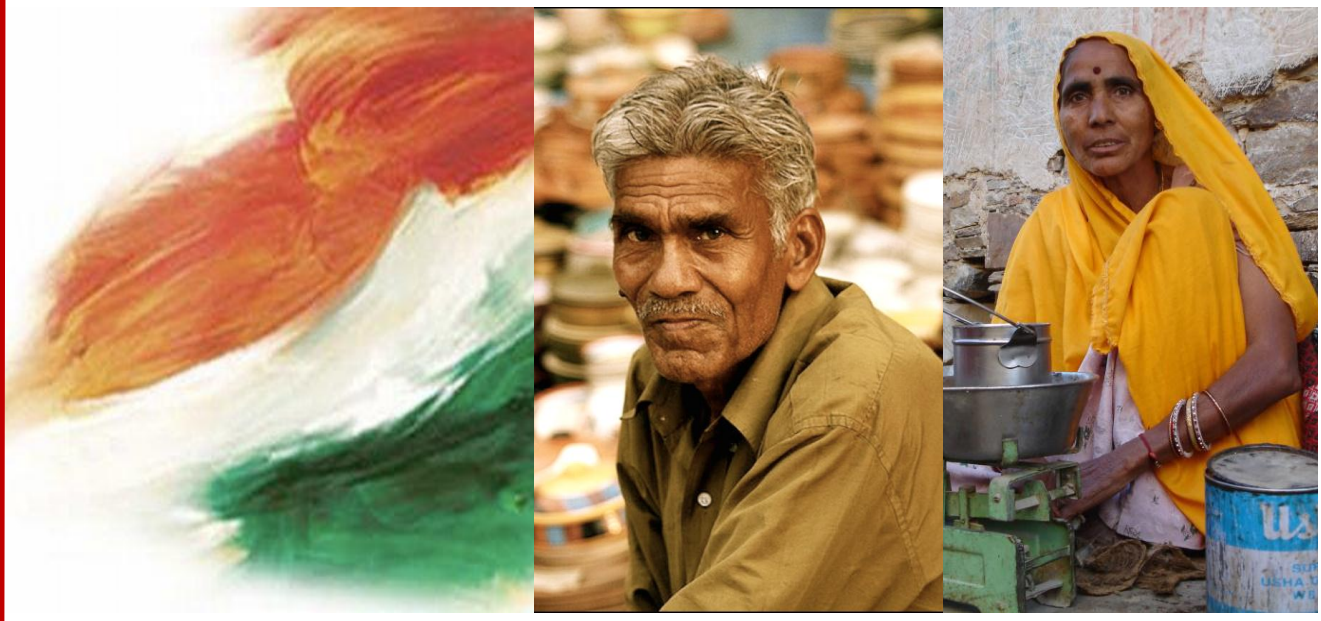
Longitudinal Aging Study in India

The 2010 Pilot Wave

User Guide



June 2012

Harvard School of Public Health

International Institute of Population Sciences

RAND Corporation

Longitudinal Aging Study in India

The Pilot Wave User Guide

This “User Guide” provides background information on the pilot round of the Longitudinal Aging Study in India (LASI) conducted in the fall of 2010. It includes an overview of the survey’s aim and questionnaire content, as well as detailed information about the sampling design of the survey. Lastly, it details a short example from a project using the LASI data to look at gender disparity in health outcomes.

Questions about the data or content of this manual can be sent to kfeeney@rand.org or metadatahelp@rand.org.

Contents

1 About the Survey

LASI is designed to be a longitudinal survey of India's aging population 45 years or older and their spouses (regardless of the spouses' age). The survey's 2010 pilot wave was funded by the National Institute of Aging and administered by Harvard University, the International Institute of Population Sciences (Mumbai, India), and the RAND Corporation, with further support from University of California, Los Angeles and National Institute of AIDS Research (Pune, India).

LASI covers demographic, health, economic, and psychosocial topics relevant to studies on aging. The questionnaire was patterned after that used in the Health and Retirement Survey (HRS) to facilitate cross-national aging research with the United States, and with similar HRS-type surveys in Korea, China, England, and several other European countries. The survey instrument has two parts: the household interview and the individual interview. Each part has several modules, listed below, that parallel those in its sister surveys:

