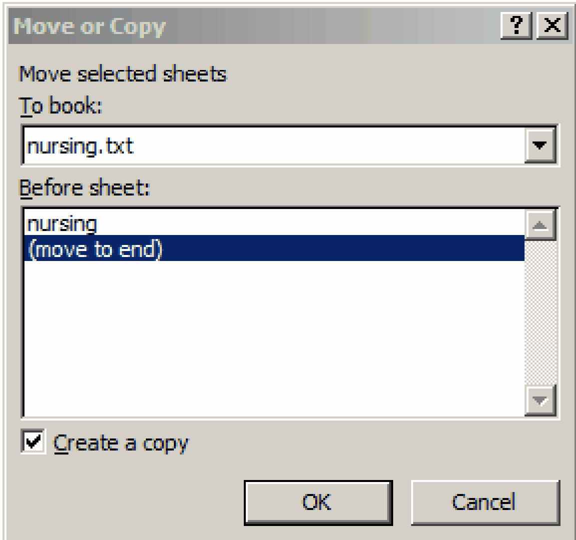
Does this look like a "Lertap workbook"?

No. We have to have a Data worksheet, and a CCs worksheet.

Fortunately, it's pretty easy to get kitted out with Data and CCs worksheets.

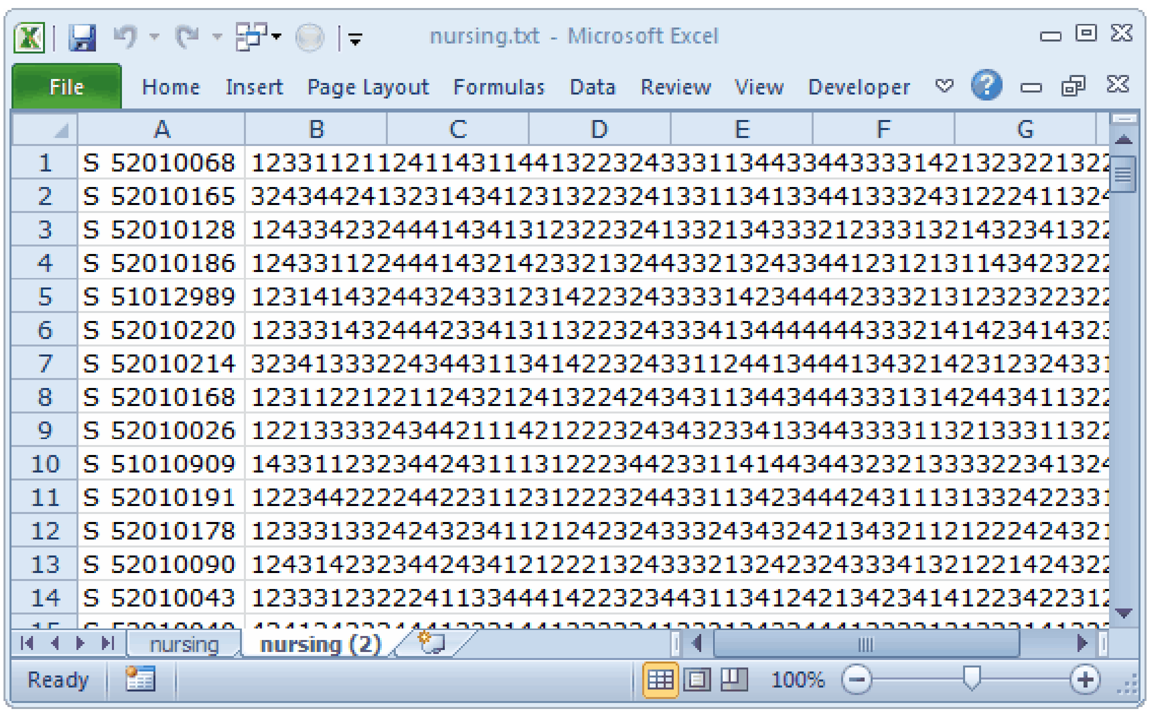
To get a Data worksheet, for example, we could simply change the name "nursing" to "Data". This would be 100% okay, but our own preference would be to make a copy of "nursing", and then turn the copy into our Data worksheet.

A right-click on the "nursing" tab will open up a little menu with an option to "Move or copy ...".

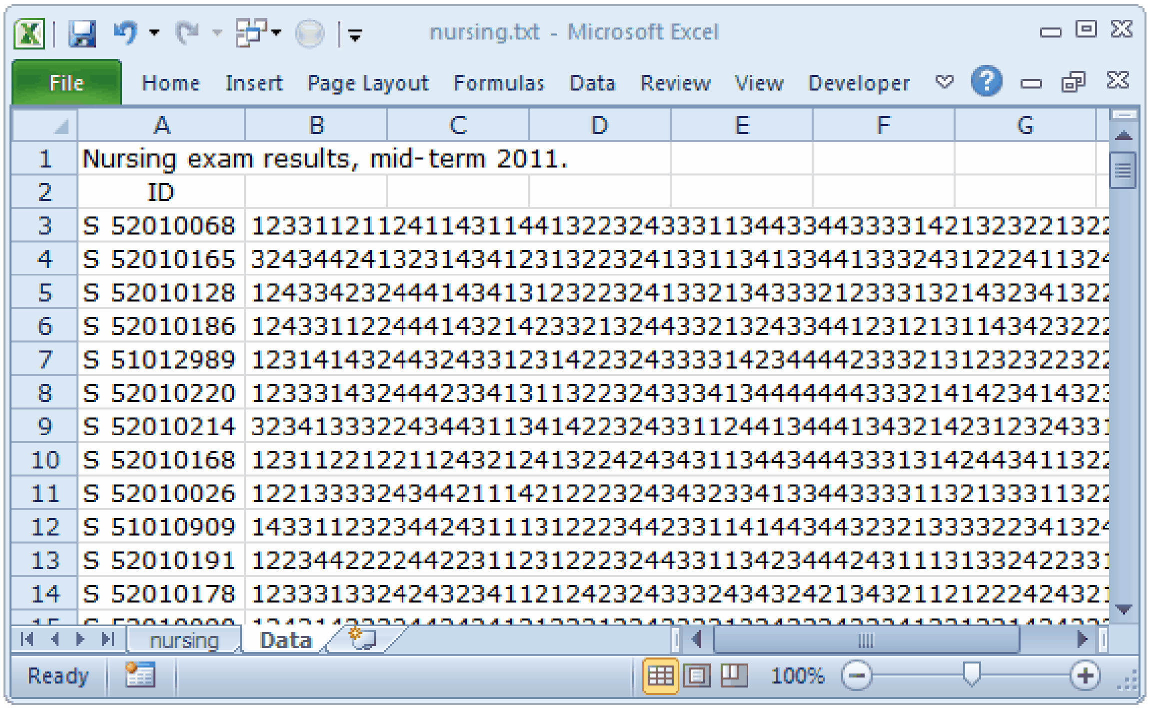


Be sure to put that tick in the "Create and copy" box, will you?

OK



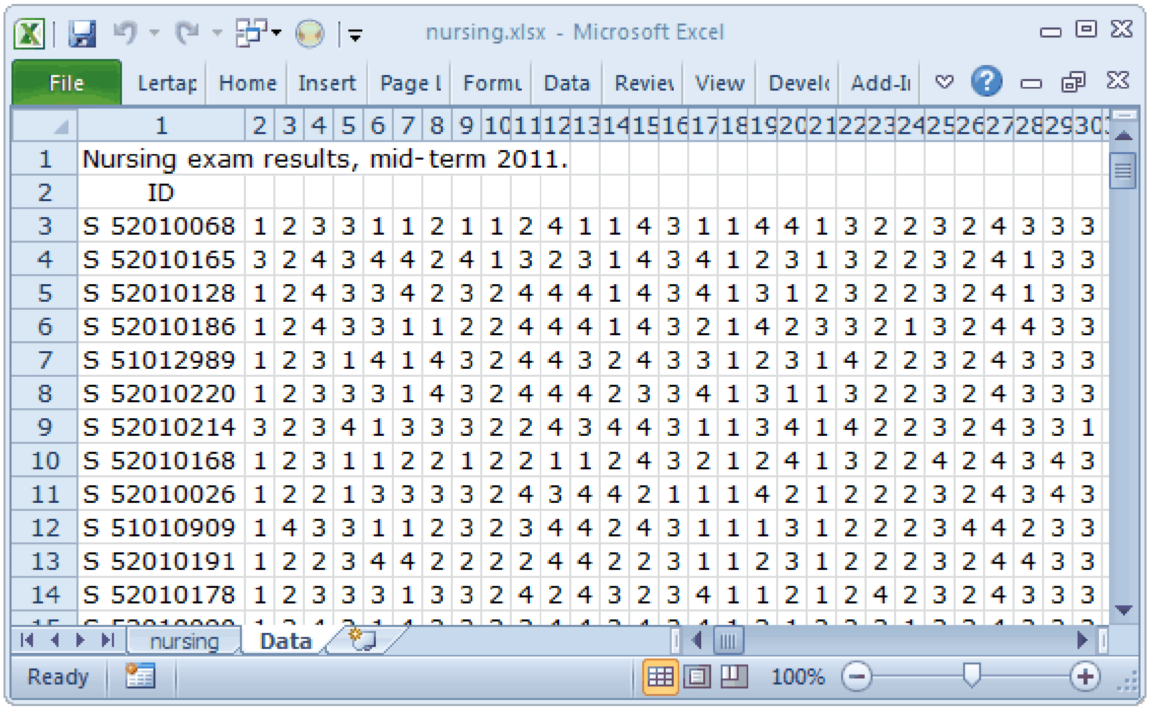
We change "nursing (2)" to "Data". We insert two new rows at the top of Data. We scribble a brief description of the data in the first row, and label the first column as "ID":



Note that Excel says we're still working with a file called "nursing.txt" (look at the very top, where it says "Microsoft Excel" in hard-to-see text). Change this to "nursing.xlsx" by going to the File tab, and using the usual "Save as" option.

Next, we want to distribute the item responses so that they're each ensconced in their own cosy little columns. If you paid attention to all the examples of Data sheets found in this website (or document), you will have noticed that item responses are always housed in individual columns.

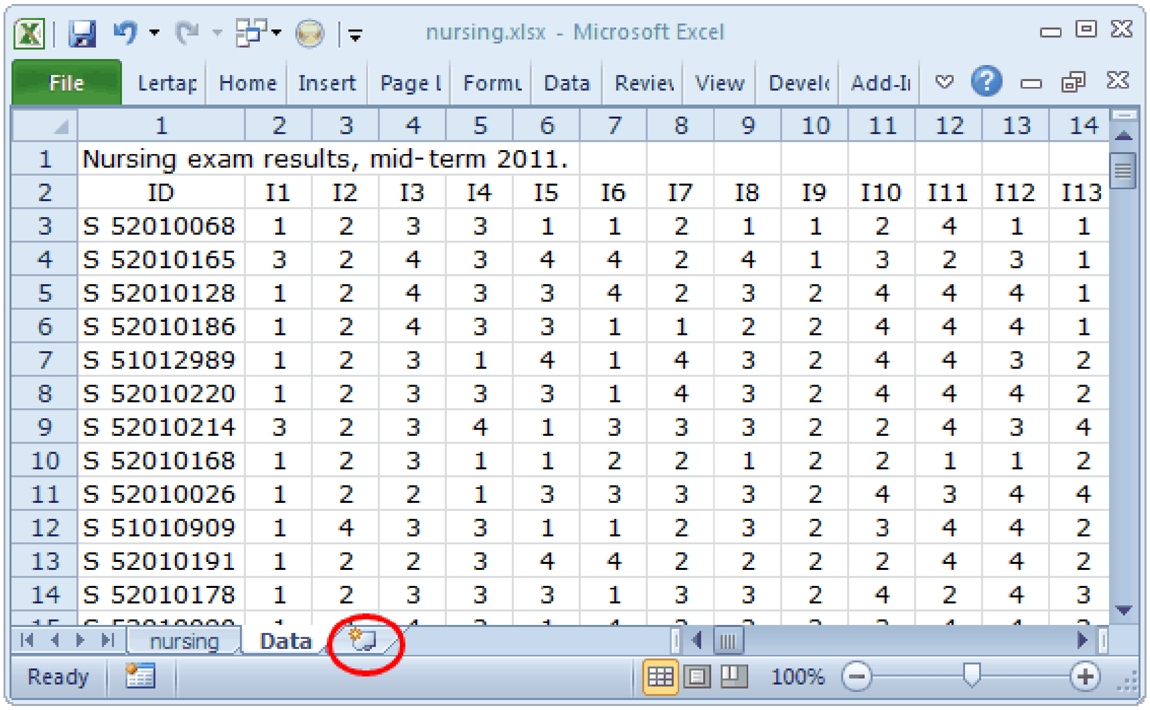
We have a need for The Spreader, a very handy little utility indeed (!). An example of its use is shown in the next topic. Here's what The Spreader did for us:



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Nursing exam results, mid-term 2011.																													
2	ID																													
3	S 52010068	1	2	3	3	1	1	2	1	1	2	4	1	1	4	3	1	1	4	4	1	3	2	2	3	2	4	3	3	3
4	S 52010165	3	2	4	3	4	4	2	4	1	3	2	3	1	4	3	4	1	2	3	1	3	2	2	3	2	4	1	3	3
5	S 52010128	1	2	4	3	3	4	2	3	2	4	4	4	1	4	3	4	1	3	1	2	3	2	2	3	2	4	1	3	3
6	S 52010186	1	2	4	3	3	1	1	2	2	4	4	4	1	4	3	2	1	4	2	3	3	2	1	3	2	4	4	3	3
7	S 51012989	1	2	3	1	4	1	4	3	2	4	4	3	2	4	3	3	1	2	3	1	4	2	2	3	2	4	3	3	3
8	S 52010220	1	2	3	3	3	1	4	3	2	4	4	4	2	3	3	4	1	3	1	1	3	2	2	3	2	4	3	3	3
9	S 52010214	3	2	3	4	1	3	3	3	2	2	4	3	4	4	3	1	1	3	4	1	4	2	2	3	2	4	3	3	1
10	S 52010168	1	2	3	1	1	2	2	1	2	2	1	1	2	4	3	2	1	2	4	1	3	2	2	4	2	4	3	4	3
11	S 52010026	1	2	2	1	3	3	3	3	2	4	3	4	4	2	1	1	1	4	2	1	2	2	2	3	2	4	3	4	3
12	S 51010909	1	4	3	3	1	1	2	3	2	3	4	4	2	4	3	1	1	1	3	1	2	2	2	3	4	4	2	3	3
13	S 52010191	1	2	2	3	4	4	2	2	2	2	4	4	2	2	3	1	1	2	3	1	2	2	2	3	2	4	4	3	3
14	S 52010178	1	2	3	3	3	1	3	3	2	4	2	4	3	2	3	4	1	1	2	1	2	4	2	3	2	4	3	3	3

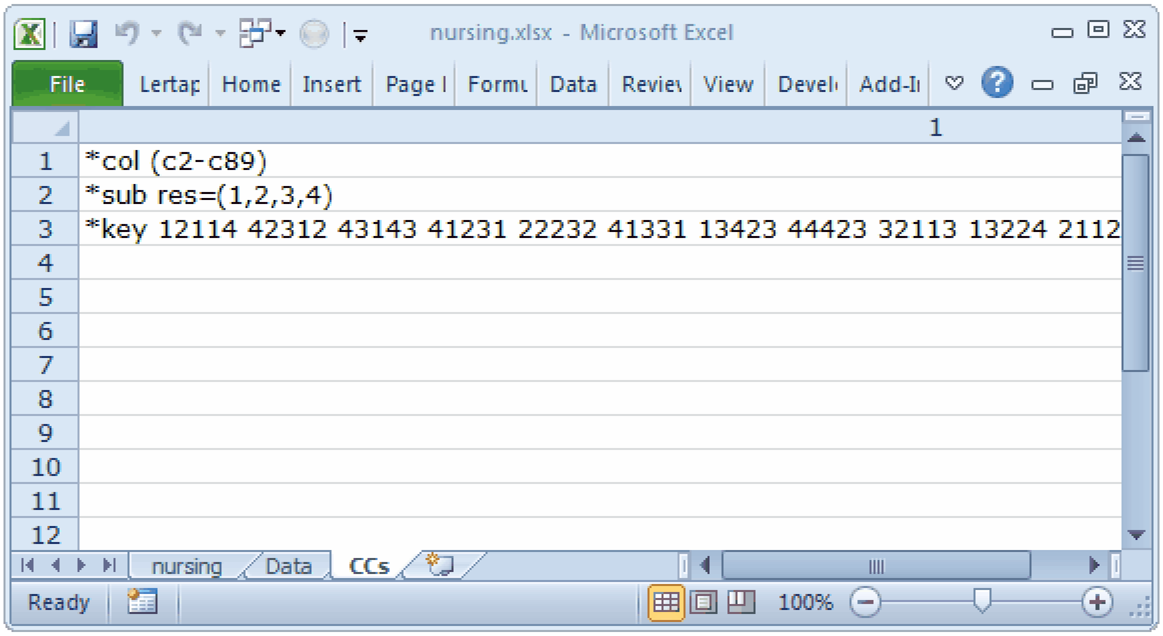
We put item numbers at the top of each column (below) . This is very very easy to do if you take the time to learn how to use Excel's "autofill" option.

