Highlighting Missing Data in Lertap5

Larry R Nelson[1], Curtin University, Australia
20 July 2022

Enhancements have been made to Lertap5 so that assessing the presence and extent of missing data is easier to do, with "NA" used to denote missing data whenever users ask Lertap5 to output a csv (comma-separated values) file with item scores[2].

This document is largely designed for Lertap5 users who might desire to export data in the csv file format for use with R scripts, RMD scripts, or data analysis packages such as JASP and SPSS, paying special attention to matters related to how missing data are processed.

I'll use the 10-item "Comfort" scale from the LertapQuiz dataset built into Lertap5.

Figure 1

Figure 1 displays the first 12 records of Comfort Scale item responses. There were a total of 60 respondents.

Each of the 10 items on the scale allowed for five responses: 1, 2, 3, 4, 5. The CCs sheet used to process results is shown in Figure 2.

Ordinarily only the first three rows in Figure 2's CCs sheet would be used. However, later on I will be talking about "MDO" and thus have defined a second Lertap5 subtest, giving it the name of "MDO Comfort", and including the MDO keyword on the second sub Aff row.

Missing data in Lertap5 arise in just two cases: a cell in the Data sheet will be empty and/or will contain a response not seen in the Res= declaration on a *sub line.

---

[1] l.nelson@curtin.edu.au
[2] Figure 8 has an example of a csv file.

Figure 1 indicates that the first record was missing a response on item Q32. The last item, Q35, turned out to have 9 missing responses – three of them are showing in Figure 1, corresponding to records 4, 6, and 7. (Readers unfamiliar with Lertap5 may benefit from this pdf document.)
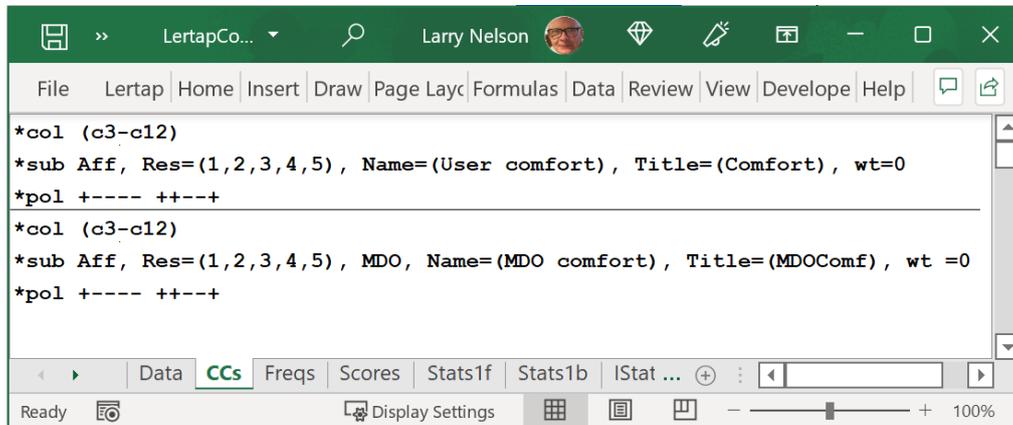

Figure 2

Lertap5's "Interpret" option will create a "Freqs" worksheet with a summary of item responses. Figure 3 shows results for Q34 and Q35. As mentioned above, Q35 was missing a response from 9 people.
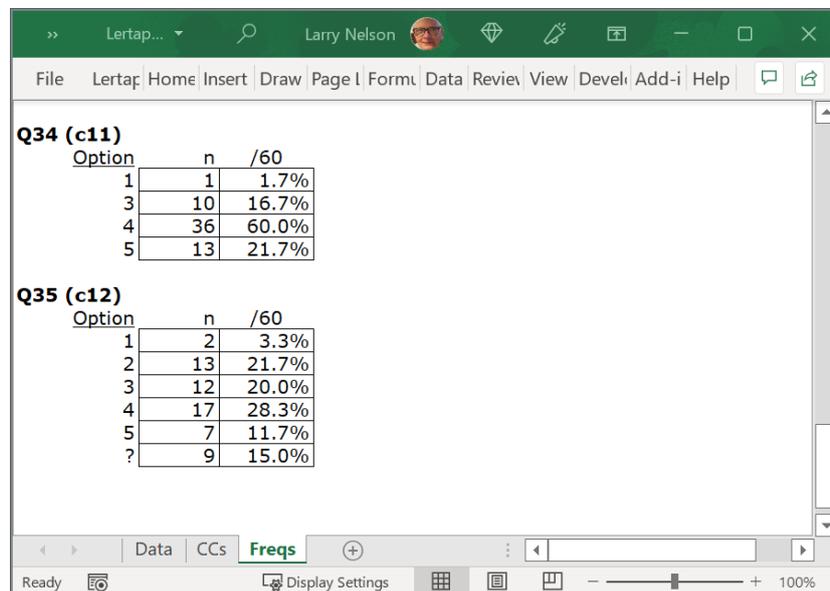