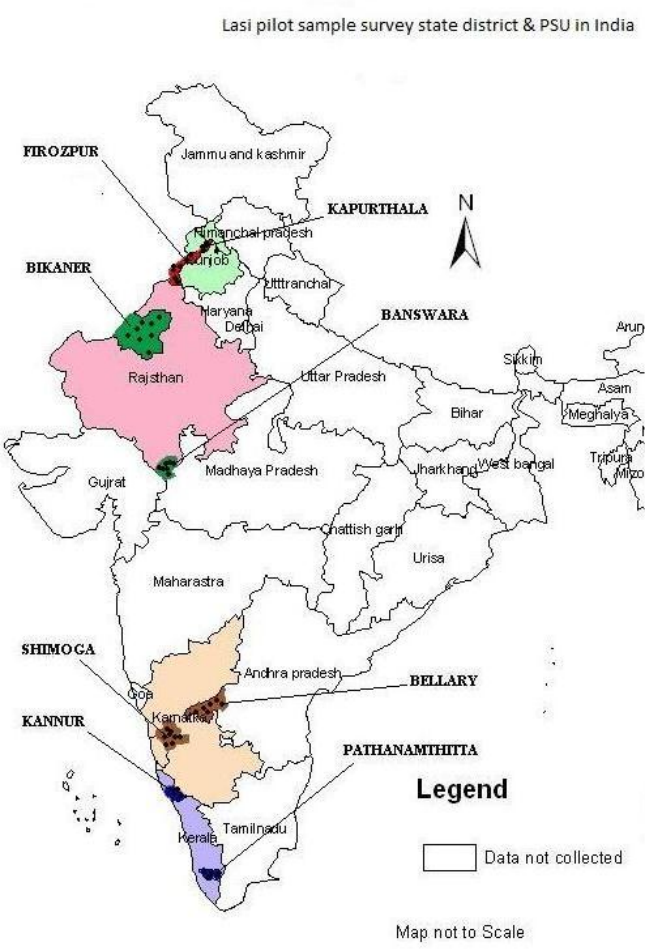
Household Modules	
Coverscreen	Household roster and basic demographic information on all household members
Housing and Environment	Characteristics of the physical dwelling and neighborhood
Consumption	Household expenditure and consumption of food and non-food items, including home-grown goods
Agricultural Income	Household agricultural activity and agricultural income
Non-agricultural Income	Income for household members, from labor and other sources, such as government transfers and remittances
Assets and Debt	Household financial and non-financial assets and debts
Individual Modules	
Demographics	Detailed individual demographics (education, marital status, etc)
Family and Social Networks	Characteristics about the respondent's family, friends, and social networks
Health	Self-reported health status, including physical, functional, emotional, and cognitive health and health behaviors
Health Care	Health insurance and health-care use
Work and Employment	Respondent's employment history
Pension	Respondent's income from and participation in pension plans
Vignettes	An experimental module designed to elicit comparative information about respondent health
Economic Expectations	An experimental module that asks about economic expectations
Social Connectedness	An experimental module that asks about a respondent's social network
Biomarkers	Biomarkers (blood samples), performance test (lung function, grip strength), and anthropometrics (height, weight, waist to hip ratio)

While the survey targets the aging population, the household survey can be completed by any knowledgeable household member at least eighteen years of age, once per household. This often included respondents who participated in the individual survey, but not always. The individual survey was completed by the consenting age-eligible individual and again by his or her spouse regardless of age. In some instances, a knowledgeable household member not otherwise eligible for an individual interview completed a proxy interview for the age-eligible

household member. Proxy interviews did not include biomarkers, psychosocial questions, vignettes, economics expectations, or social connectedness questions.

2 Sampling Plan

LASI uses a complex sampling design. Two districts were randomly chosen from each state. Within these districts eight primary sampling units (PSUs) were chosen to be surveyed. Rural PSUs with fewer than 500 households then used a two-stage sampling procedure, while urban PSUs and rural PSUs with at least 500 households used a three-stage procedure. Thus, the probability of a household being selected for the LASI is the product of the probabilities that (1) it lies within one of the two districts in each of the four states surveyed, (2) it lies within a selected PSU of the selected districts, (3) it is in a selected village/urban census enumeration block (CEB), and (4) it is a selected household within a selected village/urban CEB. Weights are created using the inverse probability of selection combined with household and individual response rates. This type of sampling plan introduces more sample-to-sample variability than simple random sampling, which has important implications for analysis. For example, standard errors on any estimates need to be adjusted to reflect the larger survey error. Nevertheless, we can stratify standard errors on state, district, and urban residence.



The survey was conducted in four Indian States and fielded in the dominant language of that state: Rajasthan and Punjab in the north, and Kerala and Karnataka in the south. Using weights for individuals or households, we can weight analysis to be representative within a particular state, or as pooled across the four states. Given such a heterogeneous sample, researchers must consider whether to pool data from all states or within states. (For a more detailed discussion of how representative the data is of national estimates and other surveys conducted in India, see Arokiasamy et al. 2012).

The sampling plan was based on the latest round of the Indian Census at the time of the survey, the 2001 round. To calculate weights for the data, we used forecast estimate of the 2012 population based of the same census round. We based weights on rural and urban population counts of the population 45 years and older within each state, further adjusting by response rates. We later discuss which weights to use in specific analyses.

3 Accessing the Data

RAND Corporation released the data through the RAND Survey Mega Meta Data Repository in January 2012. This repository is a web-based host of harmonized longitudinal studies from more than 20 countries (including the RAND-harmonized Health and Retirement Data and several other HRS-family surveys). User must create a log -in name and password to download the data from the link below.

https://mmicdata.rand.org/megametadata/?section=study&studyid=36

Once logged in, a user can navigate to the LASI home page under the “Browse Studies” tab.

Clicking on the “download” link will navigate to a new page where users can access the full questionnaire in PDF format as well as the household and individual data sets they need for analysis.

The PDF questionnaire is essential to understanding the LASI data. It contains the original survey questions in English and their question ID. This allows users to match questions to variables in the individual and the household data sets. It also allows users to observe survey flow and skip patterns in the data, e.g., to differentiate between true missing responses and questions for which certain respondents may not have been eligible. Translated questionnaires are available upon request in Hindi, Kannada, and Malayalam.

Assembling Your Data Set

Data Structure

After downloading the individual and household data sets, users must merge them to link data for individuals (e.g., health, health-care use, demographics, labor-force participation) with that for their households (e.g., income, economic status, household demographics).

