# Chapter 10

# Computational Methods Used in Lertap 5

Contents



---

## Overview

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

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

Understanding how Lertap works is aided by knowledge of Excel's basic structure, which consists, at the top level, of a "workbook". A workbook is a collection of "worksheets". Each worksheet is a matrix, or grid, of rows and columns. The intersection of a row and column produces a "cell".

Readers familiar with other spreadsheet programs, or with older versions of Excel, will want to take particular note of the "workbook" structure of Excel. This

structure was first found in the version of Excel known as "Excel 95". It permits multiple spreadsheets, or "worksheets" as Excel calls them, to exist as a bound collection within a single file. Under Windows, this file usually has an extension of "xls". These xls files are referred to as workbooks; each workbook may contain from one to hundreds of individual worksheets.

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

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

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

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

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

It is generally the case that users need only save the two primary sheets, Data and CCs, in their workbook. This is so as the other sheets, the so-called secondary ones, are derived from the primary ones. Any or all of the secondary sheets may be deleted, and then re-generated from the primary ones. Of course, there is no harm, none whatsoever, in having workbooks with many worksheets--there is no need to delete the secondary sheets at all. However, users will sometimes want to share their workbooks with colleagues, and may want to do so by sending them over the Internet. In cases such as this it is often useful to pare back on the number of worksheets. (Note that it is easy to have Excel extract a copy of any worksheet in a workbook, sending the copy to a new xls file. See Excel's on-line help for instructions.)

## The Lertap5.xls workbook

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

When Lertap5.xls is opened, all of its code modules are made available (exposed) to all other open Excel workbooks. Access to these modules is via the Lertap5 toolbar, a toolbar having a smiley yellow face towards its left side, and an 8-ball towards its right side. As of October 2000, Lertap's standard toolbar looked like this[1]:


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

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

## Interpret CCs lines (the Sub worksheets)

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

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

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

The Sub worksheets are usually hidden by Lertap. They may be unhidden easily—here again Excel's on-line help manual will be of assistance. It may be instructive to users to examine the contents of a Sub file as they contain the fundamental subtest and item data used in Lertap analyses. In fact, after the option to "Interpret CCs lines" has once been taken, the contents of a workbook's CCs worksheet are never again referred to, that is, not until the "Interpret CCs lines" option is taken again.

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

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

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

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

## The Freqs worksheet

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

*col (c3, c5, c10)

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

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

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

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

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

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

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

When the Freqs listing indicates the presence of unexpected responses, Excel's data filter capability may be used to quickly locate the Data records containing

the responses. This is a powerful utility, easy to use (after some practice)--refer to Excel's on-line help manual for instructions.

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

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

## Elmillon item analysis

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

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

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

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

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

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

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

The little "f", "b", and "ul" letters at the end of the names of these new worksheets signify "full", "brief", and "upper-lower". The full worksheets have

very detailed item and subtest performance data, while the brief worksheets have concise item performance summaries. The full worksheets have report formats which will be familiar to users of Lertap 2 and Lertap 3. The brief reports are new to this version, as are the upper-lower ones.

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

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

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

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

## Stats1f for cognitive subtests

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

**Item 7**

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

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

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

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

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

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

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

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

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

Above, options A and E both have non-zero weights, and the number of people who got some points for their answer to this item is 51 (being 20 plus 31). The total number of respondents is 60, so the item's difficulty would be 51/60, or 0.85[2].

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

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

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

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

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

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

each item. Each item's "score", or points, forms part of the subtest score, and this will serve to inflate the value of the correlation coefficient.

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

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

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

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

Methods for determining the biserial correlation coefficient may be seen in Allen and Yen (1979, p.39); in editions of Glass & Stanley (1970, p.171, eq. 9.11; 1974, p.171, eq. 9.11); in Gulliksen (1950, p.426); and in Magnusson (1967, p.205, eq. 14-7). Lertap uses the equation seen in Glass & Stanley. Values of the normal density function required in the equation are based on an algorithm known as "NDTRI", found in the IBM Scientific Subroutine Package[3].

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

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

One could make the point that the part-whole inflation issue applies to the "avg." value as much as it does to the pb(r) values. That is, in the case of internal criteria, the "avg." value is inflated if a response has a "wt." greater than zero. While this is so, Lertap does <u>not</u> apply a correction to the "avg." and "z" values.

Lertap allows cognitive items to have more than one correct answer. More precisely, Lertap allows any response to have any weight (wt.). The correction for inflation will, however, be applied to only one response, the one having the

[3] See http://pdp-10.trailing-edge.com/www/lib10/0145/

greatest positive weight. If two or more responses share the greatest positive weight, the correction is applied to the response which comes last when Elmillon outputs its results to the Stats1f worksheet.

Above, it was mentioned that Elmillon calculates statistics at three levels: response, item, and subtest. For all items, cognitive and affective, the statistics computed at the item level include the item mean, and the item's product-moment correlation with the criterion score, corrected for part-whole inflation when the criterion is internal, which is the usual case.

In the case of cognitive items, the normal situation is for each item to have one correct answer, with a weight of one point. When this is true, the p value shown for the correct response will equal the item's mean, and the pb(r) value will equal the item's product-moment correlation.

The subtest statistics found in sheets such as Stats1f have the following format:

Summary statistics

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

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

The number of scores is usually equal to the number of records in the Data worksheet. However, if a Data record is missing data for all of the items belonging to a subtest, that record is omitted from calculations.

The first standard deviation value reported by Elmillon is the "population" standard deviation, computed by using "n" in the denominator of the relevant equation. The "sample" standard deviation and variance are computed using "n-1" in the denominator. Lertap commonly uses population variance and standard deviation values in its calculations. For example, it uses the population standard deviation for z-scores.

For a discussion of variance and standard deviation calculations, see Glass & Stanley (1970, p.82 & 1974, p.82), Hays (1973, p.238), Hopkins & Glass (1978, p.78), or Magnusson (1967, p.8).

The figures shown in parentheses are calculated using the maximum possible score. Above, for example, the average subtest score, that is, the mean subtest

score, was found to be 12.63, which is 50.5% of the maximum possible score, 25.

The reliability figure reported by Elmillon is Cronbach's coefficient alpha. Lertap 2 used Hoyt's analysis of variance procedure to derive an estimate of reliability. As is now well known, Hoyt's procedure produces the same value as alpha. Lertap 3 used alpha. For a discussion of coefficient alpha, and its interpretation, Pedhazur & Schmelikn (1991, pp.92-100) have a particularly thorough presentation. It has long been known that coefficient alpha is not an index of the factorial complexity of a test; high values of alpha cannot be interpreted as meaning that a test is measuring just one factor, or that the test's items are necessarily highly homogeneous.

The index of reliability is the square root of the alpha value, while the standard error of measurement equals the standard deviation times the square root of the quantity (1-alpha).

For further discussion of the calculation and interpretation of these statistics, see Ebel & Frisbie (1986), Hopkins, Stanley, & Hopkins (1990), Linn & Gronlund (1995), Magnusson (1967), Mehrens & Lehmann (1991), Oosterhof (1990), and Pedhazur & Schmelikn (1991).

After Stats1f's subtest statistics, Elmillon summarises item difficulties using a series of 10 bands.

**item difficulty bands**

**.00:** Item 22
**.10:**
**.20:**
**.30:**
**.40:** Item 1 Item 2 Item 9 Item 11 Item 14 Item 18 Item 19 Item 20 Item 21 Item 25
**.50:** Item 3 Item 4 Item 6 Item 7 Item 10 Item 12 Item 15 Item 17 Item 24
**.60:** Item 8 Item 13 Item 16 Item 23
**.70:** Item 5
**.80:**
**.90:**

Above we see that Item 22's difficulty fell into the first band, labelled ".00". This means that its difficulty was below .10. To see what the actual value was one may either scroll up in the Stats1f worksheet to see the item's response statistics, or look up its value in the Stats1b worksheet.

After the item difficulty bands, Elmillon creates a similar display for item discrimination values:

**item discrimination bands**

**.00:**
**.10:**
**.20:** Item 4 Item 22
**.30:** Item 5 Item 14 Item 24
**.40:** Item 7 Item 9 Item 16 Item 23
**.50:** Item 3 Item 10 Item 12 Item 15 Item 17
**.60:** Item 1 Item 2 Item 6 Item 8 Item 11 Item 18 Item 21 Item 25
**.70:** Item 13 Item 19 Item 20
**.80:**
**.90:**

It may be seen, above, that Item 7's discrimination value falls in the .40 band, meaning that it was equal to or greater than .40, but less than .50.