(Notice how the columns have changed to numbers instead of letters? Lertap does this -- it much prefers this style, but you can change back to letters most easily. How?)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Nursing exam results, mid-term 2011.													
2	ID	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13
3	S 52010068	1	2	3	3	1	1	2	1	1	2	4	1	1
4	S 52010165	3	2	4	3	4	4	2	4	1	3	2	3	1
5	S 52010128	1	2	4	3	3	4	2	3	2	4	4	4	1
6	S 52010186	1	2	4	3	3	1	1	2	2	4	4	4	1
7	S 51012989	1	2	3	1	4	1	4	3	2	4	4	3	2
8	S 52010220	1	2	3	3	3	1	4	3	2	4	4	4	2
9	S 52010214	3	2	3	4	1	3	3	3	2	2	4	3	4
10	S 52010168	1	2	3	1	1	2	2	1	2	2	1	1	2
11	S 52010026	1	2	2	1	3	3	3	3	2	4	3	4	4
12	S 51010909	1	4	3	3	1	1	2	3	2	3	4	4	2
13	S 52010191	1	2	2	3	4	4	2	2	2	2	4	4	2
14	S 52010178	1	2	3	3	3	1	3	3	2	4	2	4	3

Just about ready, Freddy. We need to add a CCs worksheet. Click where the red circle indicates. The new sheet will come in as "Sheet 3" as we already have two other worksheets. Rename it as CCs. Add a *col line and a *key line. Oh yes, almost forgot: the items in this test used response codes of 1 2 3 4, so we need a *sub line too as these are not the default response codes for cognitive items (Lertap assumes response codes are A B C D unless you say otherwise):



Bob's your uncle now, for sure -- party time is here big time. Fix yourself up with a cup of hot party cocoa, drop in a marshmallow or two, and get Lertap to run.

What say? You've already forgotten how to run Lertap? Why, you just have to take a few steps:

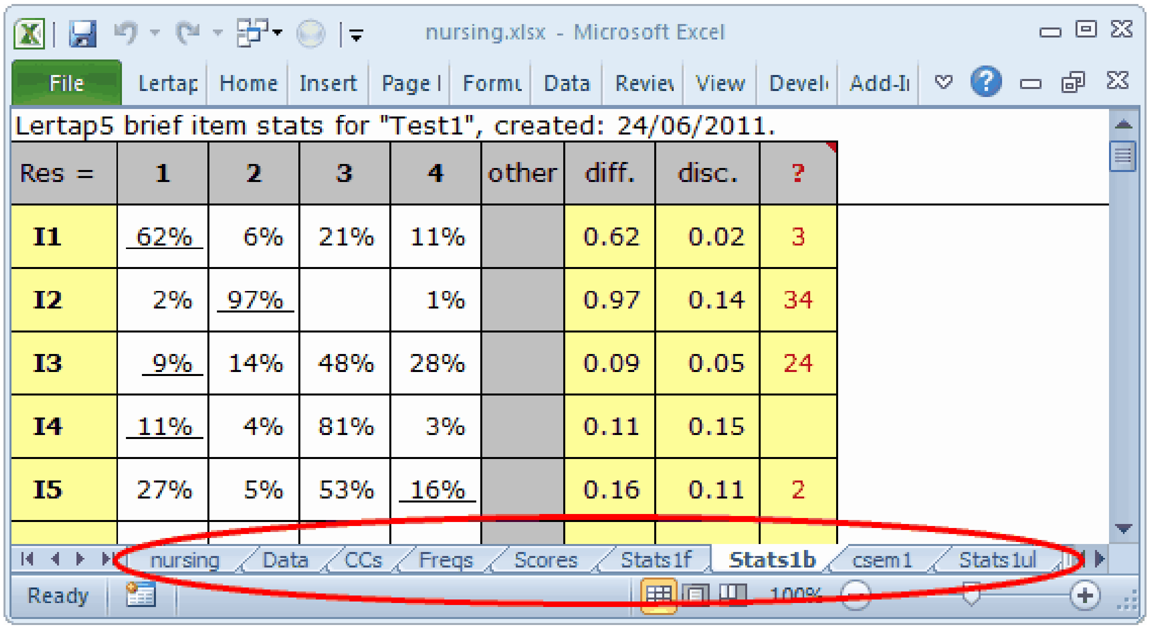
The 3 steps

1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

After stepping out, Lertap will leave all sorts of new reports to look at:



Res =	1	2	3	4	other	diff.	disc.	?
I1	62%	6%	21%	11%		0.62	0.02	3
I2	2%	97%		1%		0.97	0.14	34
I3	9%	14%	48%	28%		0.09	0.05	24
I4	11%	4%	81%	3%		0.11	0.15	
I5	27%	5%	53%	16%		0.16	0.11	2

Done.

Tidbits:

It's possible to persuade scanners to create data files which are ready for Excel, obviating the need to use the Text Import Wizard. One very common approach is to get the scanner to create a "**csv**" file, a "comma-separated values" file. These files will whiz right into Excel. See this example.

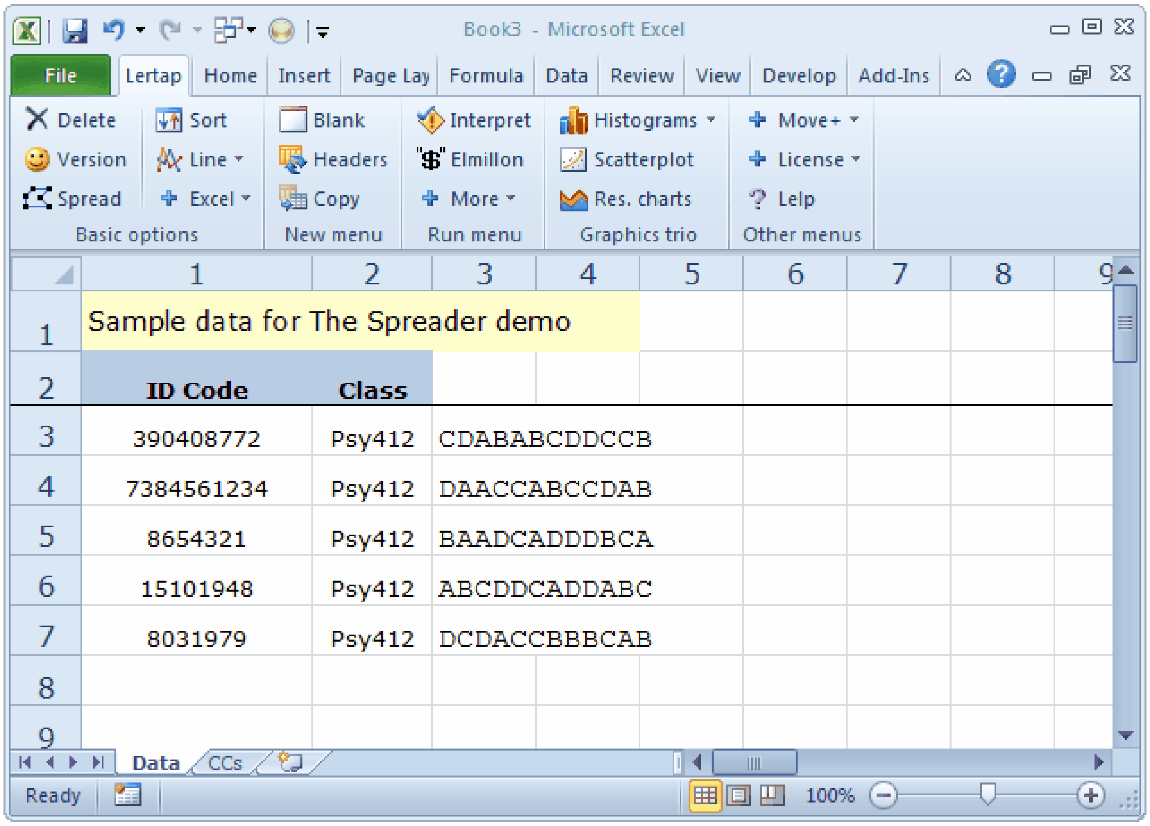
It may very well be the case that the people who run your scanner service can be a big help -- ask. It's probably true that 92.4% of the world's Lertappers get their test data prepared by a scanner. While the Text Import Wizard isn't hard to negotiate, most Lertap users seem to have solid support from the folks who run the scanners.

5.2 The Spreader

The Spreader is a handy little tool that can do a big job when it comes to data entry, Excel, and Lertap.

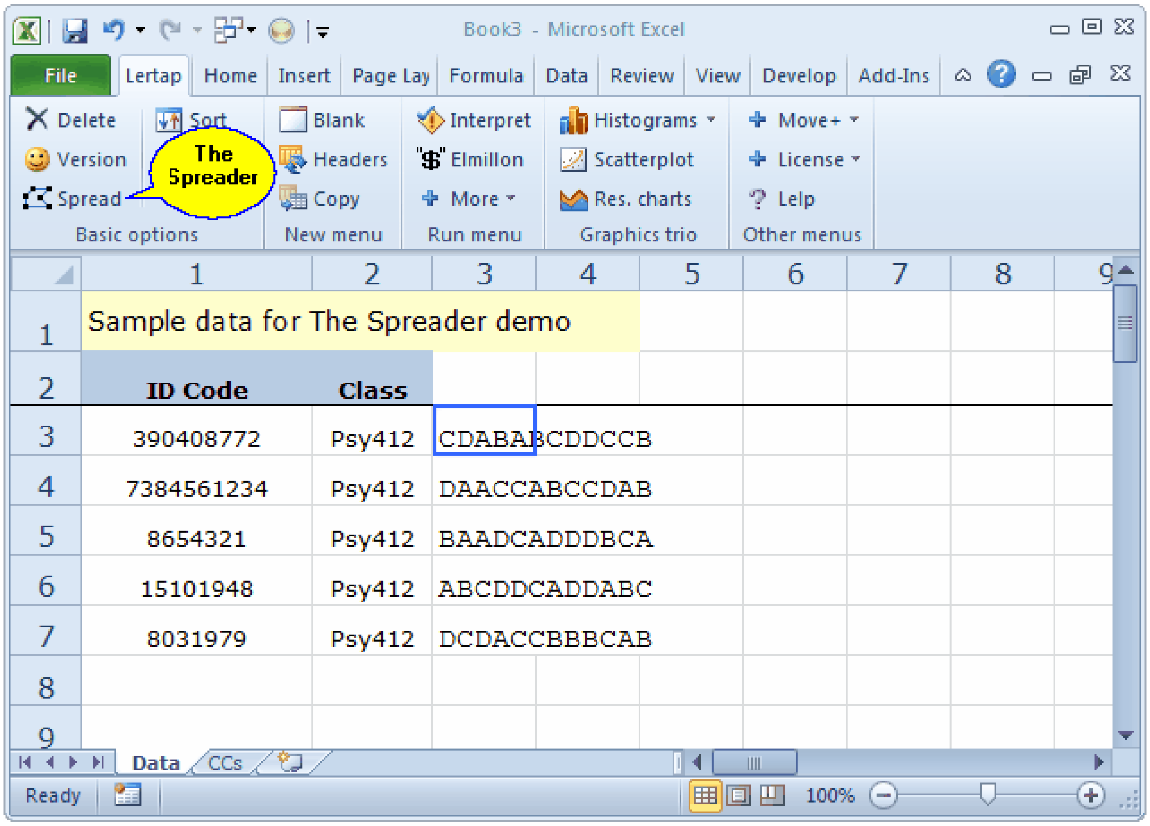
Consider the sample Data worksheet below. Note that each student's 12 item responses have been typed into an unbroken string.

The string of answers begins and ends (in this example) in column 3. At this point, all columns to the right of column 3 are empty.