To do so, first open the household data set and look through the variables. You will notice some important variables related to sampling and survey administration at the top - state, district, PSU, the language of the interview, residence (rural/urban), and the time the interview started and stopped.

The next set of variables is about household members from the coverscreen section of the interview. These include household demographics, such as age, gender, and marital status for each member, and do not directly correspond to questions in the PDF questionnaire. Rather, these were generated by the computer-assisted personal interviewing (CAPI) system, which LASI interviewers used in the field.

For example, *hmembermaritalstatus_10_* is the marital status of the tenth member of the household as reported by the coverscreen respondent. These household member variables are indexed by a number assigned to each household resident. The names of these residents have been removed, but members can be matched across variables by the numerical index tag after the underscore on these sorts of variables. So, if the marital status of the tenth household member is *hmembermaritalstatus_10_*, then his or her gender is *hmembergender_10_*. It can also be used to match demographic information about the household member given in the coverscreen to other information given in a different module, such as that on income. LASI allows up to twenty household members. Household members generally are listed from oldest to youngest although this is not always the case in large, extended-family households. For a definition of who in the household is counted as family, please consult the coverscreen questionnaire.

It is important to note that there is a second set of household demographic variables that correspond to those enumerated in the coverscreen instrument. These are indexed slightly differently, as couples are listed dyadically. In some cases, depending on the type of analysis, it may be easier to work from the coverscreen variables that are enumerated in the instrument. In others, it may be easier to work from the CAPI-generated variables. The variable *spousecvid* can be used to match spouses in the data with the CAPI-generated variables.

The type of indexing described above is used throughout other LASI modules. For example, in the “Family and Social Networks” module, LASI asks several questions about respondents’ children. Each child – whether living at home or not (e.g., whether or not they have been enumerated in the coverscreen survey) – is indexed with a numerical tag. So *fs206_3_* and

*fs207_3_* both refer to the third child mentioned by the respondent. The former indicates whether that child can read or write and the latter whether that child ever attended school.

Similarly, in the social networks experimental module, the indexing refers to individuals the respondent named in their social network. So *es007_1_* indicates the marital status of the first person the respondent listed in their social network and *es006_1_* indicates the age of the first person listed.

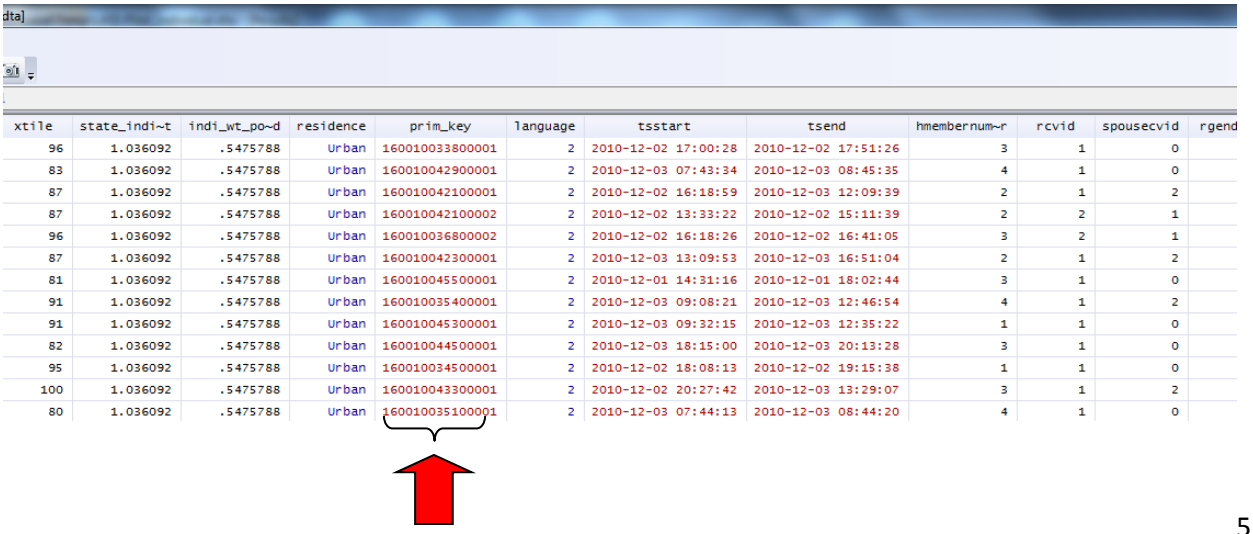
Following the household roster are variables with information about the respondents for the household interview: *financialr, familyr, housingr, consumptionr, mainr, informantr, selectedr.* FinancialR indicates if the respondent answered the income, assets, and agricultural income sections of the household module, FamilyR indicates whether the respondent answered the coverscreen section, HousingR indicates if the respondent answered the housing and environment section, ConsumptionR indicates if the respondent answered the consumption section. MainR indicates who answered the most number of the household interview modules, and SelectedR indicates if the respondent was selected for an individual interview. InformantR is blank and was not fielded. Note in some cases a proxy, non-age-eligible household resident was the respondent for a section of the household module.

Then we have variables from the instrument, enumerated by module (e.g., HE for "Housing and Environment", IN for "Income") and question number (e.g., HE001). Data in the individual data set is organized the same way. Users can use the PDF questionnaire to match questions to variables in the data set.