Here Elmillon is reporting item-level data. Item discrimination is now the item's product-moment correlation with the criterion.

For the usual case of a cognitive item with one correct answer, and a weight of one point, an item's mean will be equal to p, its difficulty, and its product-moment correlation will equal its pb(r) value for the correct answer.

The item difficulty and discrimination bands did not feature in Lertap 2; they were introduced in Lertap 3.

New to Lertap 5 is a display of adjusted alpha reliability values. This appears at the very end of the Stats1f worksheet, and has a format such as that seen below (note that to save space the display below has been truncated after the tenth item; there were 25 items in the complete display).

**alpha figures (alpha = .9149)**

| *without* | *alpha* | *change* |
|---|---|---|
| Item 1 | 0.909 | -0.006 |
| Item 2 | 0.909 | -0.006 |
| Item 3 | 0.911 | -0.003 |
| Item 4 | 0.917 | 0.002 |
| Item 5 | 0.915 | 0.000 |
| Item 6 | 0.910 | -0.005 |
| Item 7 | 0.914 | -0.001 |
| Item 8 | 0.910 | -0.005 |
| Item 9 | 0.914 | -0.001 |
| Item 10 | 0.911 | -0.003 |

The display above begins by repeating the subtest's overall alpha value, as reported earlier under summary statistics. It then indicates what the value of alpha would be if each item were omitted from the subtest. For example, if Item 1 were removed from the subtest (for some reason), the alpha figure would decrease to 0.909, a change of –0.006 from the original value of alpha.

## Correction for chance scoring

It is possible to correct cognitive test scores for the possible effects of guessing by using the "CFC" control word on a *sub card.

The correction is based on what the literature often refers to as the "standard correction". Item weights are changed so that distractors have a weight equal to minus one (-1.00) divided by the total number of distractors used by the item.

For discussions on the pros and cons of using the correction for chance formula, or, more generally, "formula scoring", see Ebel & Frisbie (1986); Hopkins (1998); Linn & Gronlund (1995); Mehrens & Lehmann (1991); Oosterhof (1990); and Wiersma & Jurs (1990).

Assessing the impact of applying Lertap's CFC scoring may be accomplished by duplicating a cognitive subtest's control cards, and using the CFC control word on one of the *sub cards. The following cards indicate how this might be done:

```
*col (c3-c27)
*sub Res=(A,B,C,D,E,F), Name=(Knowledge), Title=(Knwlg)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
*col (c3-c27)
*sub Res=(A,B,C,D,E,F), CFC, Name=(Knowledge CFC), Title=(KnwlgCFC)
*key AECAB BEBBD ADBAB BCCCB BABDC
*alt 35423 35464 54324 43344 45546
```

The two subtests referred to by these cards are identical, except that the second *sub card carries the CFC control word. The ramifications of using CFC may be seen by looking at the correlation between `Knwlg` and `KnwlgCFC`, by having Lertap make a scatterplot of these two scores, and by comparing sorted listings of the scores.

## Stats1b for cognitive subtests

It has been mentioned that Elmillon, Lertap's item analysis program, adds two worksheets for each Sub worksheet found in a workbook. The first of these is called Stats1f; snapshots from a typical Stats1f sheet are shown above.

New in Lertap 5 is a condensed summary of item-level statistics. This is found in the sheet called Stats1b; a sample from a typical Stats1b sheet for a cognitive subtest is shown below.

Lertap5 brief item stats for "Knowledge of LERTAP2", created: 23/08/00.

| Res = | **A** | **B** | **C** | **D** | **E** | **F** | other | diff. | disc. | ? |
|---|---|---|---|---|---|---|---|---|---|---|
| **Item 1** | 43% | 42% | 15% | | | | | 0.43 | 0.66 | |
| **Item 2** | 7% | 20% | 12% | 13% | 48% | | | 0.48 | 0.66 | |
| **Item 3** | 3% | 2% | 53% | 42% | | | | 0.53 | 0.54 | |
| **Item 4** | 55% | 45% | | | | | | 0.55 | 0.23 | |
| **Item 5** | 22% | 70% | 8% | | | | | 0.70 | 0.33 | |
| **Item 6** | 27% | 50% | 23% | | | | | 0.50 | 0.62 | |
| **Item 7** | 33% | 2% | 12% | | 52% | | 2% | 0.52 | 0.40 | D |
| **Item 8** | 2% | 63% | 35% | | | | | 0.63 | 0.61 | D |
| **Item 9** | 15% | 43% | 8% | 7% | 8% | 13% | 5% | 0.43 | 0.40 | |
| **Item 10** | 17% | 10% | 12% | 53% | | | 8% | 0.53 | 0.54 | |

The Stats1b display tries to fit as much item performance data as possible on a single line. It begins with a header line showing possible item responses; above these show as A B C D E and F.

Following the header row, a mixture of response and item-level statistics are given for each item. Above it may be seen that, for Item 7, 33% of the respondents selected option A, corresponding to the same response's p value of 0.33 found in the Stats1f worksheet. The underlining indicates that option A had a response weight greater than zero. (Option A was the correct answer to this item.)

Item 7's "diff." index shows above as 0.52. This value will equal the p value found in the full statistics sheet, Stats1f, if the item has only one correct answer, with a weight of one point, which is the usual case for cognitive items.

Item 7's "disc." index is an item-level statistic, equal to the product-moment correlation of the item with the criterion, corrected for part-whole inflation (discussed above). The disc. value will equal the Stats1f pb(r) value if the item has only one correct answer.

The last column in the Stats1b display is headed with ?, a question mark. If an item has an entry in this column, Lertap is pointing out that the item has one or more distractors which *may* be functioning in a questionable manner. Above we see, for example, that Lertap has flagged one of Item 7's distractors, D. No-one selected this distractor. This is usually an undesired outcome for a distractor—the role of a distractor is to serve as a foil, a decoy which will appear attractive to

weaker respondents. To Lertap, weaker respondents are those with criterion scores below the criterion mean.

To repeat, Lertap regards a distractor as effective if it is selected by respondents, and if these respondents have a below-average score on the criterion. Distractors which are not selected by anyone, and distractors which are selected by respondents whose average criterion score equals or exceeds the criterion mean, are candidates for the ? column.

This summary of distractor performance had no equivalent in Lertap 2. Lertap 3 was the first version to have a distractor adequacy index, reporting the percentage of distractors which were selected by weaker respondents.

The Stats1b report is obviously a brief one. The most complete information about item functioning is given in the Stats1f worksheet.

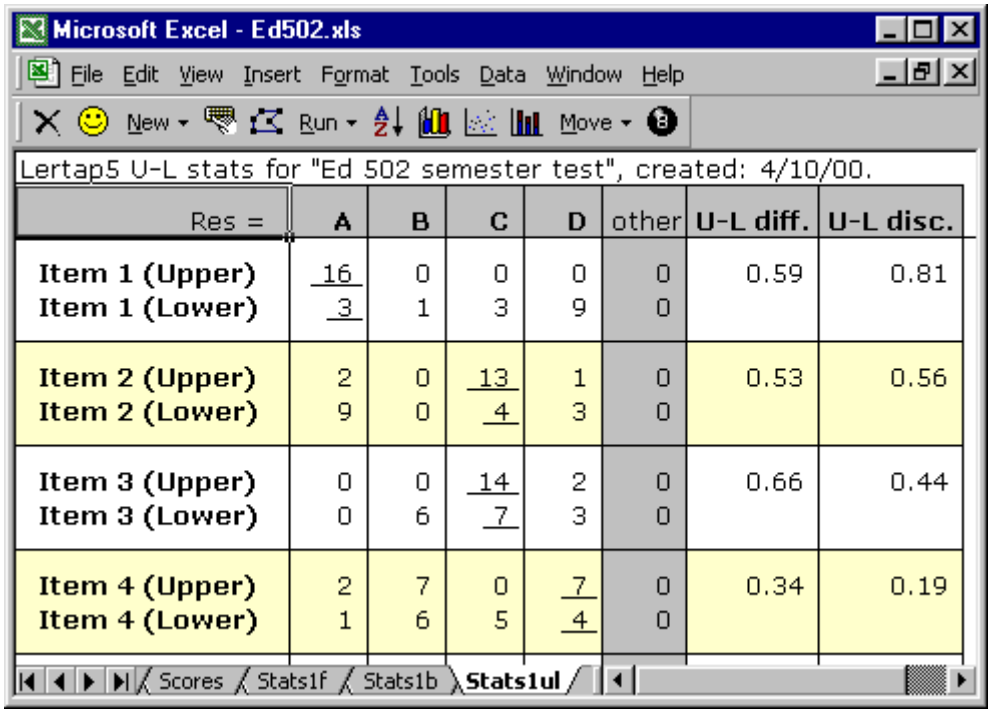
## Stats1ul for cognitive subtests

The Stats1b worksheet is new to this version of Lertap, as is another, one which is added only for cognitive subtests: Stats1ul.

