Read Me
TLMPS 2014 data version 2.0
6.22.16

Files:

Questionnaires
Questionnaires in English and Arabic
1—Household (variables denoted with an h[question #])
2A—Individual Child (variables denoted with a c[question #])
2B—Individual Adult (variables denoted with a v[question #])
3—Migration & Enterprises (variables denoted with a m[question #])

Codebooks
File: Codebook TLMPS 2014 data v2.0.xls
For created variables only. Gives variable name, label, ranges and states, universe, and notes. Please be sure to use the created variables (rather than raw variables) when they are available.

Additional Codebooks
Folder "Additional Codebooks" includes codebooks for:
- Additional variables codebook (specializations (v307, v310, v315, v332), training centers (v331), additional higher education institutions (v309, v313), additional primary schools (c313, v343), additional preparatory schools (c305, v335), additional secondary schools (v321))
- Liste pays (countries)
- Codage des établissements Ens Superieur (Higher Education Institutions; see also Additional Variables Code book for non-standard codes)
- الإبتدائي - المؤسسة حسب معطيات 2014-2015 (primary school codes)
- وتقني وثانوي إعدادي - المؤسسة حسب معطيات 2014-2015 (secondary school codes)
- Specialite ens superiur (Higher education specialization codes)
- Crop codes CTP 2009 _Structure (Ar) (Crop codes)
- Economic Activity codes NAT09_Fr (Economic Activities)
- Occupation codes NNP2014_Fr (Occupation)
- TLMPS 2014 geographic codes Tunisia (geographic codes)

Data Files
Files: TLMPS 2014 v2.0 all.dta (all variables)
    ---OR---
    TLMPS 2014 v2.0 pt 1.dta (created variables and questionnaires 1 & 2)
    TLMPS 2014 v2.0 pt 2.dta (created variables and questionnaire 3)

NOTE: the "TLMPS 2014 v2.0 all.dta" data file contains more than 2,000 variables. If you are using STATA IC, this file has too many variables to open. Use "TLMPS 2014

v2.0 pt 1.dta" and "TLMPS 2014 v2.0 pt 2.dta" instead. If you can open a file with more than 2,000 variables, you can just use ""TLMPS 2014 v2.0 all.dta"

hhid: unique household identifier
indid: unique individual identifier

Weights
expan_hh: household questionnaire expansion factor
expan_roster: household roster expansion factor
expan_indiv: child/adult questionnaire expansion factor
expan_migr_ent: household migration/enterprise questionnaire expansion factor

Note: Data is not self-weighted. Weights need to be used to get representative statistics. Weights should be used in STATA as analytical or probability weights. The weight used should be selected based on which questionnaire the *dependent* variable comes from. If data is being used across multiple questionnaires, we still recommend using the weight based on the questionnaire the dependent variable comes from, but that covariates be created so that observations missing data on those covariates are included with dummies for missing. For instance, if using receipt of remittances (covariate) to predict child schooling (dependent variable), create remittance receipt categorically/as a series of dummies for: 0 (did not receive), 1 (did receive), 2 (data missing).

Do Files:
The do files used to create the 2014 data from the raw data are available. All of the do files run in sequence based on "TLMPS Master Do File v51 CK". If you are interested in particular variable's creation, you can also search for it among the do files, but be aware that some variables are created or modified across multiple files.

Problems with the Data:
There are a large number of issues with the data that users should be aware of:

Observations Missing Corresponding Data:
There are individuals in the household questionnaire and household roster who should have answered a questionnaire (migration, adult, child) and did not. The variables stat_adult_quest, stat_child_quest, and stat_migr_quest indicate whether observations were matched successfully. Non-response to these questionnaires has been incorporated into the questionnaire specific weights.

Skip pattern problems:
Skip patterns were not always obeyed as they should be. Created variables are based on the preceding information and applying skip patterns as in the questionnaire.

Contradictory information:
Individuals often responded contradictorily in different sections (identified as a wage worker in one section, not in another). Generally, created variables were defined based on the earliest information in the questionnaire.

Non-response:
There are substantial problems with non-response to different sections/questions. While we have created weights for total non-response to a questionnaire part (individual (adult/child) or migration/enterprise), there is a great deal of non-response and missing data on individual questions. There is therefore a lot of missing data in both raw and created variables. If a variable of interest to you has substantial missing data, it may be necessary to use missing data techniques (for instance multiple imputation) to achieve representative statistics. Non-response to full questionnaires was definitely non-random and it is therefore likely that missing data on specific questions is also non-random, which will bias results.

Educational Attainment Data:
For children 6-14, in the education section those who entered basic as their highest level (304a) have been skipped over the question on the school year (304b). Likewise, all adults 15-45 who attended school were skipped over the question on the school year (308b), except those who attended the third cycle of higher education (in 308a). There were a few individuals who did not obey these skip patterns, but mostly the key year-within-level data is missing. This problem is based on the questionnaire skip pattern, but is a problem for identifying years of schooling and educational attainment. We reconstructed educational attainment from the life events calendar as much as possible.

Data cleaning philosophy:
As we generated created variables, we enforced the skip patterns that should have been applied, and when there were data contradictions, we treated the earlier information as correct. For example, individuals were asked if they had worked in the past 3 months. Those who had worked in the past 3 months were then supposed to be asked their employment status in a later section. In that later section, some individuals who had not worked in the past 3 months (per the earlier data) gave an employment status. In the created variables, we set the (later) employment status variable to missing if the (earlier) employment questions indicated the individual had not been working.

For "raw" (h*, v*, c*, m*) variables, because we wanted to allow researchers the option to make their own case-by-case decisions on contradictory information or data that did not follow the skip patterns, we undertook only "light" cleaning. Invalid responses (e.g. 0 for a 1 "yes" 2 "no" variable, "2018" for a birth year) were all removed. Responses with clear and minor entry errors (e.g. 998 for a year of birth, presumably instead of 9998 for don't know) were recoded to correct codes. Due to the programming in data collection, unfortunately it was not always possible to distinguish between a missing response and a no or zero response for some variables. For example, when individuals were asked their highest year of schooling completed within a level, everyone who did not answer the question was given a zero, but zero would also be a valid response for someone who was just starting school. When there was a zero response, we set it to missing if it was given by an individual who, per the preceding raw variables and skip patterns, should not have answered the question. However, remaining zero responses (or "no" responses for some variables) might actually represent missing data, and should be treated with some caution.