Merging the data

We want to merge our individual data (organized by module and question number) into our household data. To do this, we need to create a linking variable on which we can merge. That linking variable will be the household ID (HHID), which we will derive from the variable *prim_key* in the individual data set. Prim_Key contains information on the state, sample, household and respondent. We can remove the individual or respondent-level information from this variable so that it is unique to households in both data sets. The household data set already has this HHID variable, so we open the individual data set and create it there.

Below, the arrow shows which portion of the prim_key variable we want to keep. The last digit of the prim_key variable stores respondent-specific information for a given household. To merge households together, we want to remove this last digit.



xtile	state_indi~t	indi_wt_po~d	residence	prim_key	language	tsstart	tsend	hmembernum~r	rcvid	spousecvid	rgend
96	1.036092	.5475788	Urban	160010033800001	2	2010-12-02 17:00:28	2010-12-02 17:51:26	3	1	0	
83	1.036092	.5475788	Urban	160010042900001	2	2010-12-03 07:43:34	2010-12-03 08:45:35	4	1	0	
87	1.036092	.5475788	Urban	160010042100001	2	2010-12-02 16:18:59	2010-12-03 12:09:39	2	1	2	
87	1.036092	.5475788	Urban	160010042100002	2	2010-12-02 13:33:22	2010-12-02 15:11:39	2	2	1	
96	1.036092	.5475788	Urban	160010036800002	2	2010-12-02 16:18:26	2010-12-02 16:41:05	3	2	1	
87	1.036092	.5475788	Urban	160010042300001	2	2010-12-03 13:09:53	2010-12-03 16:51:04	2	1	2	
81	1.036092	.5475788	Urban	160010045500001	2	2010-12-01 14:31:16	2010-12-01 18:02:44	3	1	0	
91	1.036092	.5475788	Urban	160010035400001	2	2010-12-03 09:08:21	2010-12-03 12:46:54	4	1	2	
91	1.036092	.5475788	Urban	160010045300001	2	2010-12-03 09:32:15	2010-12-03 12:35:22	1	1	0	
82	1.036092	.5475788	Urban	160010044500001	2	2010-12-03 18:15:00	2010-12-03 20:13:28	3	1	0	
95	1.036092	.5475788	Urban	160010034500001	2	2010-12-02 18:08:13	2010-12-02 19:15:38	1	1	0	
100	1.036092	.5475788	Urban	160010043300001	2	2010-12-02 20:27:42	2010-12-03 13:29:07	3	1	2	
80	1.036092	.5475788	Urban	160010035100001	2	2010-12-03 07:44:13	2010-12-03 08:44:20	4	1	0	

This is an example of Stata code to remove this last digit to crease a household ID:

```
gen hhid = substr(prim_key, 1, 14)
     destring hhid, replace
     format   hhid  %20.0g
     replace hhid = hhid*10
```

Below, we show the data with both the respondent ID and the household ID. This is the same screen capture as above, now with *hhid* included.

ile	state_indi~t	indi_wt_po~d	residence	prim_key	hhid	language	tsstart
96	1.036092	.5475788	Urban	160010033800001	160010033800000	2	2010-12-02 17:00:
83	1.036092	.5475788	Urban	160010042900001	160010042900000	2	2010-12-03 07:43:
87	1.036092	.5475788	Urban	160010042100001	160010042100000	2	2010-12-02 16:18:
87	1.036092	.5475788	Urban	160010042100002	160010042100000	2	2010-12-02 13:33:
96	1.036092	.5475788	Urban	160010036800002	160010036800000	2	2010-12-02 16:18:
87	1.036092	.5475788	Urban	160010042300001	160010042300000	2	2010-12-03 13:09:
81	1.036092	.5475788	Urban	160010045500001	160010045500000	2	2010-12-01 14:31:
91	1.036092	.5475788	Urban	160010035400001	160010035400000	2	2010-12-03 09:08:
91	1.036092	.5475788	Urban	160010045300001	160010045300000	2	2010-12-03 09:32:
82	1.036092	.5475788	Urban	160010044500001	160010044500000	2	2010-12-03 18:15:

After creating the household identifier, we are ready to merge the data. We destring HHID because it is stored as a numeric float in the household file. Recall we are merging an individual data set where multiple individuals are interviewed per household into a household data set where the unit of observation is the household:

```
merge 1:m hhid using LASI_pilot_individual.dta
drop _m
```

4 Analyzing the Data

Now, you're (almost) ready to analyze the merged data. For this exercise, we examine cognitive outcomes and how they vary by socioeconomic status and gender in India. This showcases some unique features of HRS-family data sets, as well as some specific to LASI. Let's think of a research question…

Research question: Cognitive health is a growing concern in developing countries, particularly as their societies age (Prince, 1997). Unlike in developed counties, women tend to do worse than men on tests of cognitive functioning (Zunzunegui et al, 2008). Previous research suggests that this is in part because women are not given equal access to education (Maurer, 2011). We test that hypothesis in this example, seeking to answer the question how much does education account for gender disparity in cognitive functioning? [1]

[1] This research question was the basis of a study by Lee et al. (2011).

Derive Variables
The first step is to derive variables of interest for analysis.

- What variables might affect cognitive outcomes for men and women? Education is the primary explanatory variable, but age is another. Can you think of others?