The item discrimination statistics seen in both the Stats1f and Stats1b sheets are based on correlation coefficients. There is another way of indexing item discrimination, one which was originally advocated before the birth of desktop computers, and a method which remains popular with many users. It's called the "upper-lower" (U-L) method.

An advantage of the U-L approach is its conceptual simplicity. Test results are used to define two groups—the "upper" group, with high test scores, and the "lower" group, with low scores. If an item, a cognitive item, is discriminating, it should be the case that the upper group is more successful in identifying the correct answer, and less prone to fall for the distractors than is the lower group.

Lertap's U-L stats sheet, Stats1ul, has two sections. The first section summarises item results in the format seen here:



Microsoft Excel - Ed502.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 U-L stats for "Ed 502 semester test", created: 4/10/00.

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

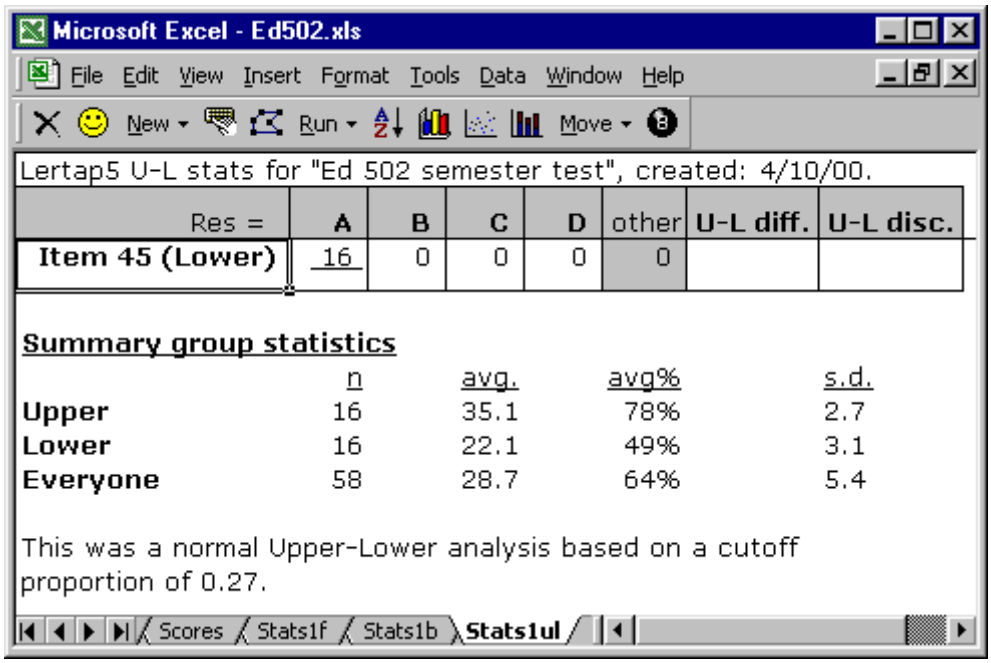
Scores Stats1f Stats1b **Stats1ul**

The U-L item stats show how many people in each group selected each item option. The keyed-correct answer is underlined (all responses with weights greater than zero are underlined—usually there's only one such).

The U-L diff. index is the number of people in both groups who selected the right answer, divided by the sum of the number of people in both groups. In this example, 19 respondents got the Item 1 right, and there was a total of 32 respondents, 16 in the upper group, and 16 in the lower.

The U-L disc. is based on computing two difficulty figures, one for each group. In this example, the difficulty of Item 1 in the upper group was 16/16, or 1.00. In the lower group, the difficulty was 3/16, or 0.19. The U-L disc. is the difference between these two, 1.00 – 0.19, or 0.81.

The lower section of the Stats1ul sheet looks like this:




The statistics shown in this summary section are based on results from the top 27% of the overall group, and the bottom 27%. In this example, the whole group, "Everyone", consisted of 58 people—27% of 58, rounded to the nearest integer, is 16, the "sample size" of each group.

The usual Lertap procedure is to use an internal criterion, the subtest score, to define the groups. To do this, Lertap scores each person's responses, forms an array in memory with copies of the scores, then sorts the scores and picks off the top and bottom groups, writing their results to a temporary, hidden, worksheet called "ScratchScores". It is possible to use an external criterion to define the groups, as mentioned later in this chapter.

The avg. shown in the table is the mean of the scores in each group. The avg.% figure is avg. divided by the maximum possible score. The s.d., standard deviation, is computed using the "population" equation (see the discussion on standard deviation calculations mentioned earlier).

Why are the groups based on 27%? It's a standard, often-recommended figure. According to Hopkins, Stanley, & Hopkins (1990, footnote on p.269), the reason for selecting 27% can be traced to Kelly (1939), who presented a case for this being the optimal value for defining the groups. Garrett (1952, footnote on p.215) wrote "*....There are good reasons for choosing 27%. When the distribution of ability is normal, the sharpest discrimination between extreme groups is obtained when item analysis is based upon the highest and lowest 27 per cent in each case....*".

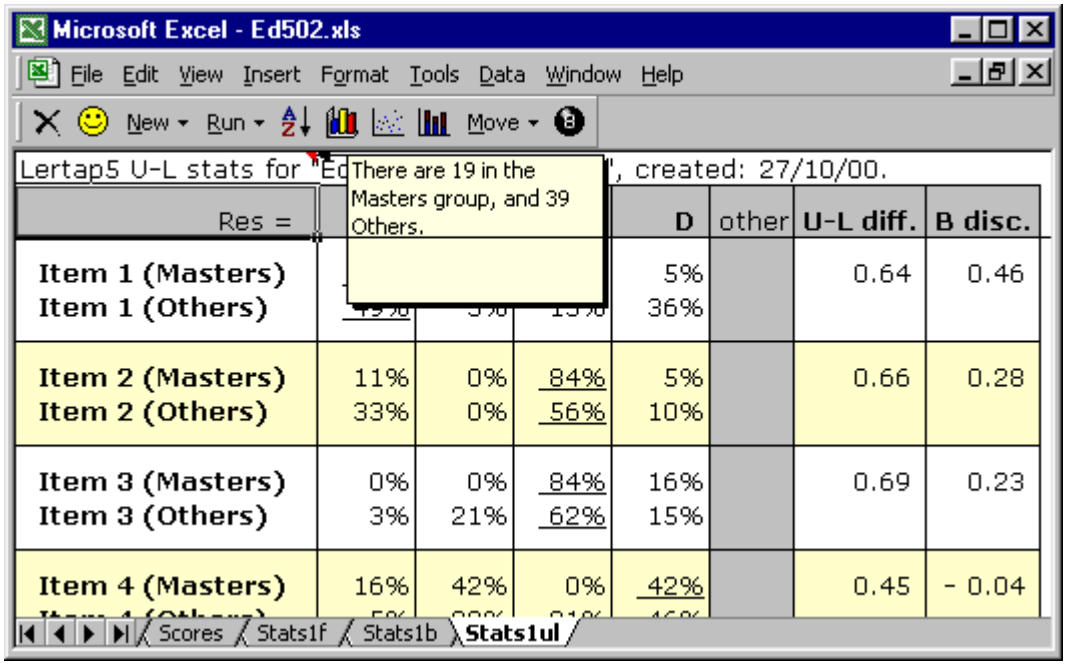
Many authors suggest that departures from 27% will not have adverse effects. Linn & Gronlund (1995), and Garrett (1952), write that 25% would be fine. Ebel & Frisbie (1986) suggest that 25% or even 33% would be workable, but then clearly state (p.229) that "*....The optimum value is 27 percent....*".

It is possible to have Lertap use something other than 27%, if wanted. The "System" worksheet holds crucial operational settings such as this. It is even possible to have the U-L analysis turned off altogether; this is also managed by changing one of the settings in the System sheet. Refer to a following section on how to access this worksheet.

**Mastery and criterion-reference testing**

Chapter 5 provides comments on the use of the Mastery and Mastery= control words. These words are used to alter Lertap's U-L analysis so that the groups are defined by reference to a criterion level.

When the Mastery word is used, the upper group will be those who have equalled or bettered the criterion level, and the lower group will be all others. Lertap's default mastery level is 70% of the criterion score. This level is easily changed by using, for example, Mastery=80, which changes the level to 80%.