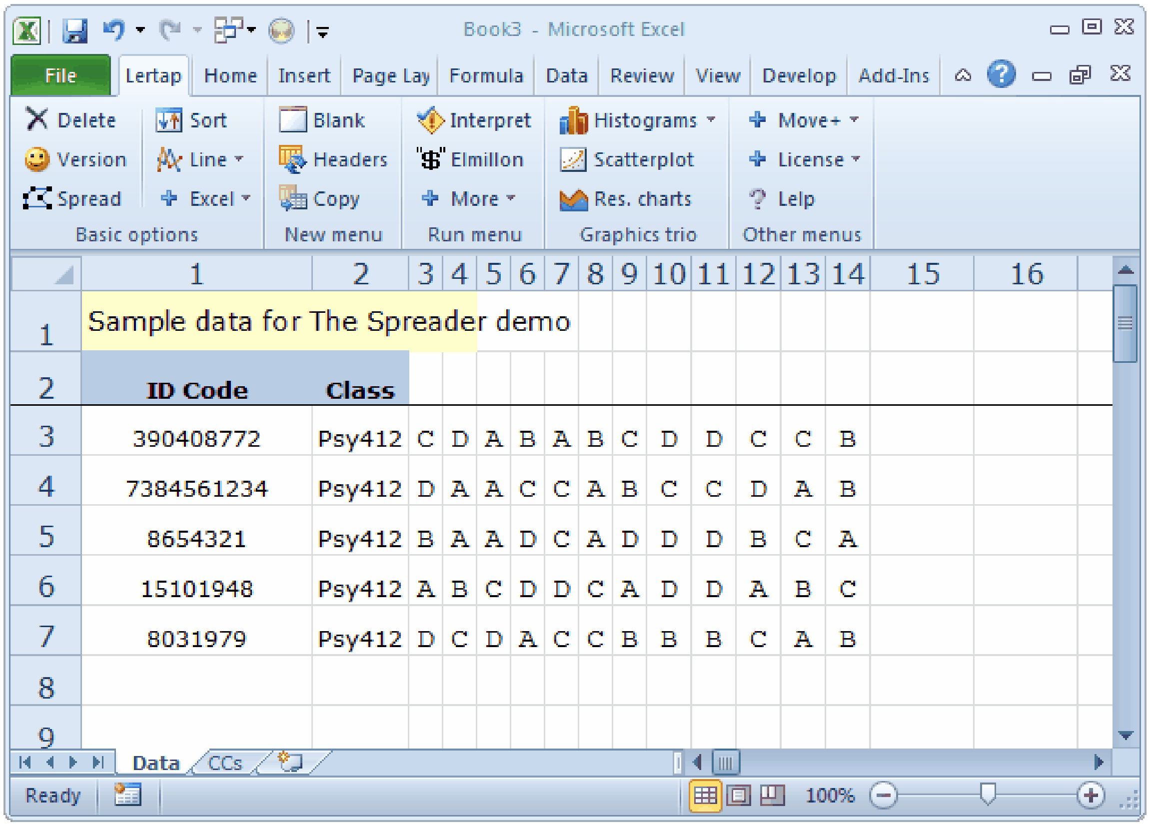


Were we to click on the Interpret option at this point, thinking that we'd start getting results, Lertap would choke. It wants each item response to be housed in its own column, one response per column. Here we have all 12 responses in a single column. It was faster to type them in this way, without having to use the tab key before entering each response. The output from scanners will very often be similar, with all item responses in a long string, with nothing separating them.

Not to worry, not to worry. The Spreader will take these strings apart.



Before clicking on "Spread", we position the cursor where the blue box is, that is, on the string of responses for the first student. Then a click on Spread, and:



How sweet it is. Now each response is parked in a single column. *Bob's your uncle*, mate!

Read more about spreading responses here.

5.3 Scanner data import (2)

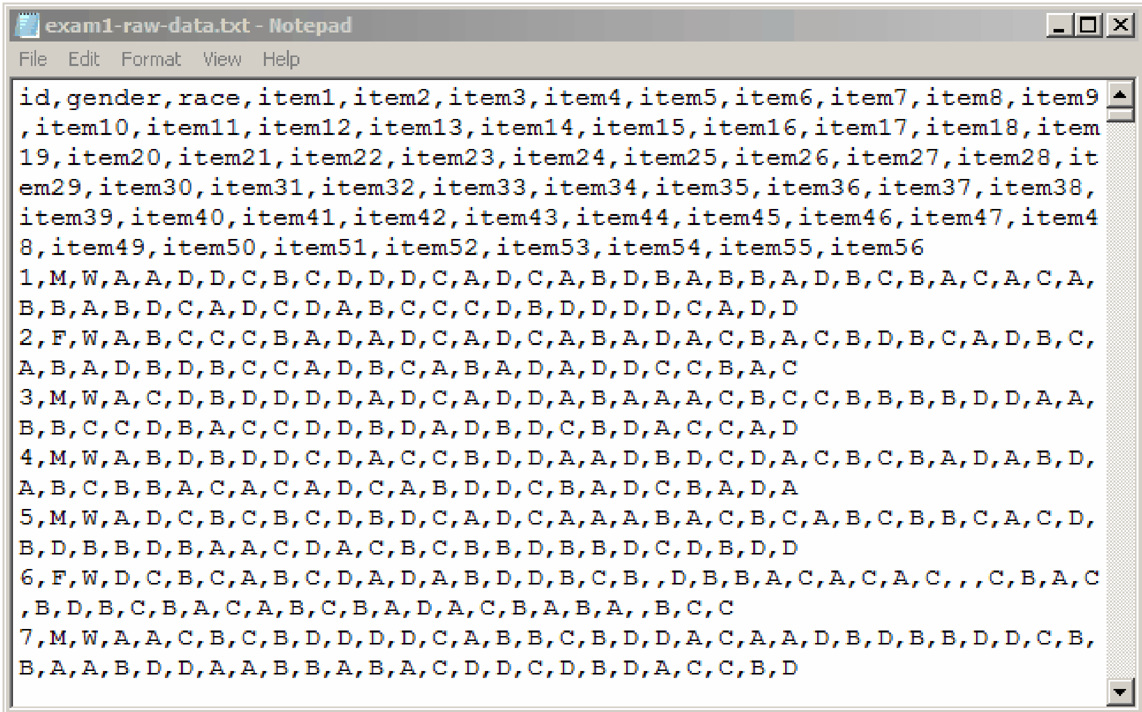
Back a topic or two it was suggested that scanners can be persuaded to create a "**csv**" file, where csv stands for "comma-separated values". We've got an example here, thanks to work done at the University of Virginia. (Get your scanner to output csv files, and you'll be in Fat City, as we will point out.)

J. Patrick Meyer, author of the jMetrik software system, has provided several sample datasets which jMetrik users may cut their teeth on.

One of these is called "**Exam1** - A data file of 6,000 cases and 56 binary items". A click on the website's option to download the dataset will see a file called "exam1-raw-data.txt" make its way to your computer.

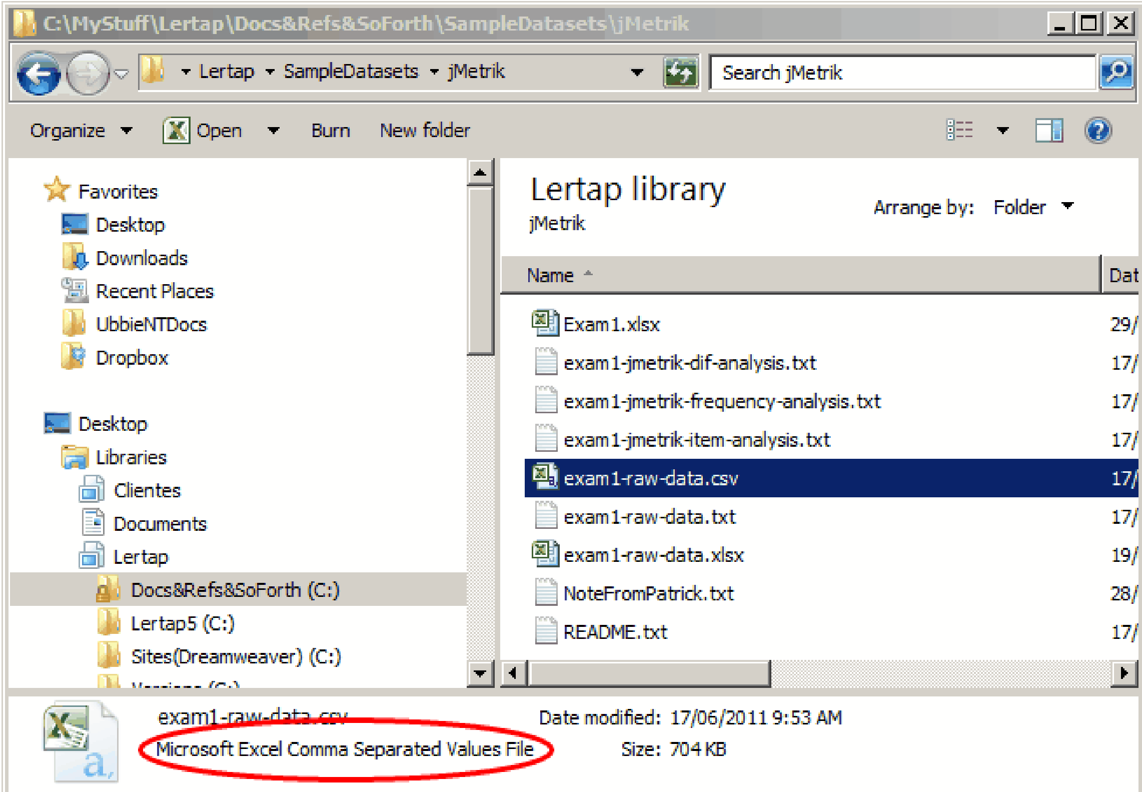
You could use Excel's Text Import Wizard to create an Excel worksheet from this file, exactly as we did before. It would do a good job, too.

But we might use the Windows **Notepad** program to have a little squiz of this file first. Doing so will work to our great advantage in this case. Have a gander:

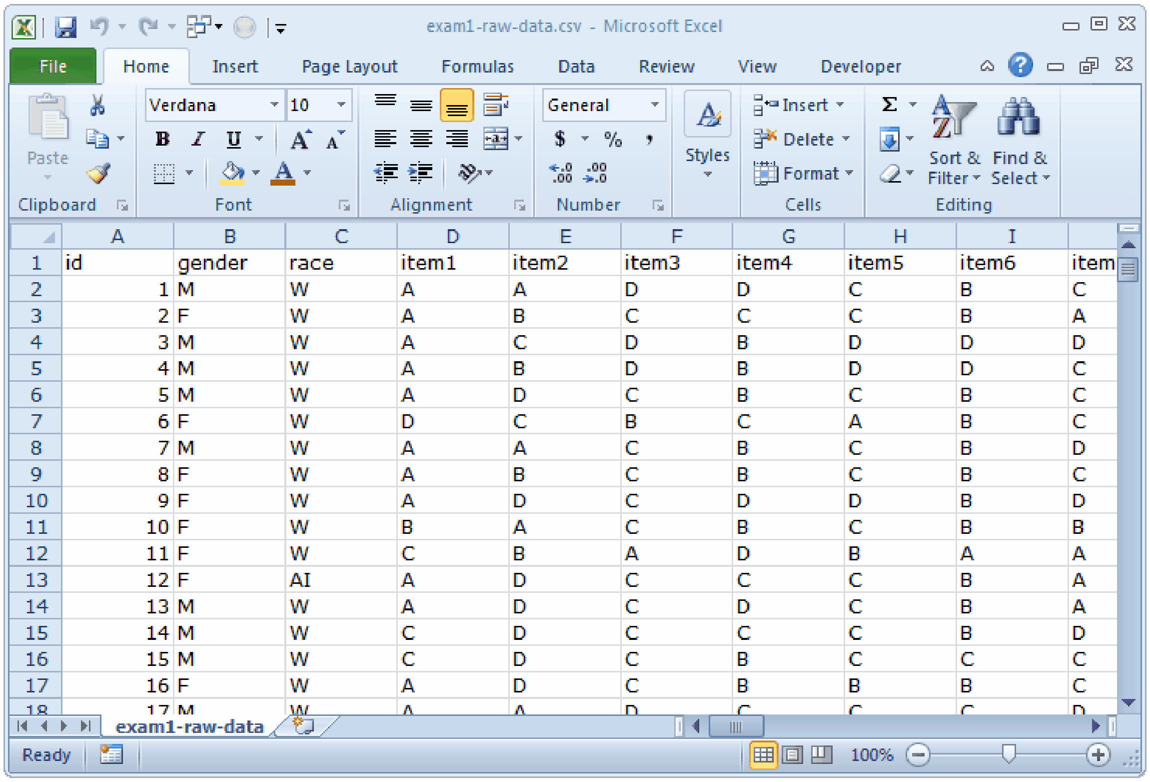


If you had never seen a csv file before, those days are gone. This is a csv file, and you have seen it.

Instead of using the Text Import Wizard to Excel-ize this file, we'll just make a copy of it, and then change the extension on the copy to csv. We did this while you were sipping on your cup of hot cocoa, making a file called "exam1-raw-data.csv". You can see this file listed here:



Window's default program for handling csv files is Excel. This being the case, all we have to do now is open the file. Here's what we see:



Is this magic, or what? What we had to do, essentially, was just one thing: change the file's extension from "txt" to "csv". No need at all to call on the Text Import Wizard with its multiple steps. Here, in a single baby step, we got what we wanted. Very neat.

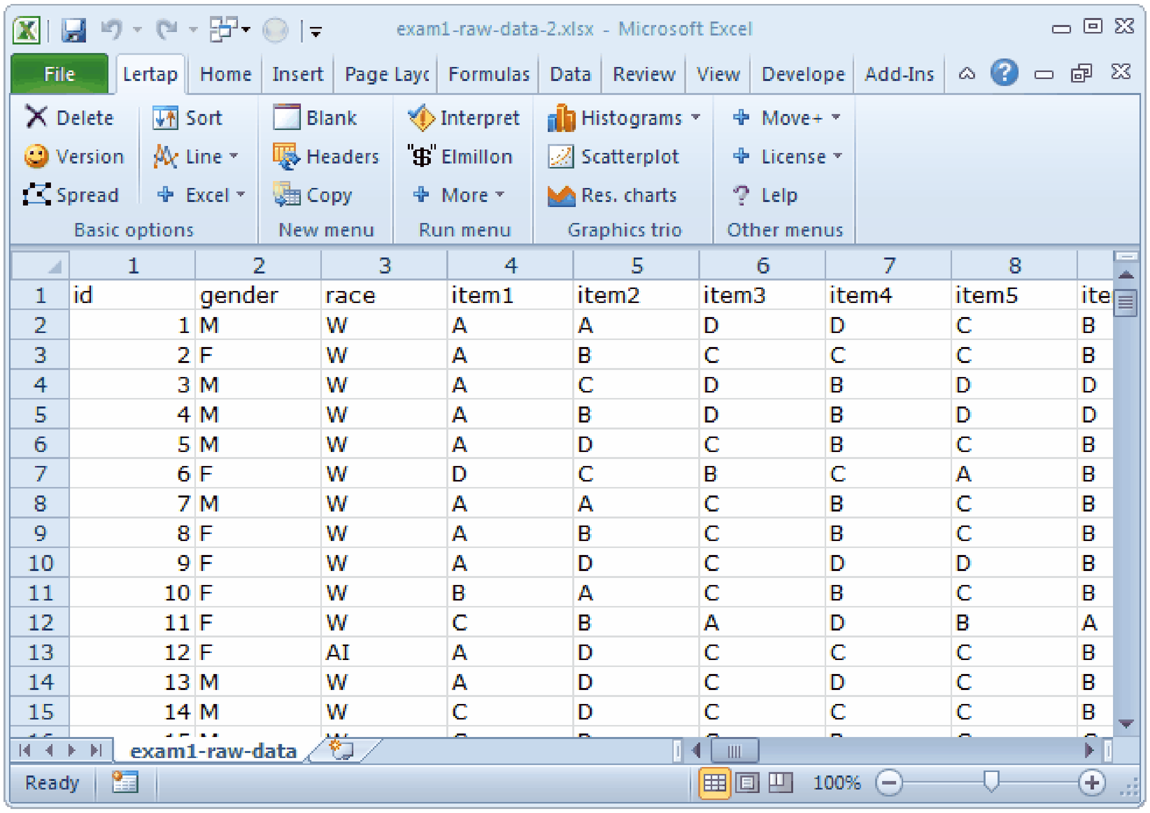
But now we ask: is this file ready for Lertap?

Most certainly not. There are a few steps to go through before Lertap will process the item responses found in the file.

We'll start by using the File menu to save this workbook as a dinkum Excel xlsx file, and then meet you in the next topic.