- Socioeconomic status is known to influence individual health as well, so we should adjust for it in our analysis. How can we measure socioeconomic status in India? Caste? Household consumption?

- What variable will we need to estimate the correct standard errors given the LASI survey design?

These questions are meant to guide your thinking about variables to derive in Stata or another statistical package for analysis. Below, we walk through some Stata syntax used to create these variables. We first derive variables we will need to estimate correct standard errors with complex survey data. We then create our dependent variable, total word recall, which is a measure of cognitive functioning based on episodic memory questions in LASI. This measure has two parts: immediate word recall, in which the interviewer reads aloud ten words and asks the respondent to repeat as many of them as possible, and delayed recall, in which the interviewer asks the respondent to repeat as many of the same ten words as as possible at the conclusion of the cognitive functioning battery. Following these two variables, we provide code for education, caste, gender, age, and other variables we might want to examine in our analysis.

Survey Variables

```
* STRATA
egen strata = concat (state district residence)
      label var strata                                                    "Strata"
```

Dependent variable

Cognitive Measures

```
* EPISODIC MEMORY
* DELAYED WORD RECALL

gen delrecall = .
      replace delrecall = ht520 if ht520>=0 & ht520<11
      label var delrecall                              "Delayed Word Recall"

* IMMEDIATE WORD RECALL
gen immrecall = .
      replace immrecall = ht505 if ht505>=0 & ht505<11
      label var immrecall                                "Imm. Word Recall"

* TOTAL WORD RECALL
gen totwordrecall = immrecall + delrecall
      label var wordrecall            "wordrecall: Total Word Recall Memory"
```

Explanatory Variables

Among other explanatory variables to include in analysis of cognitive outcomes are those for demographics, education, and socioeconomic status. Below we discuss coding for such variables as age, caste, and education.

Demographics
Age is reported twice in the LASI survey. Below we use age from the demographic module in the individual interview. Age is also reported in the coverscreen as part of the household roster (*rage*). In some cases these age variables may not match, and researchers must decide which variable is appropriate for their analysis.

Below, we show how to calculate age at the time of the interview based on the month of the interview. We first create a variable for month of interview from the variable *tsend*, which stores the time the interview ended. In some cases, as might be expected in low literacy populations, individuals could not report their birth date, and instead reported their age in years directly. In others, respondents could not report their birthdate or estimate their age, so their age is left coded as missing.

```
* AGE - CONTINUOUS
* GENERATE MONTH OF INTERVIEW
gen inw_month  = substr(tsend, 6, 2)
destring inw_month, replace
label var inw_month                              "inw_month: Interview Month"

gen age = .
      replace age = ((2010*12 + inw_month) - (dm007_year*12  +
      dm007_month))/12
      replace age = dm008 if mi(age)

      label var age              "Age: Respondents Self Reported Age (years)"

* AGE - CATEGORICAL
qui su age
qui return list
local maxage = r(max)

egen agec = cut(rage), at(45, 55, 65, 75, `maxage')
label var age_cat                                  "age_cat: Age Categories"


* GENDER: FEMALE DUMMY VARIABLE
gen female  = .
      replace female = 1 if dm002==2
      replace female = 0 if dm002==1

      label var female                              "Female: Female Gender"
```

Gender is also reported twice in the interview: once as part of the coverscreen and again as part of the individual demographic module (*rgender*). Note that many sections of the individual interview can also be completed by a proxy respondent who answers for the age-eligible respondent.

```
* URBAN/RURAL STATUS
gen urban = .
      replace urban = 1 if residency ==1
      replace urban = 0 if residency ==2

      label var urban                              "urban: Urban residency"
```

Urban residency status is recoded from the variable *residency*, which is coded 1 if urban residency and 2 for rural residency. We plan to use this variable in a regression later in our analysis and thus create a binary variable for urban status, rather than use the original residency variable. We used the same approach for gender above. Below we create some dummy variables for state of residence.

```
* STATE VARIABLE
tab state, gen(state_)

            label var state_1                          "state_1: Punjab State"
            label var state_2                       "state_2: Rajasthan State"
            label var state_3                          "state_3: Kerala State"
            label var state_4                       "state_4: Karnataka State"
```

Next is our recoding and relabeling of the *education* and *literacy* variables. LASI asks a series of questions about educational attainment and literacy, starting with DM029 which asks the respondents if they have ever attended school. Those who indicate they have attended school and received some formal education are asked how many years of education they have received and the highest degree or level they have completed. Respondents who indicate they do not have any formal education skip these questions.

This is an example of a skip pattern in the survey, which is made possible by CAPI's adaptive interviewing. Below, we need to set the years of education variable to 0 for those respondents who did not have any formal education and thus did not answer question DM030 about years of education. While they're response shows up as missing, we can better code it in our data as 0 years of education.

```
* YEARS OF EDUCATION
gen educyrs = .
      replace educyrs = 0 if dm029==2          ///Replace years of education
                                                to 0 if never attended school
      replace educyrs = dm030 if mi(educyrs)

      label var educyrs                         "educyrs: Years of Education"

* EDUCATION, CATEGORICAL
gen educ = .
      replace educ = 0 if educyrs==0
      replace educ = 1 if inrange(educyrs, 1, 5)
      replace educ = 2 if inrange(educyrs, 6, 25)

      label var educ                                  "Education Categories"


* LITERACY
recode dm028 (1 2 4 =0 "Not fully literate") (3 = 1 "Literate"),
gen(literacy)
      label var literacy                          "literacy: Literacy Status"
```

Below is our recoding and relabeling of socioeconomic status (SES) measures, including caste and marital status.