The screen capture above displays Lertap's standard "mastery" analysis format[4]. The U-L diff. index is calculated as for a normal U-L analysis. The U-L disc. is now called the B disc. index after Brennan (1972). The B disc. index is formed in a manner identical to that used to determine U-L disc: from two difficulty values. In this example, the difficulty of Item 2 in the Mastery group was 0.84, while in the other group it was 0.56. Subtracting the lower group's difficulty from the upper group's, 0.84 – 0.56, gives a B disc. value of 0.28.

[4] The box with group sizes is a comment attached to cell R1C1, flagged by a small red triangle. It displays whenever the mouse pointer moves into the cell.

The report of U-L diff. and B disc. indices is followed by small tables of additional statistics, as shown below:

**Summary group statistics**

| | n | avg. | avg% | s.d. |
|---|---|---|---|---|
| **Masters** | 19 | 34.6 | 77% | 2.7 |
| **Others** | 39 | 25.8 | 57% | 3.9 |
| **Everyone** | 58 | 28.7 | 64% | 5.4 |

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

**Variance components**

| | df | SS | MS |
|---|---|---|---|
| **Persons** | 57 | 37.91 | 0.67 |
| **Items** | 44 | 118.69 | 2.70 |
| **Error** | 2508 | 446.24 | 0.18 |

| | |
|---|---|
| **Index of dependability:** | 0.732 |
| **Estimated error variance:** | 0.005 |
| **For 68% conf. intrvl. use:** | 0.070 |
| **Prop. consistent placings:** | 0.783 |
| **Prop. beyond chance:** | 0.514 |

The "Summary group statistics" are similar to those presented in the normal, non-mastery analysis. The "Variance components" section is based on the work of Brennan & Kane (1977, 1984), who suggested that an items-by-persons analysis of variance could be used to estimate errors associated with domain-referenced measurement. Lertap's "Error" variance component is Brennan & Kane's "interaction" component (a more accurate term might be "interaction and error" since both interaction and error variance are involved, and inseparable).

The "Index of dependability" is M(C) in Brennan & Kane (1977). It may be interpreted as a reliability coefficient—it has the same zero-to-one range of a classical reliability index, with values closer to one signifying less error in the measurement process. In Brennan & Kane's approach, error variance has two main components: the standard error of measurement from classical test theory (shown in Lertap's full statistics report), and a component due to item sampling. It is this latter component which distinguishes Brennan & Kane's method from classical test theory, and makes their 68% confidence interval larger than that found in the classical case (for the example above, the classical standard error of measurement, as a proportion, was .062).

The two last lines in the section above are based on the work and recommendations of Subkoviak (1976, 1984). The "Prop. consistent placings" is the statistic commonly referred to in the literature as $\hat{p}_0$, an estimate of the proportion of test takers who have been correctly classified as either master or nonmaster. Lertap follows the Peng and Subkoviak approximation method to derive the estimate, as seen in Subkoviak (1984, pp.275-276). Algorithm 462 from Collected Algorithms of the CACM[5] is employed to obtain bivariate normal probability values, with coefficient alpha passed as the correlation between the two normal variates.

[5] Communications of the ACM.

Since it will always be expected that some will be correctly placed by chance alone, the second statistic, "Prop. beyond chance", estimates the proportion correctly classified as a result of the testing process itself. This statistic is known as kappa in the literature.

## Stats1f for affective subtests

When it goes about its calculations, Lertap does not make major distinctions between cognitive and affective subtests. In fact it uses the same methods for deriving statistics at all three levels: response, item, and subtest. In the Stats1f worksheet, differences arise in the application of the correction for part-whole inflation, and in the display of what are called "mean/max" bands instead of item difficulty bands.

Consider the following Stats1f display for an item from a 10-item affective subtest:

**Item 30**

| option | wt. | n | % | pb(r) | avg. | z |
|---|---|---|---|---|---|---|
| 1 | 5.00 | 9 | 15.0 | 0.26 | 39.8 | 1.15 |
| 2 | 4.00 | 20 | 33.3 | 0.24 | 36.1 | 0.34 |
| 3 | 3.00 | 17 | 28.3 | -0.12 | 33.6 | -0.19 |
| 4 | 2.00 | 8 | 13.3 | -0.37 | 30.1 | -0.95 |
| 5 | 1.00 | 5 | 8.3 | -0.31 | 29.8 | -1.02 |
| other | 3.00 | 1 | 1.7 | -0.15 | 29.0 | -1.19 |

The results above use a display format essentially identical to that found in Lertap's display of response-level data for cognitive items (see, for example, the results presented earlier for "Item 7"). The "p" column from the cognitive case has been replaced by "%", but otherwise the columns are the same.

The statistical procedures used to derive the values given in the columns are also the same, with one exception: at this level Lertap does not correct any of the pb(r) values for part-whole inflation.

Notice the "other" line above, and, in particular, its weight of 3.00. How did the weight of 3.00 get there?

For affective items, Lertap automatically derives and applies a missing data weight unless told not to. The weight will be equal to what Lertap considers to be the "mid-scale" value. This item portrayed above has weights which range from 5 to 4 to 3 to 2 to 1. The mid-scale weight is 3, and this is what Lertap will apply to every respondent whose answer to this item was "other", that is, not one of the response codes seen under the option column.

Had the response weights ranged from 1 to 2 to 3 to 4 to 5 to 6, Lertap's weight for "other" responses would be 3.5.

How to defeat this? Use the MDO control word on the subtest's *sub card in the CCs worksheet. Note: this missing data weight applies only to affective subtests, not to cognitive ones. Lertap 2 and Lertap 3 users will want to note that the MDO control word works opposite to what they're used to. In previous versions the missing data weight was derived and applied only when MDO was specifically mentioned; now it's always present, and MDO must be used to turn it off.

The **subtest statistics**, or Summary statistics, for affective subtests are the same as those found in the Stats1f sheets made for cognitive subtests (see above).

A difference in the contents of Stats1f for the affective case is that the item difficulty bands seen in cognitive reports do not appear. The concept of "item difficulty" is not used in the affective case. Instead, Lertap derives what it calls mean/max figures, and summarises these in a series of bands.

**mean/max bands**

**.00:**
**.10:**
**.20:**
**.30:**
**.40:** Item 34
**.50:**
**.60:** Item 26 Item 27 Item 30 Item 35
**.70:** Item 28 Item 29 Item 33
**.80:** Item 31 Item 32
**.90:**

An item's mean/max figure is calculated by dividing the item mean by the maximum item weight. For example, Item 30's mean was 3.33; its maximum weight was 5; 3.33 over 5 gives 0.67. Consequently, Item 30 appears on the mean/max band of .60, meaning that its mean/max figure was equal to or greater than .60, but less than .70. Where is an affective item's mean reported? Not in the Stats1f sheet—it will be found in Stats1b.

The mean/max bands are particularly useful in the case of Likert items, where a scale of "strongly disagree" to "strongly agree" is deployed. An item whose mean/max figure falls into the .80 or .90 band is one where respondents reported strong agreement, assuming the strongly agree response to be the one with the maximum weight, of course.

The **correlation bands** seen for affective items indicate where each item's product-moment correlation with the criterion lies. Negative correlation values will map to the first band. Where is an affective item's correlation found? In the Stats1b worksheet (see below).

The **alpha figures** for affective subtests follow the pattern, and method, used in the cognitive case, and are interpreted in the same way.

## Stats1b for affective subtests

The differences between Stats1f sheets for cognitive and affective subtests may be slight, but this is not so much the case in the Stats1b worksheet.

Lertap5 brief item stats for "Comfort with using LERTAP2", created: 23/08/00.

| Res = | **1** | **2** | **3** | **4** | **5** | other | pol. | mean | s.d. | cor. |
|---|---|---|---|---|---|---|---|---|---|---|
| **Item 26** | 13% | 22% | 25% | 23% | 17% | | + | 3.08 | 1.28 | 0.76 |
| **Item 27** | 5% | 23% | 37% | 35% | | | - | 2.98 | 0.88 | 0.55 |
| **Item 28** | 22% | 45% | 17% | 13% | | 3% | - | 3.75 | 0.94 | - 0.14 |
| **Item 29** | 32% | 35% | 25% | 5% | | 3% | - | 3.93 | 0.89 | 0.44 |
| **Item 30** | 15% | 33% | 28% | 13% | 8% | 2% | - | 3.33 | 1.14 | 0.49 |

The differences in the Stats1b sheet are seen in the right-most columns. The column labelled "pol." indicates the scoring polarity of each item, that is, whether or not it was reversed. Items which have been reversed are denoted by a minus sign. Item 30 is one such item; when response-level statistics for this item were reported earlier, its weights could be seen running from high to low down the response options.