5.3.1 Over to Lertap

We're breezing into this topic with an Excel workbook which looks like this:

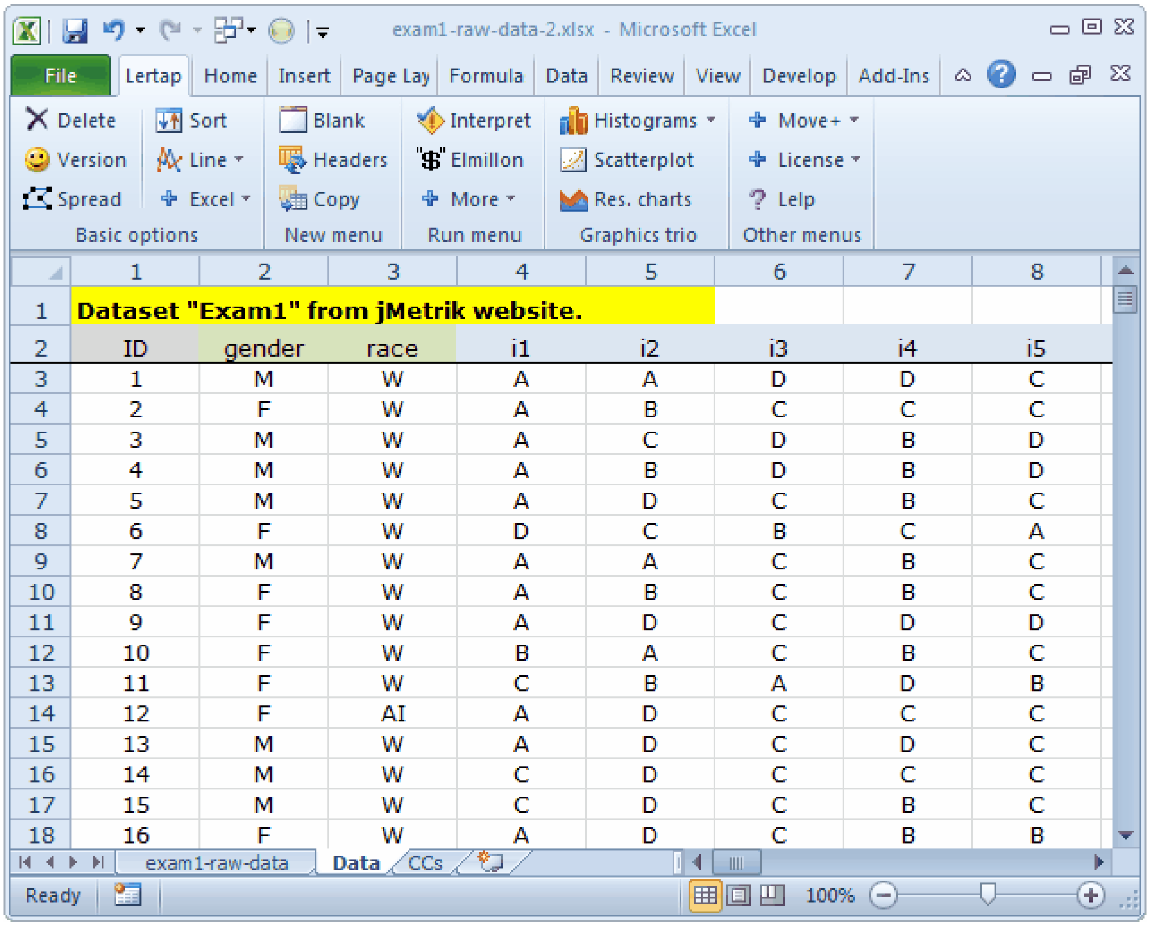


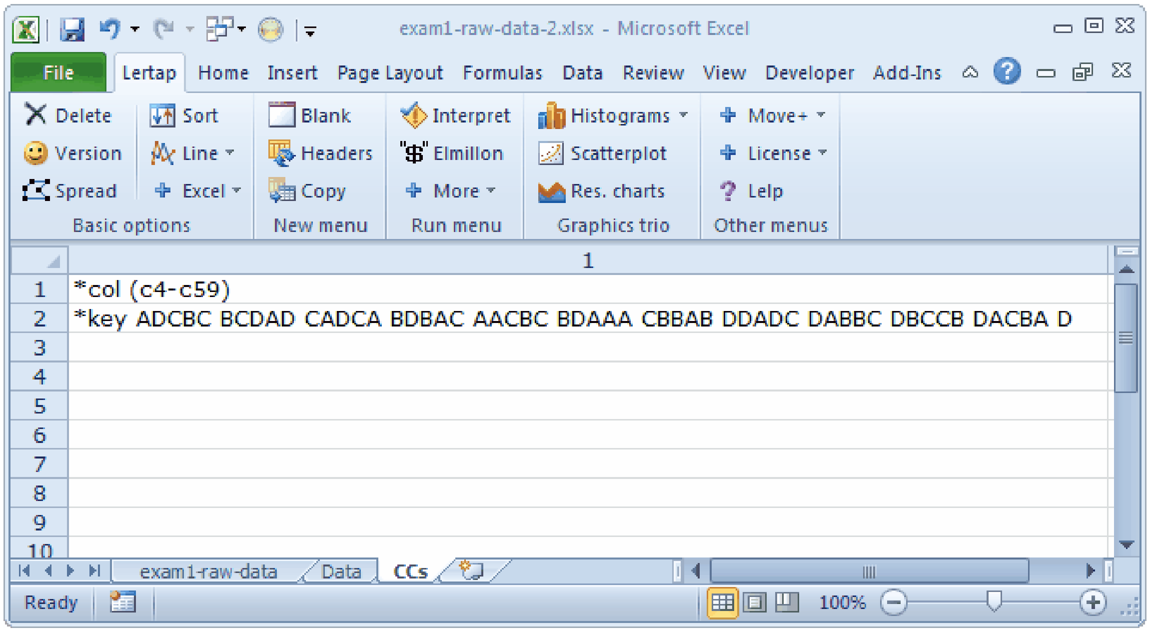
We want to use Lertap to process the data found in this worksheet. We need to have two new worksheets in order to do this. One of these new sheets has to be called "Data", while the other must be named "CCs".

The steps we started with in the last "Over to Lertap" topic serve as a great example. We'll make a copy of the present worksheet, "exam1-raw-data", and name the copy "Data". We'll insert a new row at the top because Lertap wants the columns headers (id, gender, race, item1 ...) to be in the second row.

Then we'll change the names of the items so that they will display a bit better when Lertap gets Excel to make charts. (This isn't necessary at all, but it's sort of a personal preference here at Lertap HQ. Short item handles have some practical advantages when the charts are made. Some discussion of this may be found here.)

Finally, we'll insert a new worksheet, calling it "CCs".


exam1-raw-data-2.xlsx - Microsoft Excel
File
Lertap
Home
Insert
Page Lay
Formula
Data
Review
View
Develop
Add-Ins
Delete
Version
Spread
Sort
Line
Excel
Basic options
Blank
Headers
Copy
New menu
Interpret
Elmillon
More
Run menu
Histograms
Scatterplot
Res. charts
Graphics trio
Move+
License
Lelp
Other menus
Dataset "Exam1" from jMetrik website.
ID gender race i1 i2 i3 i4 i5
1 M W A A D D C
2 F W A B C C C
3 M W A C D B D
4 M W A B D B D
5 M W A D C B C
6 F W D C B C A
7 M W A A C B C
8 F W A B C B C
9 F W A D C D D
10 F W B A C B C
11 F W C B A D B
12 F AI A D C C C
13 M W A D C D C
14 M W C D C C C
15 M W C D C B C
16 F W A D C B B
exam1-raw-data
Data
CCs
Ready
100%



With everything now set up, we can follow those Lertap steps you already know about and have come to cherish.

The 3 steps

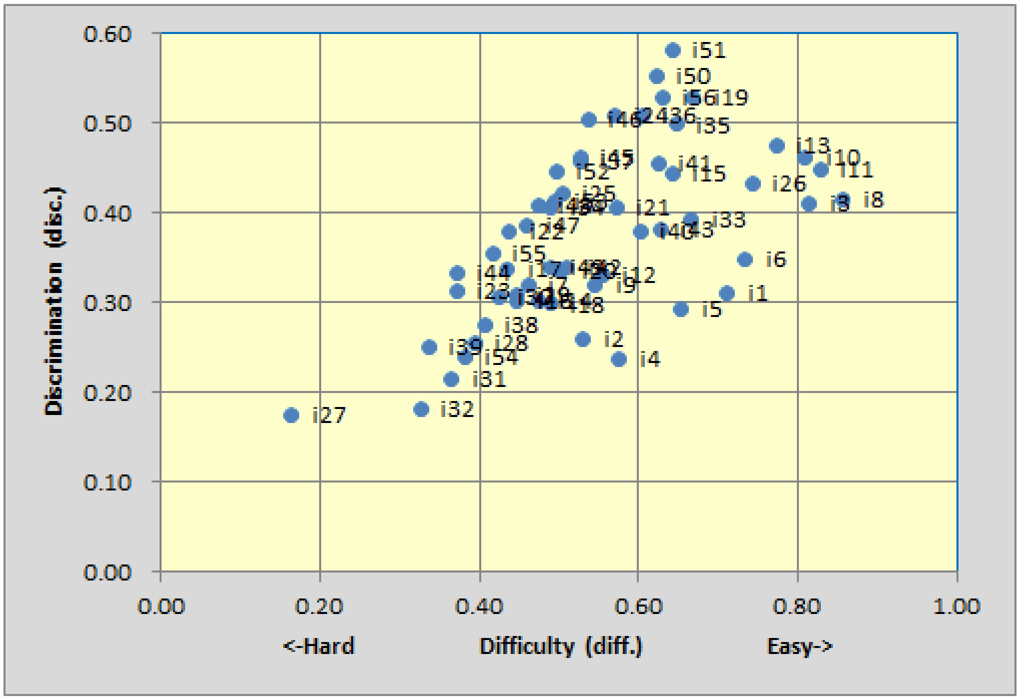
1) Make sure the workbook which has the data to be processed is active.

2) Use the **Interpret** option on the Lertap tab.

3) Use the **Elmillon** option on the Lertap tab.

Should you be stuck inside your office, supposedly marking student assignments, but just itching instead to work through this sample, we can tell you that Lertap's Stats1f report indicated a reliability figure (coefficient alpha) of 0.91 for this test.

The Stats1b scatterplot of item difficulty by discrimination looked like this:



Tidbits:

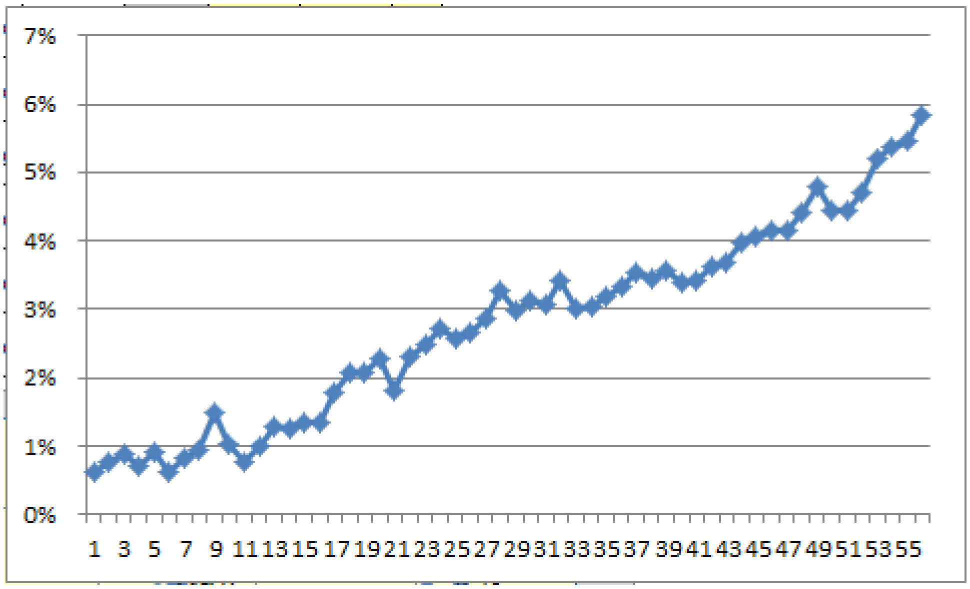
In May 2016 a special macro was created to do the work described above, making it much, much easier to bring csv files into Lertap. The macro is called "ImportCSV".

There are without doubt numerous exercises that could be undertaken with this dataset.

We might begin by looking at the quality of the data, browsing down the **Freqs** report to check make sure that the responses found in the Data sheet's columns were as expected (A B C D with some entries anticipated in the ? rows).

This seemed to be a speeded test, that is, a test with an allotted time too short for all students to get to the end of the items. If this were the case, we should see the percentage of ? entries increasing as we work down the Freqs report (and indeed we do -- if you have a close look at the data, you'll find that 20 some records were missing responses on the first 32 items of the test, a curious state of affairs).

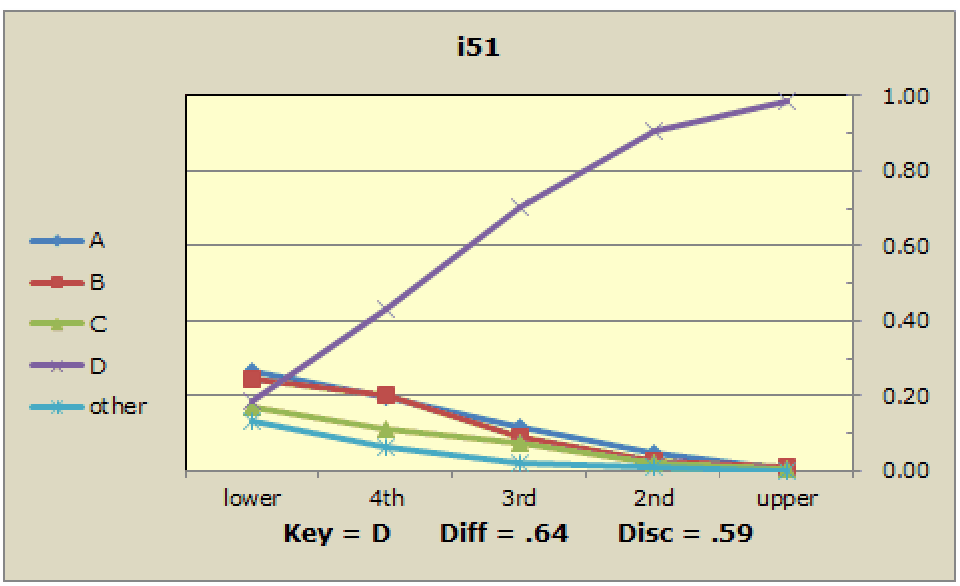
The following chart plots the percentage figure from Freqs' ? row for all of the items. How do you suppose this was done? (Hint: use a column from the Stats1b report. It takes all of two seconds to get such a graph if you use the "Line" option on Lertap's "Basic options" menu.)



Does a histogram of the test scores suggest any anomalies, such as an outlier or two or more at the low end?

How many of the test's 56 items had flags in the Stats1b **?** column? What may have been the problem with these items? How would coefficient alpha change if these problems were addressed, perhaps by double-keying (assuming the problems are substantial enough to warrant this; they may not be)?