SES Measures

```
* SOCIOECONOMIC STATUS
* CASTE IN INDIA[2]
gen caste = .
      replace caste = dm034
      replace caste = 4 if dm033==3
      label var caste                                          "caste: Caste"

tab caste, gen(caste_)
      label var caste_1                                "caste_1: Scheduled Caste"
      label var caste_2                                "caste_2: Scheduled Tribe"
      label var caste_3                                     "caste_3: OBC"
      label var caste_4                               "caste_4: Other/No caste"


* MARITAL STATUS
recode dm003 (1 3 4 5 = 0 "Not married") (2 =1 "Married"), gen(married)
      label var married                                                "Married"

* HOUSEHOLD CONSUMPTION
* USE hhexptotalpc AND f_hhexptotal
```

LASI includes both *household consumption* and household income. We use consumption for this analysis because in developing countries it is often a better measure of socioeconomic status (Strauss et al, 2010). For income and consumption measures, LASI imputes missing data, though we did not impute ownership. LASI might ask a series of questions about consumption of durable goods, like a car. Respondents are first asked if their household has purchased a car (ownership), and then in a subsequent question are asked about the purchasing price of the car (value). With this example, missing data are only imputed for households that indicate they have purchased a car, but do not have a response for the value. We used a simple hotdeck imputation procedure. We did not impute whether the household purchased a car or not if ownership is missing. To facilitate analysis with this imputed data, LASI programmers created a flag variable (f_) to indicate cases where there was imputation for at least one component of that measure. Missing responses in the flag indicate households where ownership was missing and household income and consumption could not be derived.

Note that the variable we use above in our analysis is not the total household consumption, but the *per capita* household consumption. To derive this measure (and its equivalent for household income), LASI programmers used the OECD equivalency scale which differentially weights household members by age: household heads are weighted with 1, additional adults 0.7, and children (under 16) by 0.3.

Other Variables: Above we noted that information about an eligible survey subject may be available in more than one module and thus may be available from multiple respondents. For example, if a non-age-eligible household member answered the coverscreen, we have age of the individual respondent both from him or her and the respondent of the individual interview himself. This occurs for subject matter that is reported in both the household and individual interview, e.g. income, pension income, and labor market status, available for the individual

[2] For more information about caste in India, please see Subramanian et al. (2008) referenced in the citation section of this users' guide.

respondent in both the “Employment and Pension” (EP) module and “Individual Income” (IN) module in the household interview. (This is also true for some demographic information of the age-eligible respondent, as we noted above). The user needs to decide which variables might be the best to use in his or her analysis.

Descriptive Statistics

Below we look at some descriptive statistics and bivariate associations between our outcome of interest, cognition, and demographic and SES variables among our pool of respondents.

We must first account for the LASI survey design in Stata. This is needed to calculate accurate standard errors. In doing this, we also need to consider statistical issues related to survey sampling, such as occurrences of a single primary sampling unit per strata.

Below are four commands that set weights in Stata and tell it how to calculate the standard errors. The first two, under “For households”, are (often) used when the level of analysis is the household. We have two options here. The pooled option, listed first, is used to create a sample representative of the population across the four LASI states. The second is used to examine households within a state.

The second two, under “For individuals”, are used when the level of analysis is the respondent level. The first is for constructing pooled weights used to make a sample representative across the four states. The second is used to examine a representative group of individuals within a particular state.

- For households

```
svyset psu [w= hh_wt_pooled], strata(strata)  singleunit(scaled)
svyset psu [w= state_hh_wt],  strata(strata)  singleunit(scaled)
```

- For individuals

```
svyset psu [w= indi_wt_pooled], strata(strata) singleunit(scaled)
svyset psu [w= state_indi_wt],  strata(strata) singleunit(scaled)
```

For now, we will use the individual pooled weight. Using the `svydes` command we can see some important information about the structure of the survey data, including a common problem to many surveys, single PSU per stratum. There are various ways to deal with this – we chose to adjust for this using the singleunit(scaled) commands available in some Stata versions.

```
. svydes

Survey: Describing stage 1 sampling units

      pweight: indi_wt_pooled
          VCE: linearized
  Single unit: scaled
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>

                                        #Obs per Unit
                                -----------------------------
Stratum    #Units     #Obs      min       mean      max
--------  --------  --------  --------  --------  --------
```

```
      111          2         41         15       20.5         26
      112          5        132         18       26.4         35
      121          3         75         23       25.0         29
      122          5        154         26       30.8         36
      231          3         66         15       22.0         28
      232          5        146         23       29.2         33
      241         1*         22         22       22.0         22
      242          7        183         19       26.1         32
      351          4        106         20       26.5         32
      352          4        112         16       28.0         36
      361         1*         28         28       28.0         28
      362          7        216         19       30.9         38
      471          3         60         18       20.0         23
      472          5        117         19       23.4         26
      481          3         74         21       24.7         29
      482          5        151         26       30.2         33
--------  ---------  ---------  ---------  ---------  ---------
       16         63       1683         15       26.7         38
```

We limit our analysis here to individuals at least 45 years old, excluding younger spouses. See if you can reproduce the following table on your own. How might you test the differences to see if they are significant?