Figure 3

Once the Interpret option has been run, and the Freqs report studied, the Elmillon option is then called on. It will add summary missing data information to the bottom of the Freqs report, as seen in Figure 4, and then proceed to create three new worksheet reports, including "Stats1b", found below in Figure 5.
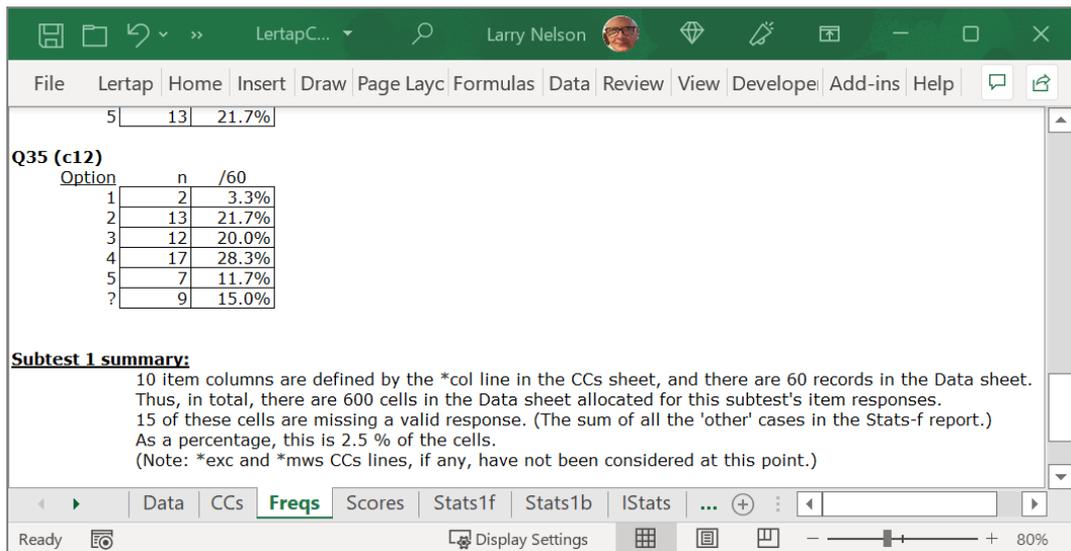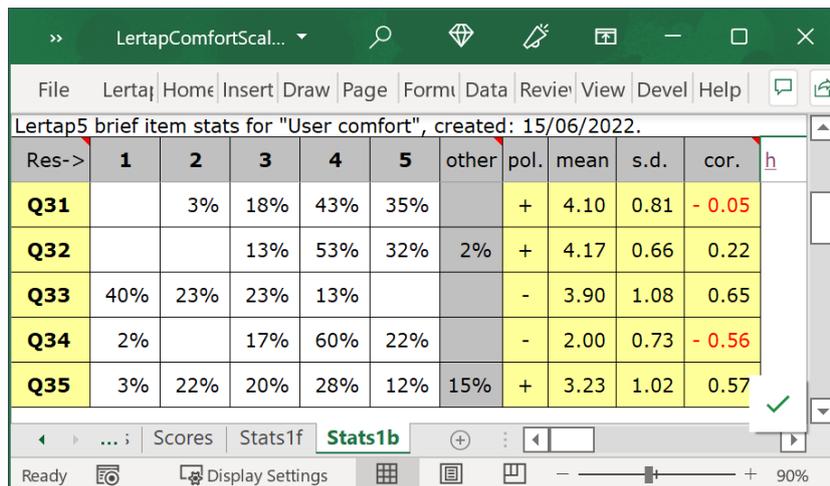
Figure 4

Figure 5

The "Other" column in a Stats1b report indicates the percentage of data records having missing data on each item. For Q35 the figure is 15%, corresponding to 9 divided by 60.

Two of Q35's results, the mean of 3.23 and "cor." of 0.57, are further discussed below. More about Stats1b figures are provided here.

Before getting into more meaty matters, now that the Interpret and Elmillon options have been taken, look anew at the Data sheet (Figure 6):

Figure 6

Colouring has now been used to highlight cells with missing data. Each time Lertap5 starts up, among numerous other behind-the-scenes events, it looks at Rows 78 and 79 of its System work-sheet to make sure that users want to have highlighting applied to cells with missing data.

The default settings are to have missing data cell highlighting on, using a light-blue background colour. Other colours are possible.

Figure 7 displays "IStats", the report created whenever the "Item scores and correlation" option is taken. It indicates the scoring points each participant got for their answer on each item. The sum of all these points, the "item scores", becomes the participant's total test score.