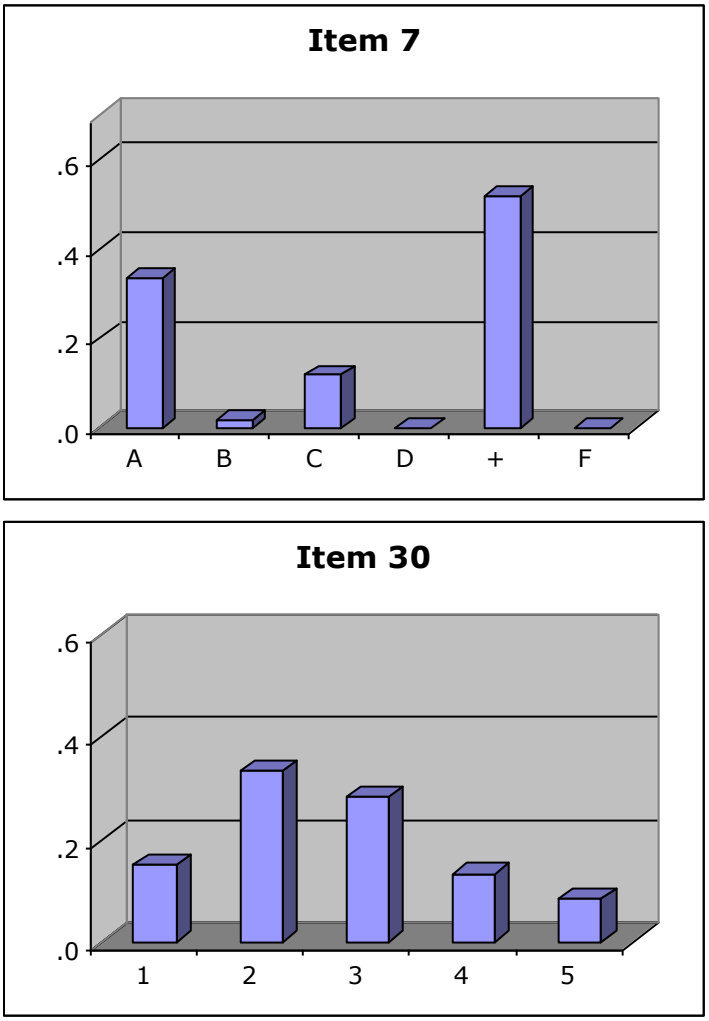
Item mean is what would be expected, that is, the average item score over all respondents, where an item's "score" is its weight—a respondent who chose option 2 on Item 30, for example, got a "score" of 4 on this item.

Item standard deviation is a population value, computed using "n" in the denominator of the appropriate equation (as discussed above).

Item correlation is the Pearson product-moment correlation between the item score and the criterion score, corrected for part-whole inflation.

## Item response charts

It is possible to have Lertap make charts from a brief stats sheet, such as Stats1b. The charts option is activated via one of the icons on the Lertap toolbar; it produces displays such as the following:




When charts are made, they generally have corresponding summary item performance data attached to them (not shown here).

Lertap calls on Excel to make these charts, passing it information from the rows of response statistics found in the Stats1b worksheet. It instructs Excel to open a new worksheet, Stats1bCht, just for the charts.

Users of Excel 97 should note that the number of charts which may be made is limited to about 50, sometimes less, unless they have installed a special fix to overcome a system limitation specific to this version of Excel. The fix may be found on the Lertap website (www.lertap.com).

## Scores

The Lertap program which creates the Stats1f and Stats1b worksheets is Elmillon. Before Elmillon creates these worksheets, it opens and starts to fill another worksheet called Scores. A sample from a Scores sheet is shown here:

| ID | Knwldge | Comfort |
|---|---|---|
| 9 | 3.00 | 32.00 |
| 31 | 12.00 | 32.00 |
| 26 | 13.00 | 37.00 |
| 27 | 11.00 | 32.00 |
| 21 | 14.00 | 33.00 |
| 59 | 19.00 | 37.00 |
| 47 | 14.00 | 42.00 |
| 42 | 20.00 | 41.00 |

The Scores worksheet always has some sort of ID information in its first column. The numbers or letters or names seen in this column may come directly from one of the initial two columns in the Data worksheet, providing the column heading in the Data worksheet begins with letters "ID" (or "id"—case is not important).

There will usually be one score for each subtest. The scores will be labelled by the characters found in the Title=() control word on *sub cards.

A respondent's score on each subtest is derived by summing over the respondent's score on each of the items which comprise the subtest.

It is possible to copy columns from the Data worksheet to the Scores worksheet, providing the columns contain only numeric data. It is also possible to copy a column from the Scores worksheet to the Data worksheet. The options for copying columns is found under the Move menu on the Lertap toolbar. It might be emphasised that these options copy and paste, they do not actually pick up a column and move it.

At the bottom of the Scores worksheet appear a variety of statistics, such as those shown below:

| | | |
|---|---|---|
| n | 60 | 60 |
| Min | 1.00 | 26.00 |
| Median | 12.50 | 33.00 |
| Mean | 12.63 | 34.48 |
| Max | 24.00 | 43.00 |
| s.d. | 6.95 | 4.61 |
| var. | 48.27 | 21.25 |
| MinPos | 0.00 | 10.00 |
| MaxPos | 25.00 | 50.00 |
| Correlations | | |
| Knwldge | 1.00 | 0.80 |
| Comfort | 0.80 | 1.00 |
| *average* | *0.80* | *0.80* |

All of these statistics, except two, are computed by Excel, not Lertap. To see how they're derived, select a cell and examine the contents of Excel's Formula Bar.

The two statistics not computed by Excel are MinPos and MaxPos, which come from Lertap. The first gives the minimum possible score on each subtest, while the second gives the maximum possible. What is the difference between "Min" and "MinPos"? The first is computed by Excel, and is simply the lowest score found in the subtest's score column. The second is calculated by Lertap using the item weights found in the subtest's Sub worksheet, and is the rock-bottom, absolute minimum score anyone could get on the subtest.

The *average* correlation shown in the very last line is the simple average of each score's correlations with all the other scores. (In this example there is only one other score.)

## Histograms

An option on the Lertap toolbar will create a "Lertap2-style" histogram from any of the columns in the Scores sheet. This is not a chart, but a special table made with text characters. A snippet from one of these histograms is shown below:

| z | score | f | % | cf | c% | |
|---|---|---|---|---|---|---|
| -1.67 | 1.00 | 1 | 1.7% | 1 | 1.7% | □ |
| -1.53 | 2.00 | 0 | 0.0% | 1 | 1.7% | |
| -1.39 | 3.00 | 6 | 10.0% | 7 | 11.7% | □□□□□□ |
| -1.24 | 4.00 | 5 | 8.3% | 12 | 20.0% | □□□□□ |
| -1.10 | 5.00 | 3 | 5.0% | 15 | 25.0% | □□□ |
| -0.95 | 6.00 | 0 | 0.0% | 15 | 25.0% | |
| -0.81 | 7.00 | 4 | 6.7% | 19 | 31.7% | □□□□ |
| -0.67 | 8.00 | 2 | 3.3% | 21 | 35.0% | □□ |
| -0.52 | 9.00 | 0 | 0.0% | 21 | 35.0% | |
| -0.38 | 10.00 | 2 | 3.3% | 23 | 38.3% | □□ |
| -0.24 | 11.00 | 3 | 5.0% | 26 | 43.3% | □□□ |
| -0.09 | 12.00 | 4 | 6.7% | 30 | 50.0% | □□□□ |

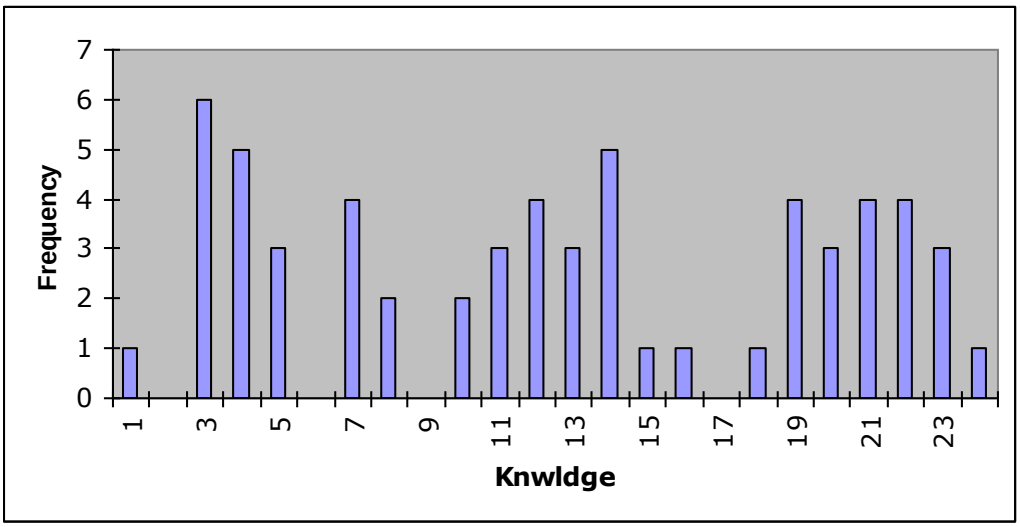
The column headed "z" shows the z-score corresponding to the numeric value found in the adjacent "score" column. The statistics required to calculate z-scores, namely the mean and standard deviation of all scores, are taken from the bottom of the Scores worksheet.