Respondent Characteristic	Mean Cognition Score [SE]
Age	
45 – 54	9.3 [0.20]
55 – 64	8.9 [0.24]
65 – 74	7.5 [0.35]
75+	5.9 [0.45]
Gender	
Male	9.1 [0.17]
Female	8.1 [0.19]
Education	
No Schooling	7.5 [0.24]
Primary	8.1 [0.29]
More than Primary	10.2 [0.18]
Caste	
SC	8.3 [0.32]
ST	7.0 [0.53]
OBC	8.9 [0.23]
Other/None	9.0 [0.20]

Above we only selected a few variables to review, but we created several more, including other measures of socioeconomic status. We chose to create categorical variables from some "continuous" measures, such as age and years of education for the purpose of descriptive analysis. Do we see similar gradients with cognition across SES when we use other measures, like consumption?

Given heterogeneous cultural and socioeconomic patterns across India, it is also important to examine regional differences. The four states in the LASI sample are very different from each other. For example, Kerala is very poor, yet has the highest rates of education and literacy in India. Punjab in the north is more economically developed, and reflects a unique religious composition. Pooling such heterogeneous populations may not be helpful. See if you can

replicate the following table on your own. Be sure to use the right survey commands and correct weights for analyzing representative samples within states.

State	Cognition Score [SE]
Punjab	10.5 [0.24]
Rajasthan	7.5 [0.37]
Karnataka	9.3 [0.28]
Kerala	7.9 [0.18]

Multivariate Analysis

A next step might be to use multivariate analysis to understand better the connection among cognitive health, gender, and education. We do so using an ordinary least squares regression analysis. Such analyses require answering methodological questions such as whether to weight the data.

In the example below, we do not weight the data, but we do adjust for the greater sample-to-sample variability in our field work and stratify the standard errors with the Stata syntax shown here. We continue to limit the analysis to respondents at least 45 years old.

```
svyset psu, strata(strata) singleunit(scaled)
```

For the first model, we see how cognitive health varies by gender, adjusting for age and region. We also adjust for caste – a variable unique to India and the LASI data set. See if you can generate the same output below.

```
. svy, subpop(if age>=45 & age<.): reg totwordrecall age female urban state_1 state_2 state_3
> /*
> */                                          caste_1 caste_2 caste_3
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        16                 Number of obs      =      1618
Number of PSUs     =        62                 Population size    =      1618
                                               Subpop. no. of obs =      1359
                                               Subpop. size       =      1359
                                               Design df          =        46
                                               F(   9,     38)    =     24.05
                                               Prob > F           =    0.0000
                                               R-squared          =    0.2434

------------------------------------------------------------------------------
             |             Linearized
totwordrec~l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0947752   .0099375    -9.54   0.000    -.1147784    -.074772
      female |  -1.125589   .1676466    -6.71   0.000    -1.463044   -.7881335
       urban |   .9115398   .2425097     3.76   0.000     .4233931    1.399686
     state_1 |   1.355013   .3763106     3.60   0.001     .5975393    2.112487
     state_2 |  -1.177952    .439407    -2.68   0.010    -2.062432   -.2934713
     state_3 |  -1.148678   .3300816    -3.48   0.001    -1.813098   -.4842582
     caste_1 |  -1.213559   .3898728    -3.11   0.003    -1.998332   -.4287853
     caste_2 |  -1.478517   .5846452    -2.53   0.015    -2.655347   -.3016874
     caste_3 |  -.1981929   .2777761    -0.71   0.479    -.7573272    .3609414
       _cons |   15.29761   .7530446    20.31   0.000     13.78181    16.81341
------------------------------------------------------------------------------
Note: variance scaled to handle strata with a single sampling unit.
```

We see some differences by gender even after adjusting for age, region, and caste. But how much of this is explained by access to education, a strong predictor of cognitive development and other health outcomes (Cagney & Lauderdale, 2002)? In the next model, we control for education and literacy. Because formal schooling is rare among these cohorts, we adjust for literacy as well.

```
. * MODEL 2: COGNITION AND EDUCATION IN INDIA
. svy, subpop(if age>=45 & age<.): reg totwordrecall age female urban state_1 state_2 state_3 /*
> */                                  caste_1 caste_2 caste_3 educyrs literacy
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        16                  Number of obs      =      1617
Number of PSUs     =        62                  Population size    =      1617
                                                Subpop. no. of obs =      1358
                                                Subpop. size       =      1358
                                                Design df          =        46
                                                F(  11,     36)    =     37.72
                                                Prob > F           =    0.0000
                                                R-squared          =    0.3290

------------------------------------------------------------------------------
             |             Linearized
totwordrec~l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0732616   .0091564    -8.00   0.000    -.0916926   -.0548306
      female |  -.6007529    .160469    -3.74   0.001    -.9237603   -.2777455
       urban |   .3353038   .2165333     1.55   0.128    -.1005551    .7711626
     state_1 |   1.779706   .3439286     5.17   0.000     1.087413    2.471998
     state_2 |  -.6442459    .378738    -1.70   0.096    -1.406606    .1181142
     state_3 |  -1.928754   .2978708    -6.48   0.000    -2.528337   -1.329171
     caste_1 |  -.3658188   .4074407    -0.90   0.374    -1.185954    .4543169
     caste_2 |  -.6353194   .6245851    -1.02   0.314    -1.892544    .6219052
     caste_3 |   .1262572   .2605282     0.48   0.630    -.3981589    .6506732
     educyrs |    .290788   .0392203     7.41   0.000     .2118417    .3697344
    literacy |  -.2516348   .3628398    -0.69   0.491    -.9819934    .4787238
       _cons |   12.41815   .7320625    16.96   0.000     10.94459    13.89172
------------------------------------------------------------------------------
Note: variance scaled to handle strata with a single sampling unit.
```