Figure 7

In Figure 7, Lertap5 has made a [missing-data substitution](), putting values (scores) of 3.00[3] in those cells that had missing data, leaving the cell highlighting in place. This is a default action. It may be turned off by using "MDO" on a *sub line in the CCs sheet. This is seen above in Figure 2. (MDO = missing data substitution off.)

Had MDO been used, the cells with missing data would still be highlighted in the IStats report, but cell values (scores) would be 0.00 .

Stats1b reports (Figure 5) indicate the correlation between each item and the total test score, with the item omitted from it (the total test score) in order to control for part-whole inflation, referred to as "spuriousness" in some other apps.

In this case, with missing data substitution in effect (Lertap5's default action), Q35's correlation with the total test score was 0.57, based on all 60 records, with the item's mean being 3.23. With MDO active, the 9 records with missing data are excluded and the correlation drops to 0.56, with mean 3.27, based on 51 records. Not a big difference *in this example*.

Finally, I move on to using the "[Omega1 macro]()", knowing that it will create a csv file with item scores, a file which I may then use with R/RMD scripts, and JASP (or SPSS).

Figure 8 has the "IScores" report created by the macro, displayed as another Excel worksheet, while Figure 9 captures the first few rows of the csv file using Windows Notepad (TextEdit is the equivalent on a Mac).

Figure 8

Figure 9

Missing data have been replaced with "NA" in Figures 8 and 9. JASP and many R/RMD scripts will readily recognise NA[4] as a missing data code and will adjust their statistics accordingly.

Lertap5 version 5.11.2 was used in this paper. It may be downloaded here. **JASP** is a capable, free data analysis system available here. Appendix B in this paper exemplifies the use of JASP.

Suggested reading:

Dai, S (2021). Handling missing responses in psychometrics: methods and software. Psych 2021, 3, 673-693 (https://doi.org/10.3390/psych3040043).

Appendix

The small note at the bottom of Figure 4 advises that the results in Subtest 1's summary have been derived without consideration of any CCs lines which may have been used to delete an item from the subtest, such as a *exc or *mws line.

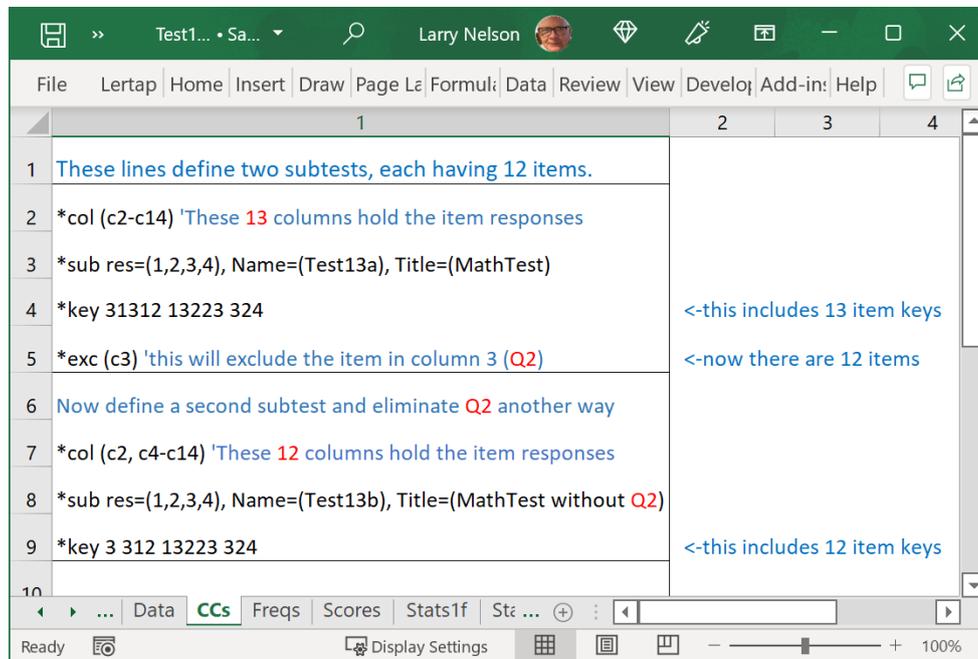Look at this little example with two subtests. It has to do with the Test13 sample dataset:

Figure 10

Test13 was a 13-item multiple-choice test of elementary mathematics ability given to primary school students, with items 'numbered' from Q1 to Q13. Student responses were entered in columns 2 to 14 of the Data worksheet (see line 2 in Figure 10), with 9s used if a student did not respond.

The second item, Q2, with responses found in Data column 3, was found to be of questionable relevance, making it desirable to exclude it from Lertap5's analyses. This can be done in two ways.