The "cf" column shows the cumulative frequency associated with each score, that is, the number of scores at and below the given score, divided by "n", the total number of scores. The value of n is taken from the bottom of the Scores worksheet.

The symbols used to "plot" the frequency bars change, becoming thinner as the bars become longer. If the frequency, "f", of any score exceeds 200, the bars are rescaled so that each symbol represents more than one case. This maximum length of 200 may be changed—see the section below on the System worksheet.

Lertap2-style histograms are provided in worksheets called, for example, Histo1L.

If users have installed an Excel option, the Analysis ToolPak, Lertap gets Excel to make the Histo1E worksheet, into which it places an Excel chart, such as this one:




Charts like this one are based on an Excel-created "Bin / Frequency" table which appears in the first two columns of the Histo1E sheet, immediately to the left of the chart. The table looks like this:

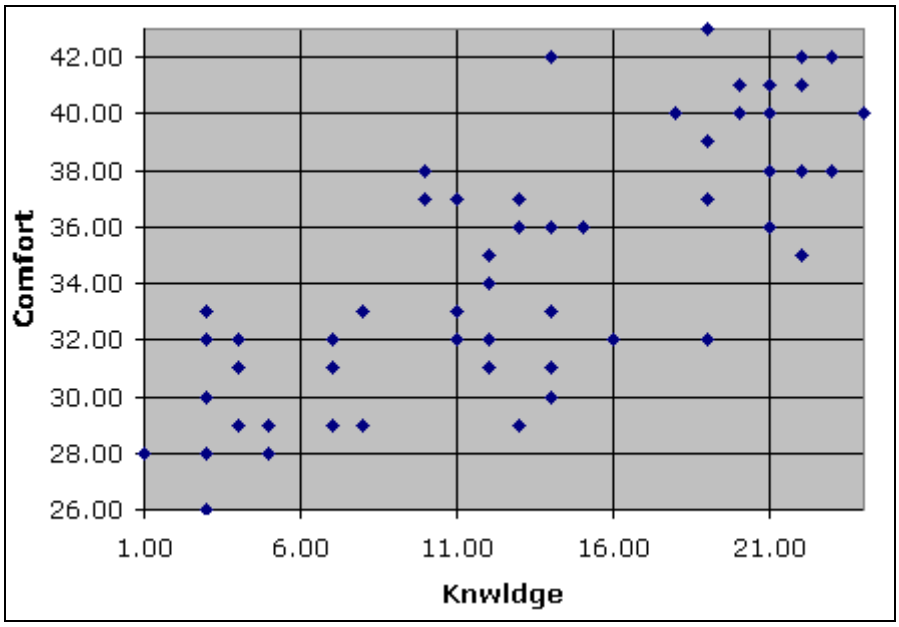
| *Bin* | *Frequency* |
|---|---|
| 1 | 1 |
| 2 | 0 |
| 3 | 6 |
| 4 | 5 |
| 5 | 3 |

This table is live—change one of its values and the chart will also change. Because of this, the table cannot be deleted. However, the chart may be dragged over the table, and the table may be shifted to the right of the chart, if desired.

The Analysis ToolPak is supplied with Excel, but has to be added in before it will work. In fact, it's called an "Add-in"—see Excel's Help for instructions.

## Scatterplots

Excel has many built-in charting routines, and "XY (Scatter)" is one of the easiest to use. Lertap's scatterplot icon asks users to nominate the two scores to be plotted, and then gets Excel to make an XY (Scatter) chart from the corresponding columns of the Scores worksheet.




Lertap uses the Min and Max figures from the bottom of the Scores sheet to get Excel to start and end its X and Y axes at corresponding scores. Users may alter the way the scatterplot looks—there are *many* options which may be applied via Excel's Chart menu.

The Analysis ToolPak is not needed for charts such as this one.

## External criterion statistics

It is possible to have the items of any subtest correlated with any of the scores in the Scores worksheet. The option for doing this is found under the Run menu on the Lertap toolbar.

**Item 30**

| option | wt. | n | p | pb/ec | avg/ec | z |
|---|---|---|---|---|---|---|
| 1 | 5.00 | 9 | 0.15 | 0.43 | 19.78 | 1.03 |
| 2 | 4.00 | 20 | 0.33 | 0.18 | 14.45 | 0.26 |
| 3 | 3.00 | 17 | 0.28 | -0.04 | 12.24 | -0.06 |
| 4 | 2.00 | 8 | 0.13 | -0.38 | 5.88 | -0.97 |
| 5 | 1.00 | 5 | 0.08 | -0.35 | 4.60 | -1.16 |
| other | 3.00 | 1 | 0.02 | 0.01 | 13.00 | 0.05 |
| | | | **r/ec:** | **0.63** | | |

The table above shows the figures which resulted by having Lertap correlate each of the items on an affective subtest, "Comfort", with the "Knwldge" score found in the Scores worksheet.

The pb/ec column gives the point-biserial correlation for each response with the criterion (Knwldge). The avg/ec column indicates the average criterion score for

all those respondents who selected each response option. For example, the average criterion score for the 9 respondents who selected option 1 on Item 30 was 19.78. The "z" value of 1.03 is 19.78 expressed as a z-score. (The average criterion score was 12.63, standard deviation 6.95.)

The **r/ec** figure is the product-moment correlation between the item scores and the criterion scores.

External criterion statistics for cognitive items are identical to those given for affective items. In the case of cognitive items having one correct answer, the r/ec figure will equal the pb/ec value found for the item's correct answer.

External criterion reports are given in worksheets with names such as ECStats1f, ExStats2f, and so forth.

After the item statistics are given, summary statistics for the external criterion are provided. These should be checked with the same statistics which appear at the bottom of the respective Scores column—the two sets of statistics should be identical (if not, an error has occurred—deleting or changing records in the Data and Scores worksheets without going all the way back to the "Interpret CCs lines" option, for example, will cause errors).

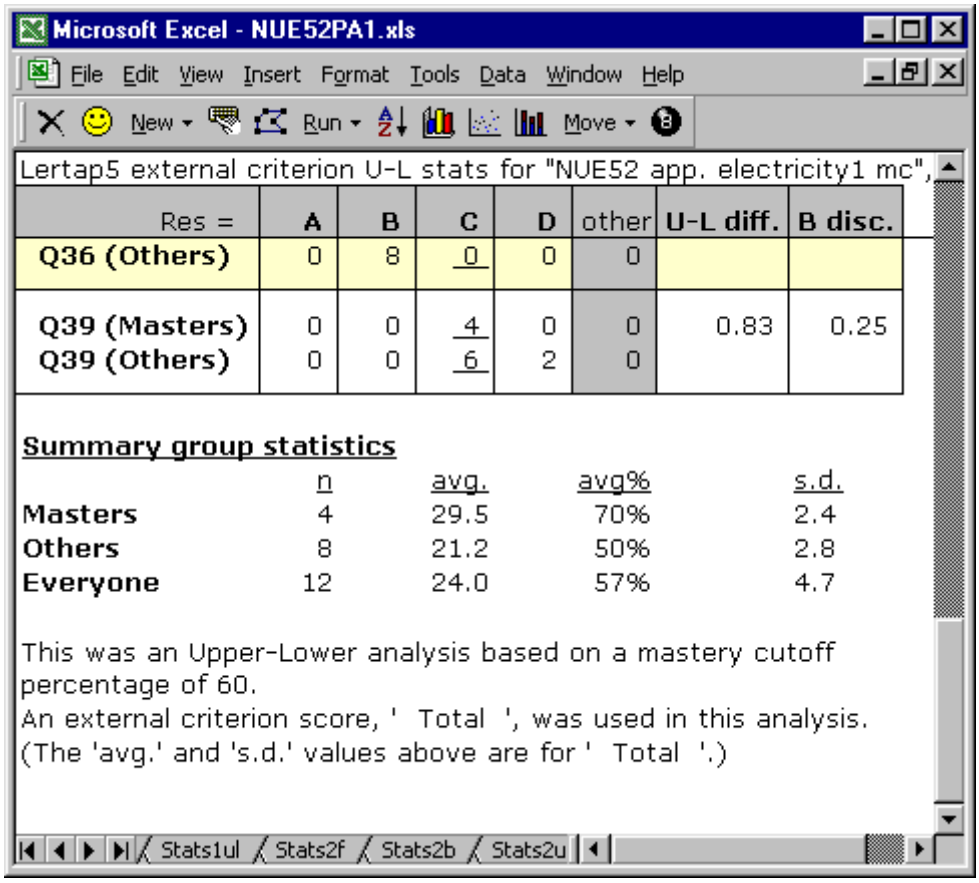
**correlation bands (with external criterion)**