What do the quintile plots for these items look like? Here's one:



What about the matter of group differences? This dataset has two demographic variables, gender and race. If you need more of an excuse to keep from marking assignments, you have it.

5.3.2 Download

This example is based on Patrick Meyer's **jMetrik** work at the University of Virginia.

The dataset may be downloaded here.

(This link was current July 2011. Note that, unlike the other datasets referenced in this website, this one is **not** from Lertap HQ.)

5.4 Iteman

Iteman 3 users may import their DAT files directly by using a special Lertap macro.

More information about this macro may be found at this page:

http://www.lertap5.com/HTMLHelp/HTML/index.html?iteman.htm

5.5 Fonts

If you use the "Blank" option on Lertap's New menu to make a fresh workbook, it will try to set the Data worksheet's font to **Verdana**, and the CCs sheet's font to **Courier New**. It does this as these two fonts seem to work well with both Windows and Macintosh computers. As Lertap runs, it will try to use the Verdana font for almost all of its many reports (which are of course Excel worksheets).

(Using a fixed-pitch font in the CCs worksheet, such as Courier New, has the advantage of getting things to line up. CCs sheets which use, for example, *key, *alt, *wts, and *pol lines will often be easier to check for errors when a fixed-pitch font is used. But this is no big deal. In practical terms, it doesn't matter a hoot what font is used in the CCs worksheet.)

The "Headers" and "Copy" options on the New menu will create workbooks having Data and CCs worksheets with inherited fonts. Whatever fonts were in the original workbook's Data and CCs worksheets will carry over to the new workbook.

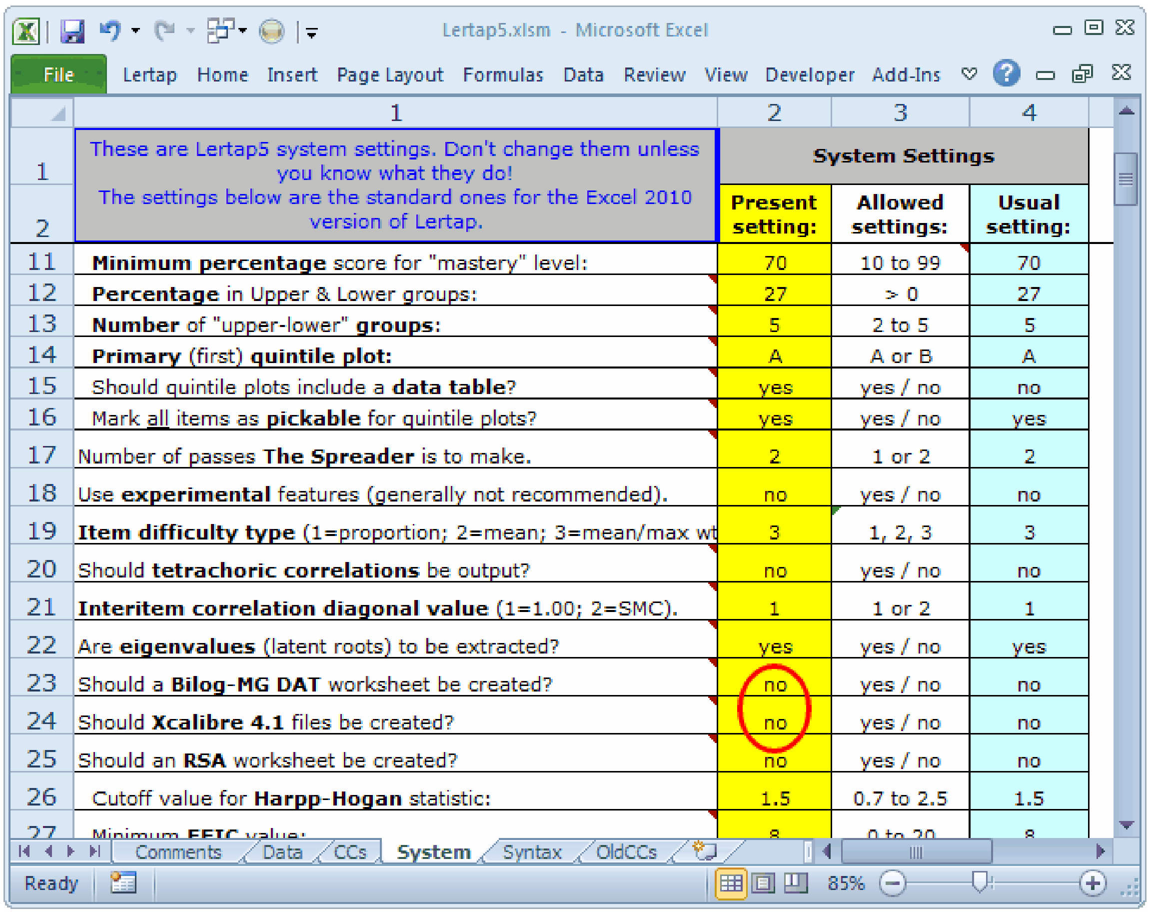
As mentioned in the previous topic, you don't have to use Lertap's New menu to set up a new workbook. You can do it with Excel, using the New option on the File tab. If Excel is used to create a new workbook, its sheets will use the "body" font (probably Calibri), or whatever has been set as the default font to be used "When creating new workbooks", a setting found by working through the File tab to get to Options, then to General, then to the "Use this font:" setting.

The matter of fonts is important as using some fonts, such as Times New Roman, will result in Lertap reports which may be rather poorly formatted.

6 Exporting Data

We'll mention just two common activities for life after Lertap (if that's possible): moving from Lertap to an IRT (item response theory) program such as **Bilog MG** and **Xcalibre**, and getting the **SPSS** package to pick up data from a Lertap workbook. (Note for **Iteman 4** users: Iteman 4 uses the same input files as Xcalibre.)

The way to get data from Lertap into a format acceptable to Bilog MG and / or Xcalibre begins with some settings in Lertap's System worksheet:



Change one or both of the no's, circled in red, to yes, save the Lertap5.xlsm workbook, and the next time the "Item scores and correlations" option is taken, new worksheets will be made for one or both of these programs. And note: the format of these new worksheets is such that, with minor alterations, they may also be workable with other programs.

Read more about Lertap and Bilog MG here, and about Xcalibre here.

SPSS is discussed in a later topic.

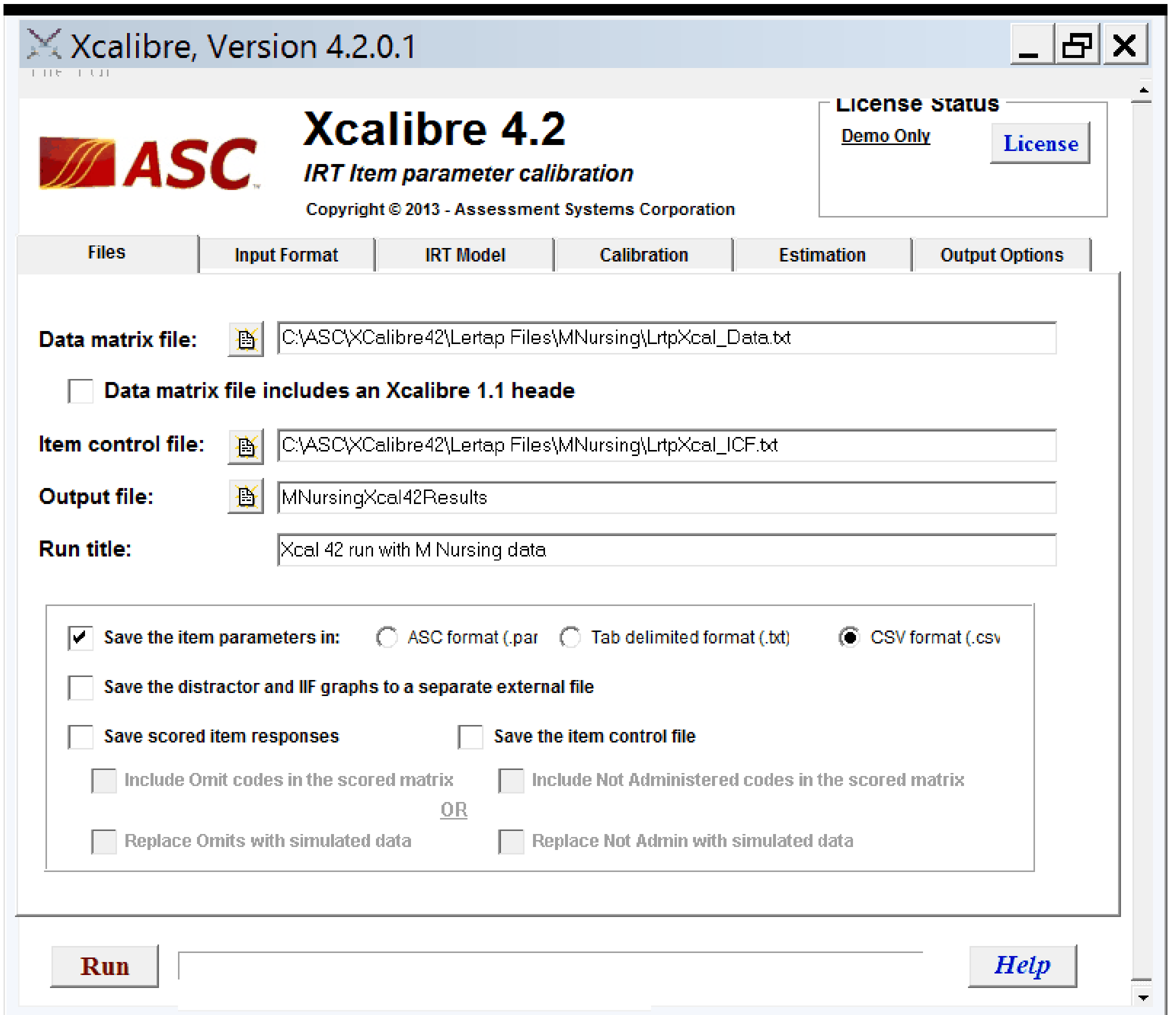
6.1 To Xcalibre 4.1

Xcalibre is a program created by Assessment Systems Corporation, ASC.

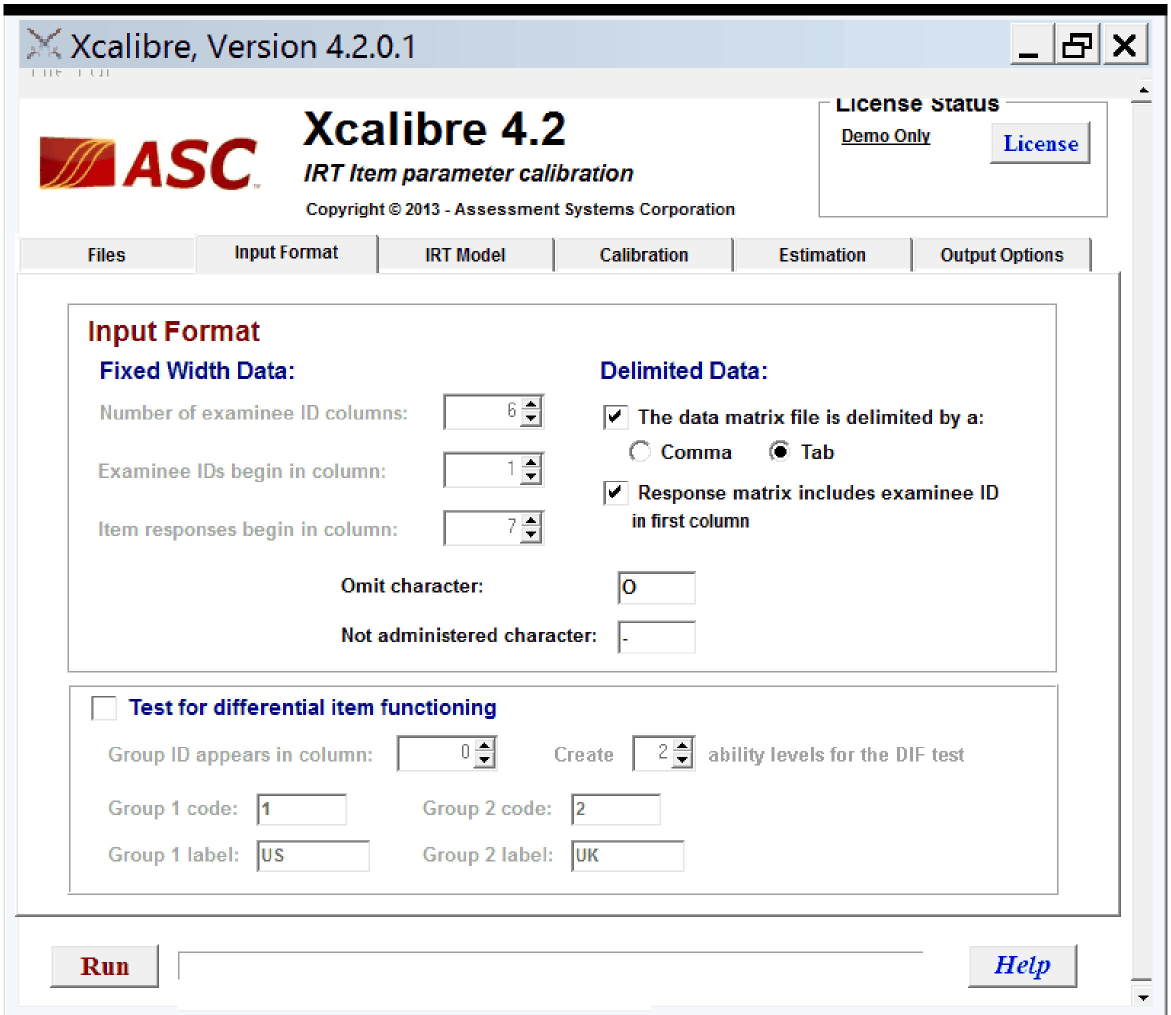
At August 2014, the latest version of this program was Xcalibre 4.2.

When Lertap's Xcalibre option is set to "yes", as mentioned in the previous topic, Lertap will create the "Data matrix file" and the "Item control file" required by Xcalibre 4.2 whenever the "Item scores and correlations" option is taken from the Run menu. The name of the data matrix file will be **LrtpXcal_Data.txt**, while the item control file's name will be **LrtpXcal_ICF.txt**. Both of these files are what's known in the business as "tab-delimited text files".

