The LTA Data Aggregation Program

Brian P. Flaherty

The Methodology Center

The Pennsylvania State University

August 1, 1999

I. Introduction

Getting data into response pattern format (grouped data) has often been a difficulty for new users of latent transition analysis (LTA; Collins, Wugalter & Rousculp, 1992). Typically, researchers have data files which are at the level of the individual. Individual level data files have a separate set of responses for each person in the sample. If one collected all the individuals who made identical responses to each of the items, one could record that set of responses once and follow it with the number of people making those particular responses. In other words, one could cross-tabulate the item responses and record the number of subjects in each cell of the contingency table. This is the format in which LTA expects data. The program documented here creates, from a rectangular, individual level ASCII text file, a response pattern data file useable by LTA.

II. Preprocessing the Data

The individual level data need to meet certain criteria before this program can accurately aggregate the data. The responses must be coded correctly, including any missing values, and the variables must be written out in a certain order.

LTA is a categorical data program, so all variables in an analysis must be coded categorically. While the coding of categorical variables is arbitrary, LTA requires that response categories be consecutive positive integers starting with one. So even if the codes you are working with are currently letters or some other set of labels, these must be recoded in order for

this program and LTA to handle them properly. If missing values will be included in the LTA analysis, the missing values must be coded as zero. This is the only code telling LTA that the response to a particular item is missing. This recoding is easily performed in a statistical package like SPSS, SAS, etc. The individual level data should be written out with at least one space between the variables, as shown in figure 1.

LTA expects the items in the analysis to appear in the data file in a certain order. For example, if one were analyzing four items measured three times, the items must appear in the same order within each time, and the four time 1 items must precede the time 2 items, which must precede the time 3 items. If latent class indicators are also being used, they must come before all items measuring the latent statuses.

Finally, some people think that it is incorrect to have people in the analysis with all responses missing. If one does not wish to have a completely missing response pattern in the analysis, then they should also be removed during this preprocessing step.

III. Running the Data Aggregation Program

The data aggregation program is invoked at the command prompt, either DOS or UNIX, by the command dataagg.exe. Once started, the user is asked four questions and then the program runs. The user must type in

```
0 0 0 0 0 0 0 0 2 2 2 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 2 2 2 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 2 1 2 1 2 1 1 1 0 0 0 0
0 0 0 0 0 0 0 0 2 1 1 1 2 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 2 2 2 2 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 2 2 2 2 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 1 2 2 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 1 1 1 1 2 2 2 1 2 1 1 1
.
.
.
2 1 2 1 2 2 2 2 2 1 2 1 2 1 2 2 0 0 0 0
2 2 2 2 2 2 2 2 0 0 0 0 0 0 0 0 2 2 2 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
.
.
.
```

Figure 1. An example individual level data file.

the name of the input (individual level) data file, the name of the output data file, the number of respondents in the input data file, and the number of variables (columns) in the input data file.

For example, if an LTA problem has four items measured at five times, the input data might look like Figure 1. These are the data for the subjects comprising the analysis. When running the data aggregation program, the name of the input and output files need to be specified, as well as the number of subjects and variables, here equal to 20. Figure 2 shows part of the response pattern file created by the data aggregation program.

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1          11
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 1           7
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 2           1
.
.
.
1 1 0 0 2 2 1 1 0 0 0 0 2 1 1 1 2 2 1 1           1
1 1 0 1 2 1 2 2 0 0 0 0 0 0 0 0 0 0 0 0           1
1 1 1 0 0 0 0 0 2 1 2 1 2 2 2 2 2 2 2 2           1
1 1 1 0 0 0 0 0 2 2 2 1 0 0 0 0 2 2 2 1           1
1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1           1
1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0         162
1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0           2
.
.
.
```

Figure 2. An example of the resultant response pattern file.

IV. Limitations

Currently, there are a few limitations to the program. The upper limit on the number of variables is 9999 at one time of measurement. It is unlikely that this will prove to be a constraint. The only real limitation is that the program is slow with a large data file. The data needs to be sorted, and the sort routine is not the fastest available. The sort routine is a selection sort (adapted from Ellis, Philips, & Lahey, 1994). Future versions of the program may have an improved sorting algorithm.

V. Bug reports

Send bug reports, suggestions, and/or feature requests to Brian P. Flaherty in either of these methods.

Email: bxf4@psu.edu

US Post Mail: The Pennsylvania State University

The Methodology Center

S-159 Henderson Bldg.

University Park, PA, 16802

References

Collins, L. M., Wugalter, S. E., & Rousculp, S. S. (1992). *Latent Transition Analysis (LTA) Program Manual (Version 1.0)*. Los Angeles: University of Southern California.

Ellis, T. M. R., Philips, I. R., & Lahey, T. M. (1994). *Fortran 90 Programming.* New York: Addison-Wesley.