**.00:**
**.10:**Item 4
**.20:**Item 5 Item 22
**.30:**Item 7 Item 14 Item 23 Item 24
**.40:**Item 3 Item 8 Item 9 Item 10 Item 11 Item 12 Item 15 Item 17 Item 18
**.50:**Item 1 Item 16 Item 20
**.60:**Item 6 Item 13 Item 19 Item 21 Item 25
**.70:**Item 2
**.80:**
**.90:**

These correlation bands appear at the very end of the ECStats1f worksheet. They summarise the r/ec values for the items. If an item's r/ec value is negative, the item will be listed in the **.00** band.

**External statistics for U-L analyses**

The sample screen snapshot below exemplifies the report format seen when an external criterion score is used with a U-L analysis:



Microsoft Excel - NUE52PA1.xls

File Edit View Insert Format Tools Data Window Help

New Run Move

Lertap5 external criterion U-L stats for "NUE52 app. electricity1 mc",

| Res = | A | B | C | D | other | U-L diff. | B disc. |
|---|---|---|---|---|---|---|---|
| Q36 (Others) | 0 | 8 | 0 | 0 | 0 | | |
| Q39 (Masters) | 0 | 0 | 4 | 0 | 0 | 0.83 | 0.25 |
| Q39 (Others) | 0 | 0 | 6 | 2 | 0 | | |

**Summary group statistics**

| | n | avg. | avg% | s.d. |
|---|---|---|---|---|
| Masters | 4 | 29.5 | 70% | 2.4 |
| Others | 8 | 21.2 | 50% | 2.8 |
| Everyone | 12 | 24.0 | 57% | 4.7 |

This was an Upper-Lower analysis based on a mastery cutoff percentage of 60.
An external criterion score, ' Total ', was used in this analysis.
(The 'avg.' and 's.d.' values above are for ' Total '.)

Stats1ul Stats2f Stats2b Stats2u

In cases such as this, the upper and lower groups are defined with reference to the external criterion. Here (above) the external criterion was a score called 'Total', and the Masters group was defined as those having a score of 60% or better on this criterion.

## Item scores matrix

An option on Lertap's Run menu will produce matrices of item scores and intercorrelations for any subtest. An example of item scores is shown here:

Lertap5 IStats matrix, last updated on: 24/08/00.

| ID | Item 26 | Item 27 | Item 28 | Item 29 | Item 30 |
|---|---|---|---|---|---|
| 9 | 2.00 | 3.00 | 5.00 | 4.00 | 4.00 |
| 31 | 3.00 | 3.00 | 2.00 | 3.00 | 3.00 |
| 26 | 4.00 | 4.00 | 4.00 | 5.00 | 3.00 |
| 27 | 2.00 | 2.00 | 3.00 | 5.00 | 4.00 |
| 21 | 2.00 | 2.00 | 3.00 | 5.00 | 4.00 |
| 59 | 4.00 | 3.00 | 4.00 | 4.00 | 4.00 |

This example is from an affective subtest. It displays the "score", or "points", earned by each respondent on each item of the subtest. When the subtest is a cognitive one, this matrix will usually consist of zeros and ones, as shown below:

Lertap5 IStats matrix, last updated on: 24/08/00.

| ID | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 |
|---|---|---|---|---|---|---|---|
| 9 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 31 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 |
| 26 | 0.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| 27 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 |
| 21 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| 59 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Summary item statistics and intercorrelations are given at the bottom of the IStats worksheet, and have the format shown here:

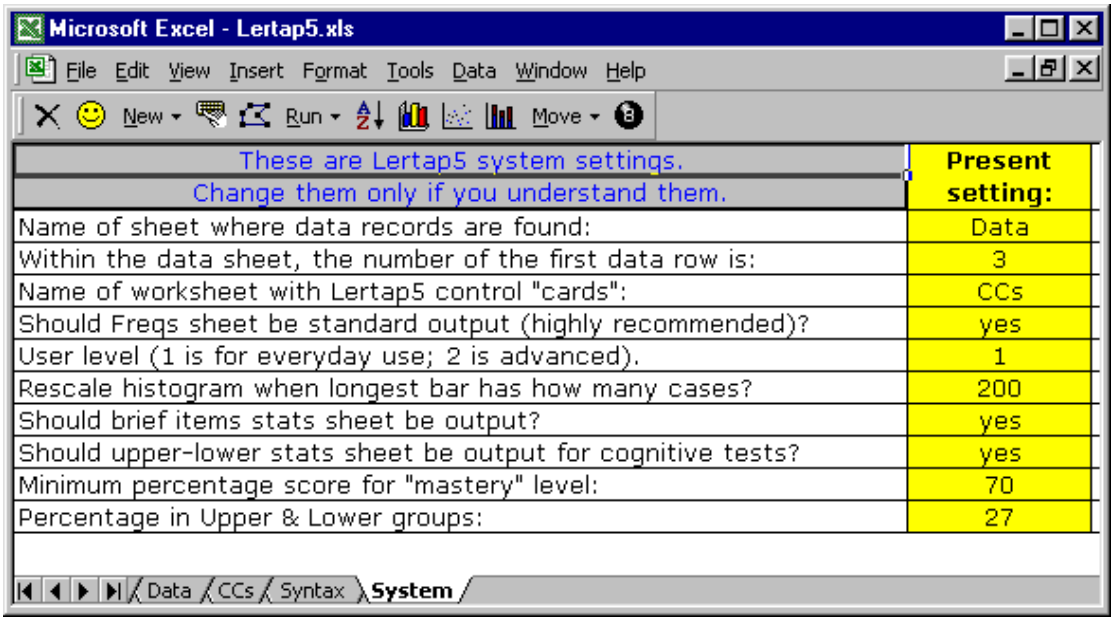
| | | | | | |
|---|---|---|---|---|---|
| n | 60 | 60 | 60 | 60 | 60 |
| Min | 1.00 | 2.00 | 2.00 | 2.00 | 1.00 |
| Median | 3.00 | 3.00 | 4.00 | 4.00 | 3.00 |
| Mean | 3.08 | 2.98 | 3.75 | 3.93 | 3.33 |
| Max | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |
| s.d. | 1.28 | 0.88 | 0.94 | 0.89 | 1.14 |
| var. | 1.64 | 0.78 | 0.89 | 0.80 | 1.29 |
| Correlations | | | | | |
| Item 26 | 1.00 | 0.57 | -0.02 | 0.41 | 0.45 |
| Item 27 | 0.57 | 1.00 | 0.17 | 0.23 | 0.34 |
| Item 28 | -0.02 | 0.17 | 1.00 | -0.14 | -0.12 |
| Item 29 | 0.41 | 0.23 | -0.14 | 1.00 | 0.37 |
| Item 30 | 0.45 | 0.34 | -0.12 | 0.37 | 1.00 |
| Item 31 | -0.01 | 0.14 | -0.05 | 0.01 | -0.04 |
| Item 32 | 0.24 | 0.06 | -0.07 | 0.27 | 0.04 |
| Item 33 | 0.74 | 0.45 | -0.19 | 0.43 | 0.55 |
| Item 34 | -0.41 | -0.34 | -0.02 | -0.31 | -0.36 |
| Item 35 | 0.66 | 0.34 | -0.20 | 0.36 | 0.51 |
| *average* | *0.29* | *0.22* | *-0.07* | *0.18* | *0.19* |

These statistics are all produced by Excel. Respective equations may be seen, in Excel, by clicking on any of the statistics cells, and then examining the contents of Excel's Formula Bar.