We had Lertap create these two files using the M.Nursing workbook. We then opened up Xcalibre 4.2 and filled in the appropriate boxes in its **Files** tab, as pictured here:

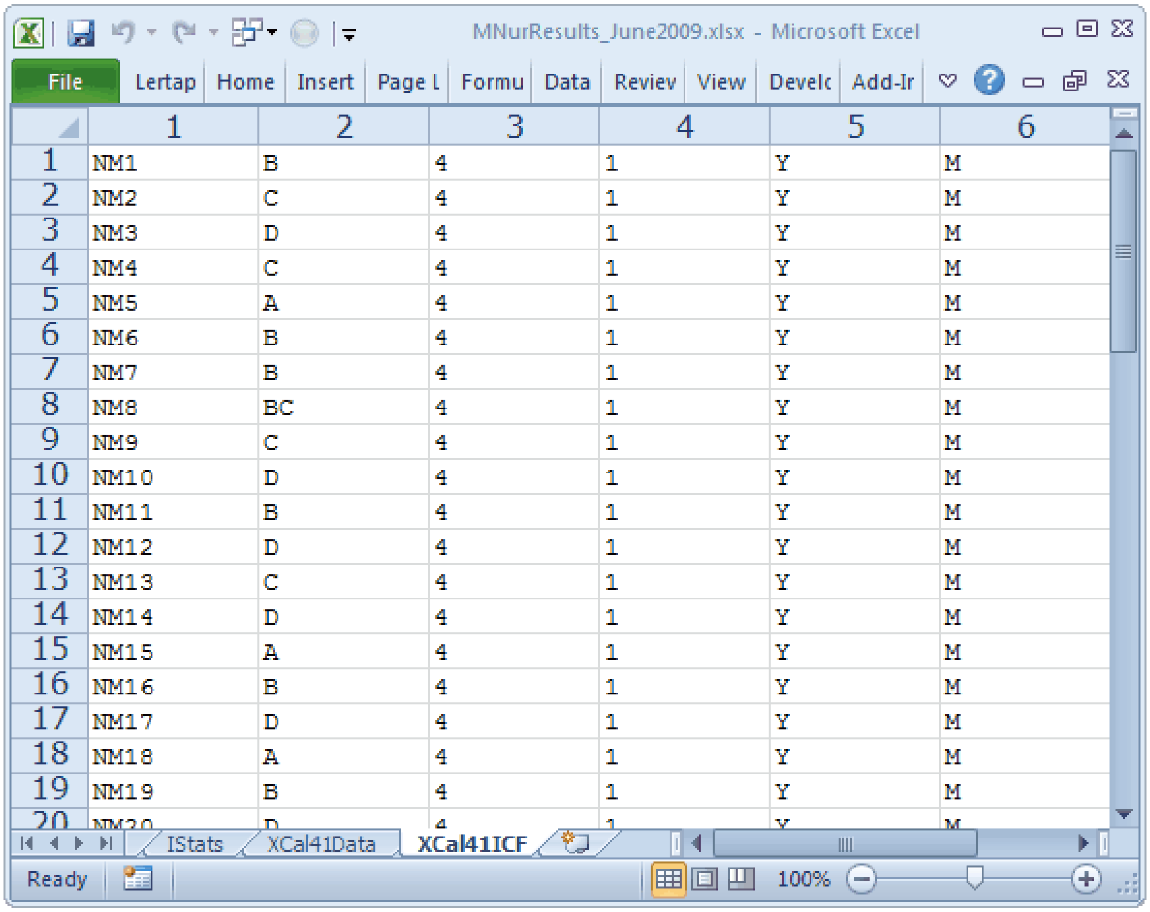


As mentioned above, Lertap's two Xcal files, one for data and one for "ICF", are both "tab delimited" files. We need to tell Xcalibre this, and the place to do it, in Xcalibre 4.2, is on the **Input Format** screen:



By the way, Lertap also creates two new worksheets when the Xcalibre option is set to yes. They're called **XCal41Data** and **XCal41ICF**. They have exactly the same data as contained in the two corresponding files, LrtpXcal_Data.txt and LrtpXcal_ICF.txt. (Technically, the worksheets should be called Xcal42Data and Xcal42ICF, but the information expected in these two files is the same for both Xcalibre 4.1 and Xcalibre 4.2, so it makes no difference.)

Here's a sample XCal41ICF worksheet:



	1	2	3	4	5	6
1	NM1	B	4	1	Y	M
2	NM2	C	4	1	Y	M
3	NM3	D	4	1	Y	M
4	NM4	C	4	1	Y	M
5	NM5	A	4	1	Y	M
6	NM6	B	4	1	Y	M
7	NM7	B	4	1	Y	M
8	NM8	BC	4	1	Y	M
9	NM9	C	4	1	Y	M
10	NM10	D	4	1	Y	M
11	NM11	B	4	1	Y	M
12	NM12	D	4	1	Y	M
13	NM13	C	4	1	Y	M
14	NM14	D	4	1	Y	M
15	NM15	A	4	1	Y	M
16	NM16	B	4	1	Y	M
17	NM17	D	4	1	Y	M
18	NM18	A	4	1	Y	M
19	NM19	B	4	1	Y	M
20	NM20	D	4	1	Y	M

Item IDs are in the first column. The Xcalibre 4.1 manual says that item IDs cannot have embedded spaces. In Lertap they can but, if they do, Lertap strips the spaces out before it writes the XCal41ICF worksheet.

Item keys (the keyed-correct answer(s)) are found in the second column; note that I8 has been double-keyed. This was done by using *mws lines in the CCs worksheet, as exemplified in this topic).

The third column has the number of options used by the items. This is followed by the "domain" column, column 4. A "domain" in Xcalibre is a "subtest" in Lertap. In this case, the first M.Nursing subtest was used as the source of the data seen in the worksheet above, so a "1" appears in the column.

Xcalibre will entertain multiple domains. In a basic test of mathematics, for example, one domain might be "addition", one "subtraction", one "division", and another "multiplication". In Lertap these "domains" would be "subtests". In order to get Lertap's LrtpXcal_ICF.txt file and the XCal41ICF worksheet to include multiple subtests, you'd have to have one global subtest in Lertap which incorporates all test items (quite easy to do). Then, you'd have to edit column 4 so that it refers to the number of the subtest, or "domain", which each item belongs to (write to support@lertap.com for assistance if needed).

If an item has been excluded from a subtest ("domain"), column 5 will have "N", for no. If an item is a "pretest" item, the "Y" in column 5 would want to become a "P", and you'd have to put it there.

The last column, 6, indicates item type. In this example, all items are of type "M", multiple-choice. If the Lertap subtest is an affective one, column 6 will have "R" (for "rating"), and column 2 will have a + or - sign to signal forward- or reverse-scoring.

Note of caution inserted 6 August 2015: be careful with the contents of column 6 -- there should be a "P" in this column if the items have already been scored on a right/wrong basis with corresponding scores of (0,1) as would often be the case with dichotomously-scored items. The "**IStats**" and "**IScores**" worksheets created by Lertap 5 will usually have (0,1) item scores for cognitive subtests.

More caution: Xcalibre will appear to work okay even when column 6 has "M" when it should have "P", but the results will be wrong.

Note that the contents of the XCal41ICF worksheet are mirrored in the corresponding text file created by Lertap, LrtpXcal_ICF.txt. Making changes in the worksheet will not automatically ripple through to the text file. Since it's the text file that get used by Xcalibre, you could open it, and make changes there instead of in the worksheet. Or, you could make them in the worksheet, and then save the worksheet as a tab-delimited txt file, overwriting the original txt file created by Lertap.

There's more about using Xcalibre in a later topic.

6.2 To SPSS

Note: also see this chapter from the Lertap manual. It might be somewhat more up to date; if it seems dated, have the browser refresh the page.

SPSS is a data analysis package widely used in the social sciences.

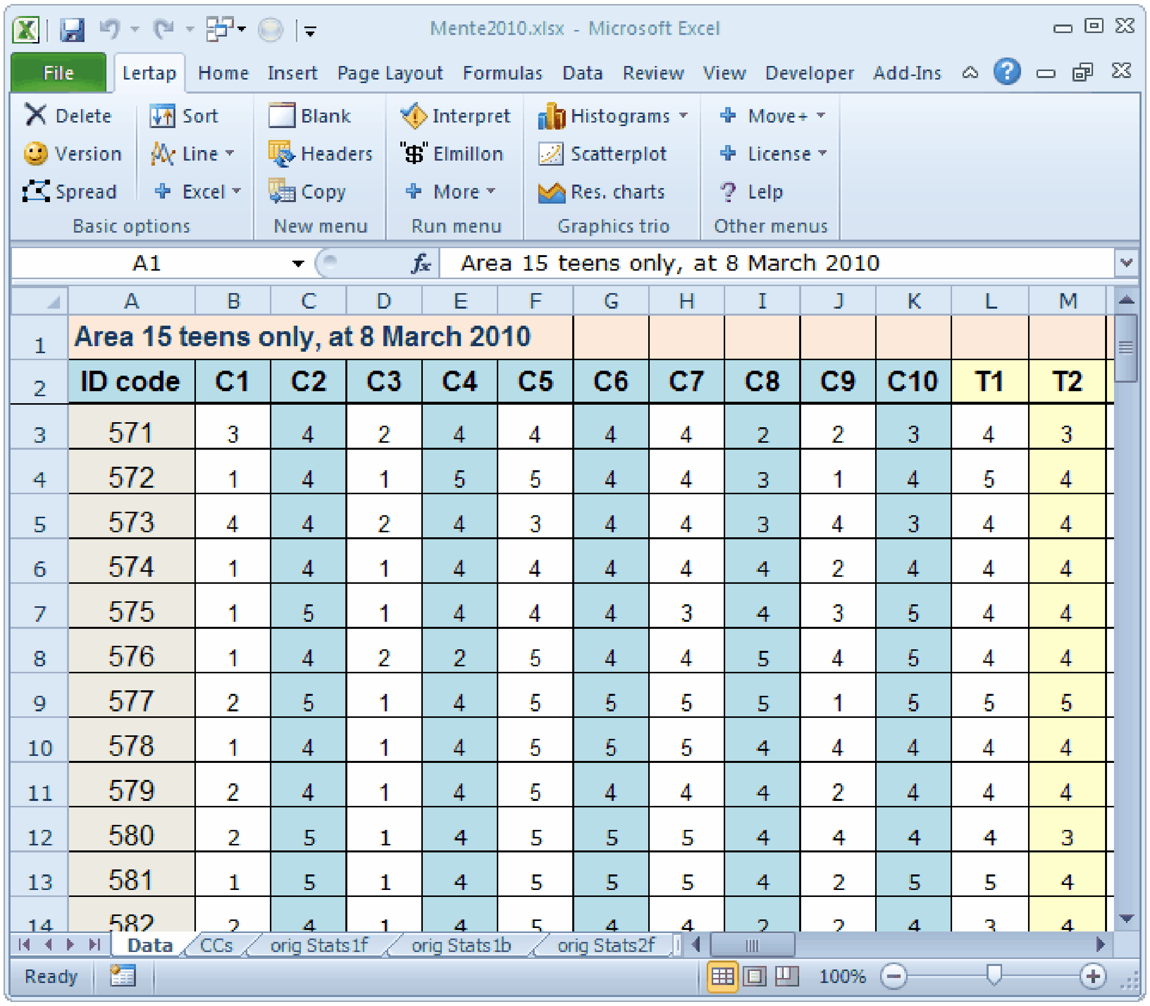
We hauled out SPSS 17 for the work described in this topic.

What we wanted to do was, first of all, copy the Data sheet from the Mente2010.xlsx workbook into an SPSS data table.

In days gone by, select, copy, and paste was a guaranteed way to get data from an Excel worksheet into an SPSS data editor table. This worked okay providing there was sufficient capacity in our computer's clipboard, which was not always the case when a large amount of data was involved.

These days SPSS has more smarts. It is capable of actually opening an Excel workbook, and grabbing data all by itself, without any need to select, copy, and paste.

To get started, look at the top of the Data worksheet in Mente2010.xlsx:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Area 15 teens only, at 8 March 2010												
2	ID code	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	T1	T2
3	571	3	4	2	4	4	4	4	2	2	3	4	3
4	572	1	4	1	5	5	4	4	3	1	4	5	4
5	573	4	4	2	4	3	4	4	3	4	3	4	4
6	574	1	4	1	4	4	4	4	4	2	4	4	4
7	575	1	5	1	4	4	4	3	4	3	5	4	4
8	576	1	4	2	2	5	4	4	5	4	5	4	4
9	577	2	5	1	4	5	5	5	5	1	5	5	5
10	578	1	4	1	4	5	5	5	4	4	4	4	4
11	579	2	4	1	4	5	4	4	4	2	4	4	4
12	580	2	5	1	4	5	5	5	4	4	4	4	3
13	581	1	5	1	4	5	5	5	4	2	5	5	4
14	582	2	4	1	4	5	4	4	2	2	4	3	4

Notice two things, please. We have the formula bar showing, and we have changed the "referencing style" so that worksheet columns use letters instead of numbers.

The formula bar is displaying the contents of cell A1 above, which is row 1, column 1, or, column A, row 1.

To see how to quickly toggle the formula bar so that it either shows or does not show, and to see how to change Excel's referencing style so that columns are letters instead of numbers, look here.

Having letters for the column "numbers" is handy when using SPSS. Extremely handy.

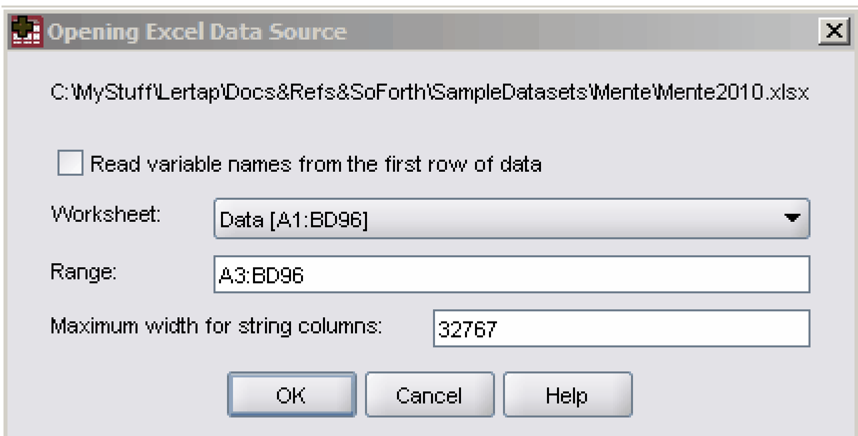
Do you have SPSS warmed up? Good. Do you have the Mente2010.xlsx workbook open in Excel? Could be Bad. You may have to close Mente2010.xlsx if you have in mind getting SPSS to read data from it; you'll know -- try the stuff below without closing Mente2010.xlsx; if SPSS complains, close Mente2101.xlsx and start again.

Now, where do the data begin in the Mente2010.xlsx Data worksheet? In cell A3. What's the last cell with data? Cell BD96. We'd like SPSS to import data from cell A3 to cell BD96. In Excel shorthand, this range is denoted as A3:BD96.

Ready. In SPSS click on File. Then on Open. Then on Data.

Pick out files of type "Excel (*.xls, *.xlsx, *.xlsm). Navigate through your computer's folders until you find Mente2010.xlsx.

A box similar to the one below should appear. Into it we have already typed the Range.



And that should do it. Click on OK, and okay is what you should be, free to SPSS-away the rest of the day.

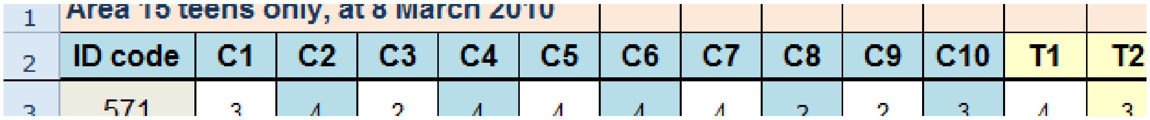
How about some factor analysis activity? Say we got an IStats report for Mente2011.xlsx. We did. It turned out to have item scores in the range B3:AT96. It's an easy job pick up these item scores in SPSS.