Line 5 uses *exc (c3) to exclude Q2 from the first subtest (c3 means column 3).

Line 7 effectively excludes Q2 from the second subtest by not referring to its corresponding Data column (c3). Note that the *keys seen in line 9 now have 12 entries – the key for Q2 is missing.

Figure 11 has the missing data summary seen at the bottom of the Freqs report.

The summary for the first subtest covers 13 items while that for the second subtest covers 12 items.

Both summarise the extent of missing responses in the Data sheet. However, the first summary may be of little or no interest as it references all items, including Q2. Lertap5 has ignored the *exc statement in line 5 of the CCs sheet (an oversight which may be corrected in a future revision).

Finally, compare the number of cells found to be missing a valid response: 3715 for subtest 1 and 3316 for subtest 2, a difference of 399. Then look at Figure 12. In this dataset, a code of 9 was used if a student failed to answer a question. Figure 12 shows the 399 missing on Q2, the excluded item.
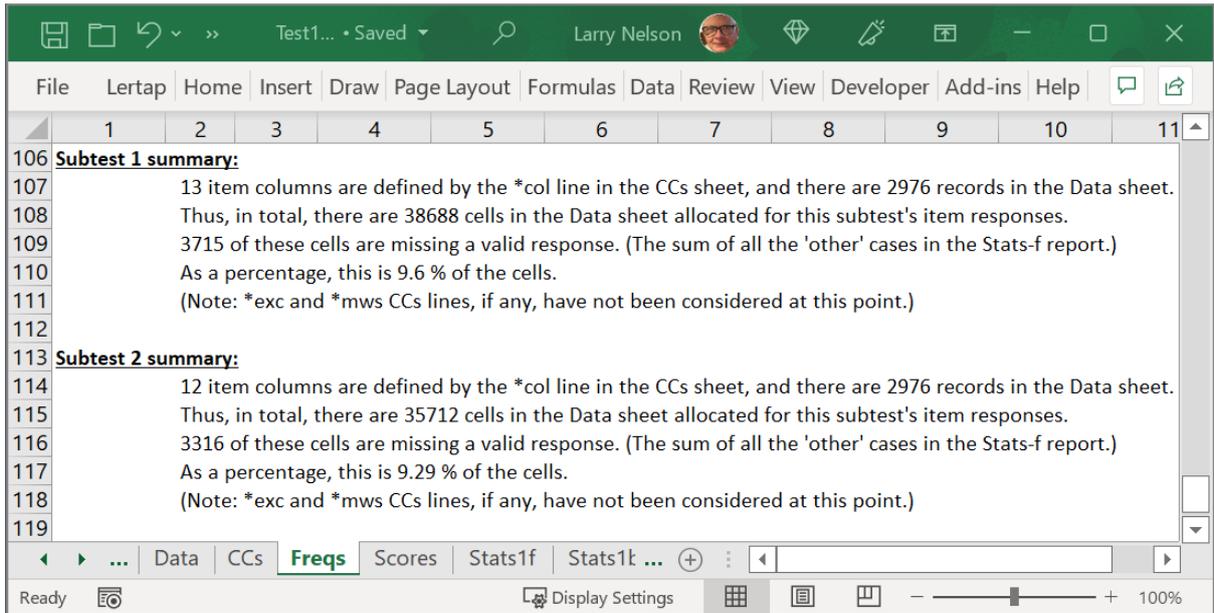
Subtest 1 summary:

106 **Subtest 1 summary:**

107   13 item columns are defined by the *col line in the CCs sheet, and there are 2976 records in the Data sheet.

108   Thus, in total, there are 38688 cells in the Data sheet allocated for this subtest's item responses.

109   3715 of these cells are missing a valid response. (The sum of all the 'other' cases in the Stats-f report.)

110   As a percentage, this is 9.6 % of the cells.

111   (Note: *exc and *mws CCs lines, if any, have not been considered at this point.)

112

113 **Subtest 2 summary:**

114   12 item columns are defined by the *col line in the CCs sheet, and there are 2976 records in the Data sheet.

115   Thus, in total, there are 35712 cells in the Data sheet allocated for this subtest's item responses.

116   3316 of these cells are missing a valid response. (The sum of all the 'other' cases in the Stats-f report.)

117   As a percentage, this is 9.29 % of the cells.

118   (Note: *exc and *mws CCs lines, if any, have not been considered at this point.)

119

Figure 11

Option	n	/2976
2	291	9.8%
3	2,235	75.1%
4	140	4.7%
9	124	4.2%

Q2 (c3)

Option	n	/2976
1	1,048	35.2%
2	637	21.4%
3	526	17.7%
4	366	12.3%
9	399	13.4%

Q3 (c4)

Figure 12