The *average* item correlation figure, shown in the last row, is the mean of an item's correlations with the other items in the subtest.

## The System worksheet

One of the five worksheets in the Lertap5.xls file is a hidden one called System.




| These are Lertap5 system settings. Change them only if you understand them. | Present setting: |
|---|---|
| Name of sheet where data records are found: | Data |
| Within the data sheet, the number of the first data row is: | 3 |
| Name of worksheet with Lertap5 control "cards": | CCs |
| Should Freqs sheet be standard output (highly recommended)? | yes |
| User level (1 is for everyday use; 2 is advanced). | 1 |
| Rescale histogram when longest bar has how many cases? | 200 |
| Should brief items stats sheet be output? | yes |
| Should upper-lower stats sheet be output for cognitive tests? | yes |
| Minimum percentage score for "mastery" level: | 70 |
| Percentage in Upper & Lower groups: | 27 |

Data CCs Syntax **System**

The System sheet contains settings which are used as parameters for Lertap's operations. All of these settings may be changed; however, changing one of the first three settings ("Data", "3", "CCs") is *not* recommended at all, and may result in unpredictable consequences.

There are three "yes" settings, each referring to the creation of one Lertap's reports. If these are changed to "no", the corresponding report (worksheet) will not be created.

The user level setting determines the number of icons which display on the Lertap toolbar—see the following section for more details.

The rescale histogram setting makes reference to the length of the bars produced in the Lertap2-style histogram known as "Histo1L". In large data sets these bars can become too long, going off the screen, and off the printer. Whenever a single bar becomes longer than 200 symbols, Lertap increases the number of cases represented by the symbol so that the bar shortens. For example, if the frequency corresponding to a particular test score is, say, 800, Lertap will divide this value by 200, and let each bar symbol represent four (4) cases.

How to access the System worksheet? It's a hidden sheet in the Lertap5.xls workbook. The workbook is protected with a password of "shack25"—use Excel's Tools / Protection option to Unprotect Workbook, type in the password without the quotation marks, and then use the Format / Sheets option to Unhide System.

**Advanced level toolbar**

One of the settings in the System worksheet has to do with what is called the "user level". If this setting is changed from its default value of 1 (one) to 2, a different Lertap toolbar results[6]:


This toolbar differs from the standard one in the two icons seen immediately to the right of the Run option. The first of these icons toggles hide / unhide for a workbook's Sub sheets, sheets which are normally hidden. This icon makes it easy to unhide all of them with just one click.

The second icon, one we call the "Liberty Bell", gets Elmillon to act on a single Sub sheet, as opposed to all of them. Normally, when the "Elmillon item analysis" option is taken, all of a workbook's Sub sheets are read, and scores and reports are produced for each. Use of the Liberty Bell makes it possible to act on just one selected Sub sheet.

The idea here is to give users the power to make alterations to their subtests at the most basic level. To investigate the effects of different item response weights, for example, a user could use Excel to copy a given Sub worksheet, go to the copy, change response weights, and then apply the Liberty Bell to get results. For some users this will be more convenient than having to go all the way back to the CCs sheet, add new control cards, and then go through the "Interpret CCs lines" and "Elmillon item analysis" options again.

Readers might be surprised to hear that they can get Lertap to do its things without having to create lines in a CCs worksheet. The program which does the real analysis work for Lertap is Elmillon. Elmillon does not read CCs lines—it looks for Sub sheets, from which it derives all of its information. If users can create their own Sub sheets, using the right format, they need not be concerned about lines of job definition statements in the CCs file.

## Exporting worksheets

An option on Lertap's toolbar will take any of the worksheets in the currently-active workbook, and turn them into a single Excel 4 worksheet. This option is provided as Excel 4 worksheets are quite easy to import in other programs, such as SPSS, something which cannot always be said of worksheets from Excel 95, 97, 98, and 2000. For example, version 9 of the SPSS package will hiccup a bit if asked to import from Excel 97—users generally have to work through an ODBC (open data base connectivity) interface, which is not overly straightforward. (It seems that version 10 of SPSS may have fixed this problem, but read on.)

The Lertap worksheets which users may want to convert to Excel 4 format will usually be Data, Scores, or IStats. These worksheets have their two top-most rows reserved for titles and headers. A program such as SPSS will want to see just one such row, so Lertap's Excel 4 converter will strip the first row from the Data, Scores, and IStats worksheets' leaving just the row of column headers, which SPSS is happy to see (it uses them as variable names—when opening an xls file from within SPSS, be sure and tick the SPSS box which says Read variable

[6] The workbook has to be saved and then re-opened before the change will become effective.

names). Lertap's Excel 4 converter will also strip the statistics rows from the bottom of the Scores worksheet.

## Time trials

In October 2000 we selected five data sets and ran them through Lertap 5 on three quite different computers. The results reported below should be seen as relative ones—the system used in October was not the final version.

### The data sets

"LRTP Quiz" is the standard data set which comes built into the Lertap5.xls file. It has 60 respondents and two subtests. The first subtest is a cognitive one, with 25 items, while the second is an affective one with 10 items.

"Ed 502" is typical of the type of results which many teachers collect. It's from an end-of-semester test given to 48 graduate students. It had one subtest, a cognitive one with 63 items.

"Nanta" is the name of a data set from Dr Nanta Palitawanont of Burapha University, Thailand. It involved 300 students responding to an affective instrument with 10 scales, with an average of seven items per scale.

"UCV 1" is from la Universidad Central de Venezuela, and involved 1,494 students responding to three cognitive tests, each with 20 items[7].

"UCV 2" is from the same university. This data set had results from 11,190 students who sat an entrance exam having two 25-item cognitive subtests.

### The computers

"LeoPAD1" was a generic laptop computer running a Pentium MMX CPU, 166 MHz clock, and 32 M RAM. It was running Excel 2000 under Windows 98. This computer would, these days, be considered an "old" Pentium.

"G3" was a Macintosh G3 computer, running at 350 MHz, and having 132 M RAM. G3 was using Excel 98. Readers should note that we're not experienced Mac users. The G3 did very little Lertap work for us until we increased Excel's memory slice to 50 K (which seemed a small amount, but got the job done; a larger memory chunk may have resulted in better performance).

"Marek" was a Pentium III machine with 128 M RAM, running at 450 MHz. It was running Excel 2000 and NT 4.0.

### The results

We present results for three levels of processing, "Int. CCs", which means "Interpret CCs lines", "Elmlln", which refers to "Elmillon item analysis", and "No U-L", which means a job with U-L analyses disabled. We included the latter level as it is likely some users will turn the U-L analysis off; we knew it added to the processing burden, and results confirm this.

[7] Thanks to Carlos Gonzalez of UCV for making these data sets available for our benchmarking tests.

In the table, "secs." means seconds, and "mins." means minutes. The reason "n.a." is given for the Nanta data set has to do with the fact that this data set involved only affective subtests; U-L analyses are only relevant to cognitive tests.

| | LeoPAD1 | G3 | Marek |
|---|---|---|---|
| **LRTP Quiz** | | | |
| Int. CCs | 30 secs. | 12 secs. | 3 secs. |
| Elmlln | 63 secs. | 23 secs. | 8 secs. |
| No U-L | 34 secs. | 14 secs. | 5 secs. |
| **Ed 502** | | | |
| Int. CCs | 28 secs. | 13 secs. | 5 secs. |
| Elmlln | 81 secs. | 22 secs. | 9 secs. |
| No U-L | 50 secs. | 13 secs. | 5 secs. |
| **Nanta** | | | |
| Int. CCs | 3 mins. | 1 min. | 18 secs. |
| Elmlln | 7 mins. | 2 mins. | 37 secs. |
| No U-L | n.a. | n.a. | n.a. |
| **UCV 1** | | | |
| Int. CCs | 7 mins. | 2 mins. | 28 secs. |
| Elmlln | 9 mins. | 4 mins. | 53 secs. |
| No U-L | 6 mins. | 2 mins. | 41 secs. |
| **UCV 2** | | | |
| Int. CCs | 33 mins. | 10 mins. | 3 mins. |
| Elmlln | 45 mins. | 24 mins. | 5 mins. |
| No U-L | 31 mins. | 19 mins. | 4 mins. |

It is likely readers will have a variety of reactions to these performance figures. Some may be surprised that the analyses can take minutes to complete in some cases. Others, perhaps more experienced data analysts, may marvel that data sets with more than ten thousand cases can be processed so quickly on a desktop. It used to be that we'd have to carry five boxes of punch cards to the "computing centre", or a magnetic tape, and generally plan to spend at least a couple of hours waiting for results.