If desired, the interitem correlation matrix, located in the IStats range B105:AT149, could also be picked up and brought into SPSS. Easy.

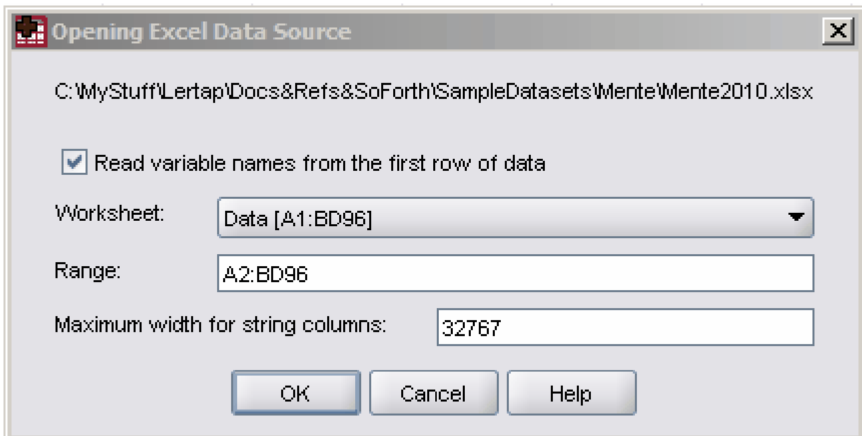
Now, what's not so nice is having to type all the "variable names" in SPSS. You'll find that SPPS defaults to variable names starting with "V1".

Wouldn't it be fine if we could get SPSS to read our Lertap column headers and use them for its variable names?

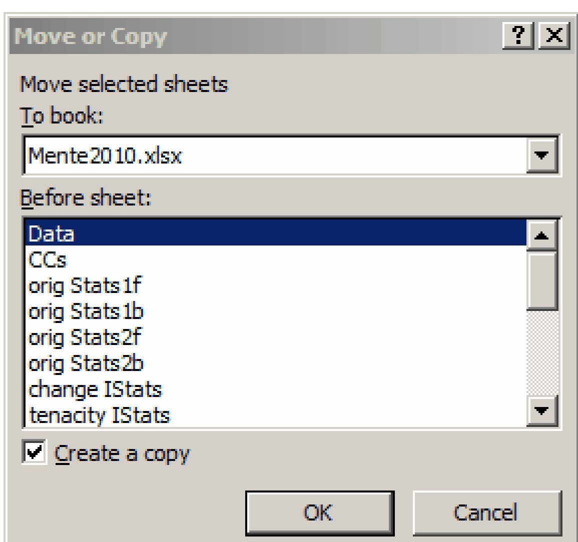
For the Mente2010.xlsx Data worksheet, the column headers start with "ID code", and, in the snippet below, go through "T2".



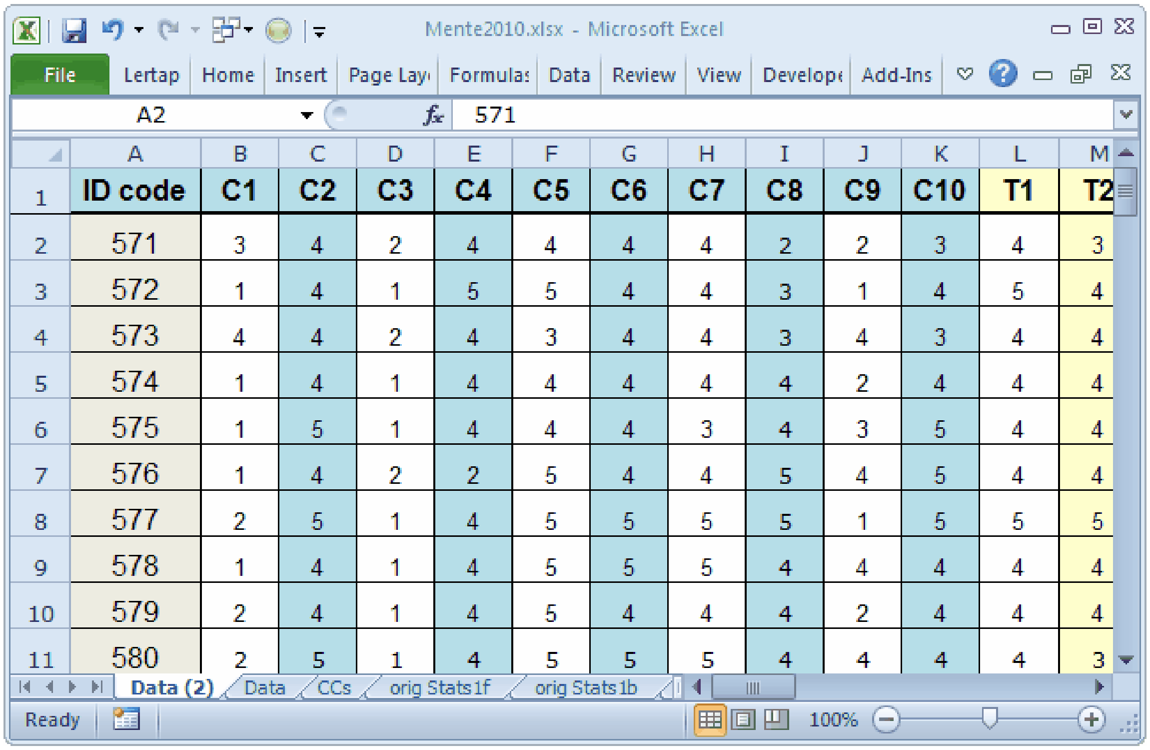
SPSS has an option to recognize and use column headers as its "variable names". We tried the following; notice the tick in the check box, and how our range now starts with A2:

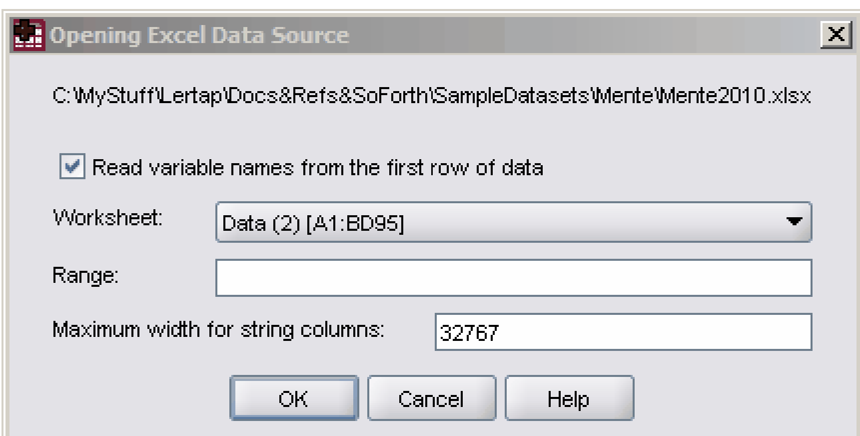


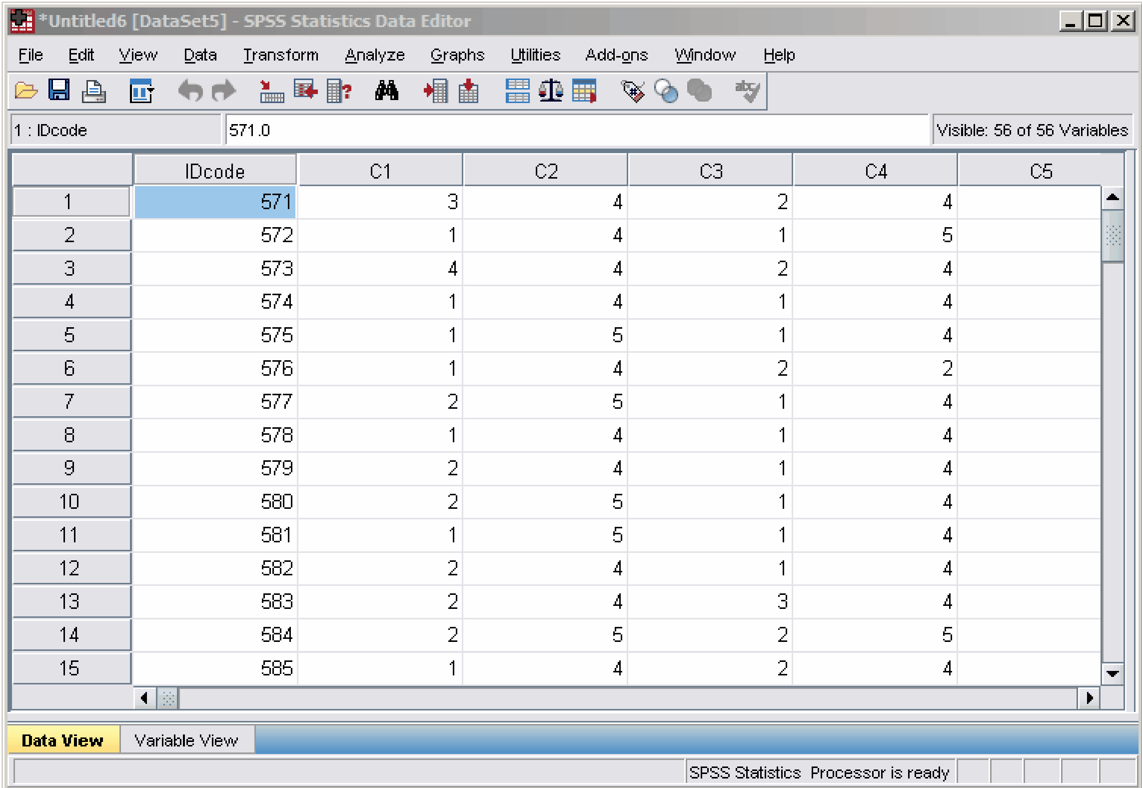
It didn't work. We've tried this a few times in recent years, and have come to settle for a work-around: we get Excel to make, for example, a copy of the Data worksheet. This is easy to do. Right-click on the Data worksheet's tab at the bottom of the screen. From the menu which drops down, select "Move or Copy ..."



Be sure to tick the "Create a copy" option. Then, in the copy of the Data sheet which results, delete the first row so that the column headers we want to have SPSS use as its variable names will be at the very top of the worksheet. The data we want to import, including the first row with its column headers, will now be in the range A1:BD95.







	IDcode	C1	C2	C3	C4	C5
1	571	3	4	2	4	
2	572	1	4	1	5	
3	573	4	4	2	4	
4	574	1	4	1	4	
5	575	1	5	1	4	
6	576	1	4	2	2	
7	577	2	5	1	4	
8	578	1	4	1	4	
9	579	2	4	1	4	
10	580	2	5	1	4	
11	581	1	5	1	4	
12	582	2	4	1	4	
13	583	2	4	3	4	
14	584	2	5	2	5	
15	585	1	4	2	4	

Bob's your uncle, mate!

6.3 Exercises

There are a number of exercises which could be thought of as activities for a rainy day. They won't hurt, and an ability to shuffle data around, between this program and that program, is something which can be handy to cultivate.

In this topic, **CTT** refers to "classical test theory", while **IRT** refers to "item response theory".

Both theories provide estimates of item difficulty and discrimination (although IRT's 1PL, or "Rasch", model does not -- it assumes all items have the same discrimination, and it consequently concerns itself only with item location, which is directly analogous to item difficulty).

Let's say that we wanted to compare the item parameter estimates from the two approaches. Yes, this has been done quite a number of times, and you might already be familiar with the literature in this regard. But pretend you're not.

What we could do is take the CTT estimates of difficulty and discrimination from Lertap, and the IRT estimates from Xcalibre, paste them into Excel columns, and then get some scatterplots.

Both programs make it a pretty simple task to do this. Lertap's Stats1b report, an Excel worksheet, has the CTT estimates. Xcalibe 4 has an option to "Save the item parameters" in a csv (comma-separated values) file, which Excel likes.

We'd need to select a dataset to work with. The larger the sample, the better. There are four datasets in this website that have a substantial number of data records: MathsQuiz (n=999), M.Nursing (n=1,768), LenguaBIg (n=5,504), and LaFlorida (n=11,109). Of these, LaFlorida is not suitable as its items were exceptionally difficult, with many students failing to answer numerous items.

Let's say we take M.Nursing. We get Lertap to produce the data matrix and item control files required by Xcalibre 4. We turn on Xcalibre's option to "Save the item parameters". We open a new Excel workbook. In column 1 we put item number. In columns 2 and 3 we paste the diff. and disc. values from Lertap's Stats1b sheet. In columns 4 and 5 will go the "a" and "b" parameter estimates from Xcalibre. *Bob's your uncle*; we're set to party.

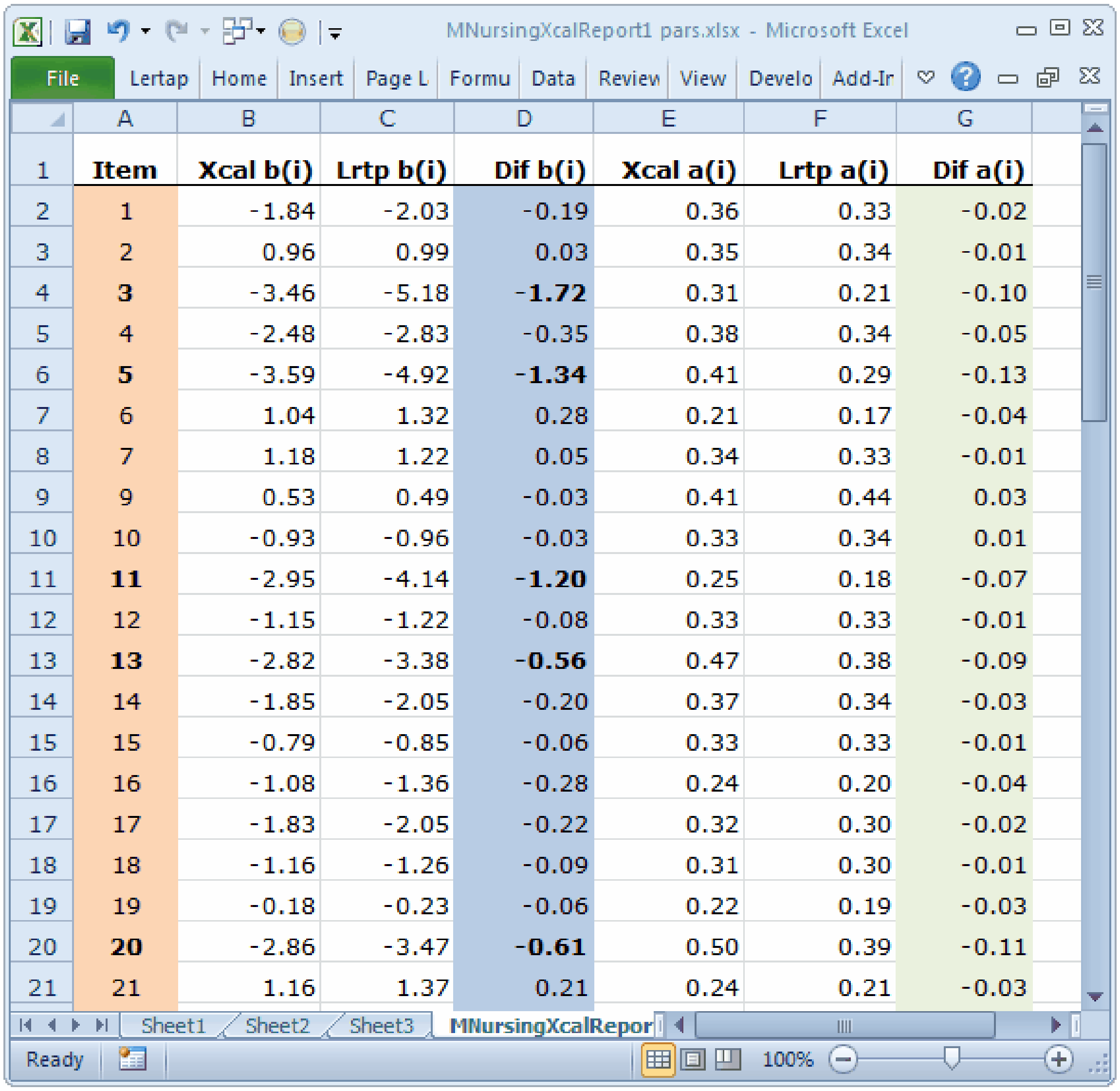
(Note: "**To halve and hold**" is an option used to create random halves from any Data worksheet. Add life to the party by splitting another dataset, LenguaBIg, into halves, and then comparing results. You could even split the halves into halves, getting quarters, and then compare results. And then, you could split the quarters ....)

Here's another example; same theme, different leg action:

Lertap has an option to produce estimates of 2PL IRT item parameters. It's activated by turning on the "experimental features" option.

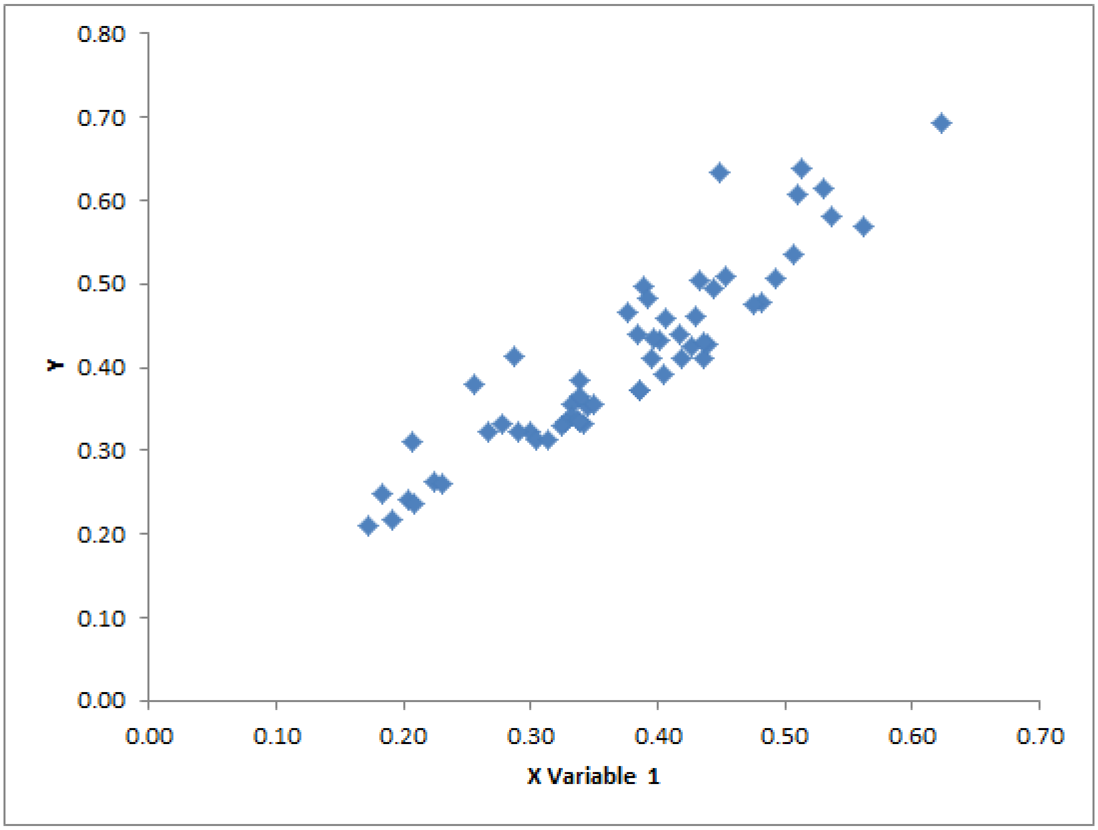
We could compare these Lertap estimates with Xcalibre's, setting the "2-parameter (a and b)" option via Xcalibre's "IRT model" tab.

Here's how our working Excel workbook looked:

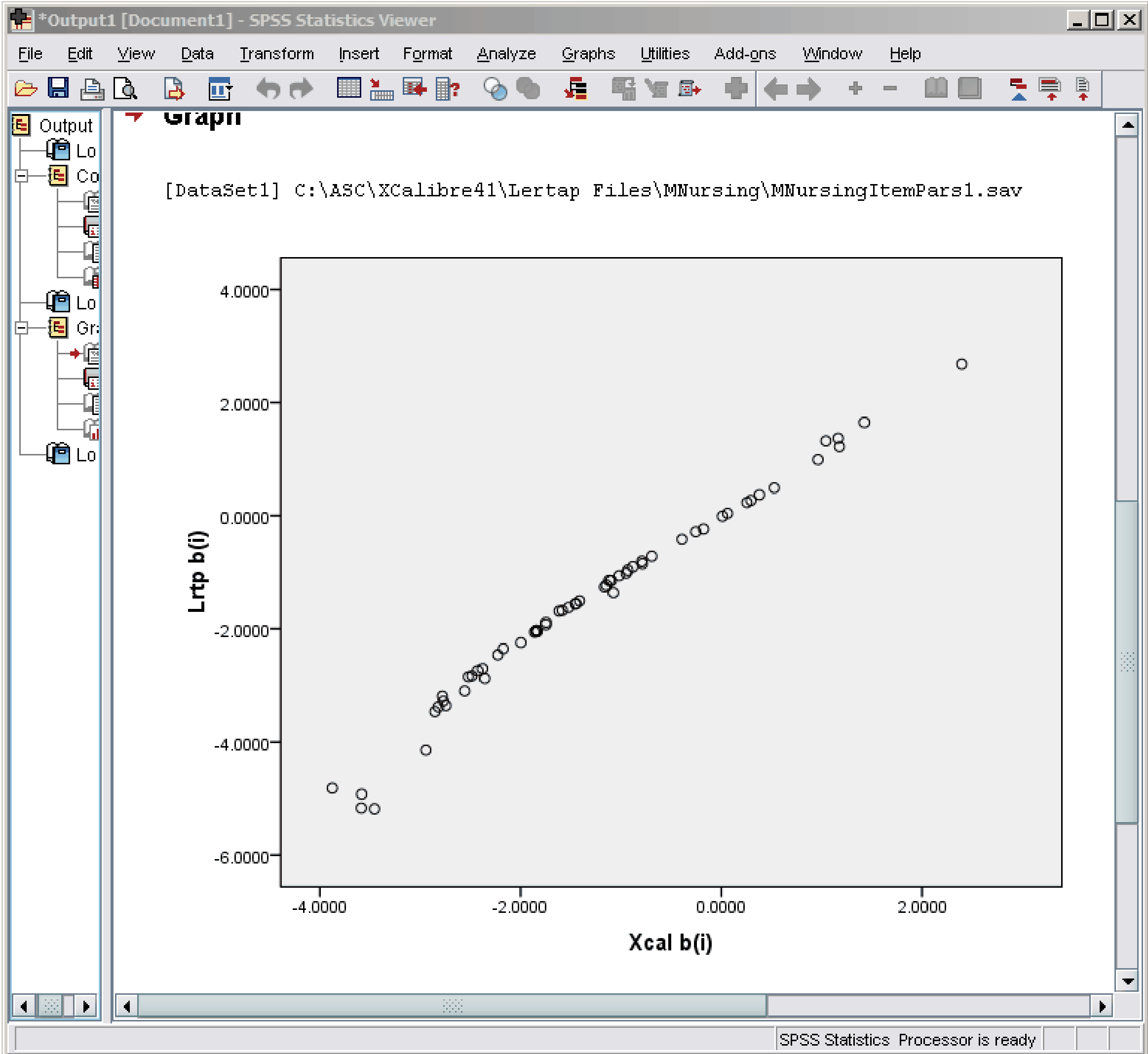


Item	Xcal b(i)	Lrtp b(i)	Dif b(i)	Xcal a(i)	Lrtp a(i)	Dif a(i)
1	-1.84	-2.03	-0.19	0.36	0.33	-0.02
2	0.96	0.99	0.03	0.35	0.34	-0.01
3	-3.46	-5.18	**-1.72**	0.31	0.21	-0.10
4	-2.48	-2.83	-0.35	0.38	0.34	-0.05
5	-3.59	-4.92	**-1.34**	0.41	0.29	-0.13
6	1.04	1.32	0.28	0.21	0.17	-0.04
7	1.18	1.22	0.05	0.34	0.33	-0.01
9	0.53	0.49	-0.03	0.41	0.44	0.03
10	-0.93	-0.96	-0.03	0.33	0.34	0.01
11	-2.95	-4.14	**-1.20**	0.25	0.18	-0.07
12	-1.15	-1.22	-0.08	0.33	0.33	-0.01
13	-2.82	-3.38	**-0.56**	0.47	0.38	-0.09
14	-1.85	-2.05	-0.20	0.37	0.34	-0.03
15	-0.79	-0.85	-0.06	0.33	0.33	-0.01
16	-1.08	-1.36	-0.28	0.24	0.20	-0.04
17	-1.83	-2.05	-0.22	0.32	0.30	-0.02
18	-1.16	-1.26	-0.09	0.31	0.30	-0.01
19	-0.18	-0.23	-0.06	0.22	0.19	-0.03
20	-2.86	-3.47	**-0.61**	0.50	0.39	-0.11
21	1.16	1.37	0.21	0.24	0.21	-0.03

An Excel-made scatterplot of the two a-parameter estimates:

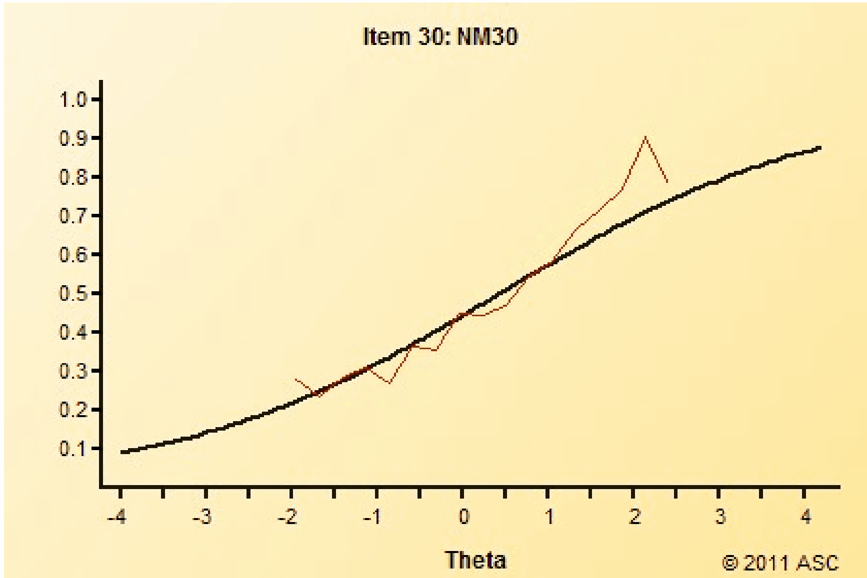


An SPSS scatterplot of the two b-parameter estimates:



What might be your guess as to the values of the correlation coefficients for these two scatterplots? Excel and SPSS output the values, but we'll leave you guessing.

Another exercise to consider: trying to discern similarities between Lertap's quintile plots, and Xcalibre's item characteristic curves. Here is just one example:



Item information

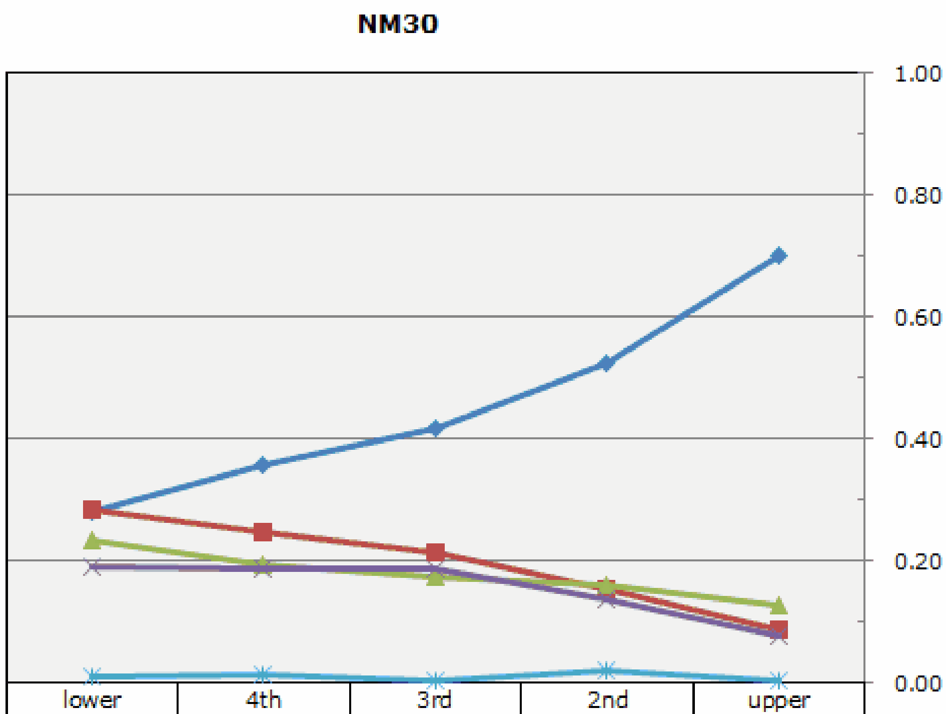
Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
30	NM30	2PL	A	Yes	4	1	

Classical statistics

N	P	S-Rpbis	T-Rpbis	Alpha w/o
1768	0.456	0.238	0.292	0.827

IRT parameters

a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
0.312	0.377	0.095	0.093	3.689	12	0.988	0.510	0.610



	lower	4th	3rd	2nd	upper
A	0.28	0.36	0.42	0.52	0.70
B	0.28	0.25	0.21	0.15	0.09
C	0.24	0.20	0.17	0.16	0.13
D	0.19	0.19	0.19	0.14	0.08
other	0.01	0.01	0.01	0.02	0.01

Key = A Diff = .46 Disc = .24

Note: this is not to suggest that Lertap is a dinkum IRT program. It is not. It does not strive to be. However, its estimates of the 2PL IRT parameters might be of use at times.

Index

- * -

- ? -

- 2 -

- A -

- B -

- C -

- D -

- E -

- F -

- G -

- H -

- I -

- R -

- S -