In the model above, we see access to education accounts for some of the disadvantage Indian women have in cognitive ability. Below we adjust for household per capita consumption and marital status. Socioeconomic status and social environment, both of which can affect cognitive health specifically and health generally, might also account for the disparity we observe.

```
. * MODEL 3: COGNITION AND EDUCATION IN INDIA
. svy, subpop(if age>=45 & age<.): reg totwordrecall age female urban state_1 state_2 state_3 /*
> */                                  caste_1 caste_2 caste_3 educyrs literacy    /*
> */                                  hhexptotalpc f_hhexptotal rmarried
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        16                  Number of obs      =      1617
Number of PSUs     =        62                  Population size    =      1617
                                                Subpop. no. of obs =      1358
                                                Subpop. size       =      1358
                                                Design df          =        46
                                                F(  14,     33)    =     31.65
                                                Prob > F           =    0.0000
                                                R-squared          =    0.3348

------------------------------------------------------------------------------
             |             Linearized
totwordrec~l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0679697   .0086493    -7.86   0.000    -.0853799   -.0505596
      female |  -.4740033   .1549292    -3.06   0.004    -.7858596   -.1621469
       urban |    .388909   .2216448     1.75   0.086    -.0572388    .8350568
     state_1 |   1.648629   .3549047     4.65   0.000     .9342434    2.363016
     state_2 |  -.6313993   .3603602    -1.75   0.086    -1.356767    .0939681
     state_3 |  -1.958521   .3056301    -6.41   0.000    -2.573723   -1.343319
     caste_1 |  -.3123355   .3996644    -0.78   0.439    -1.116818    .4921471
     caste_2 |  -.6710674   .6264951    -1.07   0.290    -1.932137    .5900019
     caste_3 |   .0902314   .2604799     0.35   0.731    -.4340874    .6145502
     educyrs |   .2831212   .0393048     7.20   0.000     .2040047    .3622377
    literacy |  -.2325218   .3755967    -0.62   0.539    -.9885588    .5235152
hhexptotalpc |   3.66e-07   1.14e-06     0.32   0.750    -1.93e-06    2.66e-06
f_hhexptotal |  -.5291748   .2794533    -1.89   0.065    -1.091685    .0333355
    rmarried |   .5053711   .2313104     2.18   0.034     .0397675    .9709748
       _cons |   11.76776   .7112575    16.55   0.000     10.33608    13.19945
------------------------------------------------------------------------------
Note: variance scaled to handle strata with a single sampling unit.
```

In the output above, we see we are able to explain about half of the “female disadvantage” observed in our first model. We see education accounts for the disparity by caste, but not by gender. Social standing might also have an effect on cognitive status as well.

This is a basic model which we can use with LASI to answer other questions, such as: What other measures of socioeconomic status or social standing might explain variation in cognitive health? What other respondent characteristics might confound the relationships in the model? How might we control for important contributors to disparities such as comorbidities and health-service utilization? What might explain the regional variation we see? On your own, explore other model specifications to become familiar with the data. Be sure to ask questions.

References

Arokiasamy, P., Lee, J., Bloom, D., Feeney, K., & Ozolins, M. (in press, 2012) Longitudinal Aging Study in India: Vision, Design, and Implementation. In *Aging in Asia: Findings from New and Emerging Data Initiatives.* Committee on Policy Research and Data Needs to Meet the Challenge of Aging in Asia. J.P. Smith and M. Majmundar, Eds. Washington, DC: The National Academies Press.

Cagney, K.A., & Lauderdale, D.S. (2002). Education, Wealth, and Cognitive Function in Later Life. *Journal of Gerontology: Psychological Sciences*, 57B (2):163-172.

Lee, J., Shih, R., Feeney, K., & Langa, K. (2011) Cognitive Health of Older Indians: Individual and geographic determinants of female disadvantage. RAND Working Paper Series WR-889.

Maurer, J. (2011). Education and Male-Female Differences in Later-life Cognition: International Evidence from Latin America and the Caribbean, *Demography*, 48 (3): 915 – 930.

Prince, M. (1997). The Need for Research on Dementia in Developing Countries. *Tropical Medicine and International Health*; 2(10); 993-1000.

Strauss, J., Lei, X., Park, A., Shen, Y., Smith, J.P., Yang, Y., & Zhao, Y. (2010). Health Outcomes and Socio-economic Status among the Elderly in China: Evidence from the CHARLS Pilot, *RAND Working Paper WR-774*, RAND Corporation: Santa Monica, CA.

Subramanian, S.V., Ackerson, L.K., Subramanyam, M.A., & Sivaramakrishnan, K. (2008). Health Inequalities in India: The Axes of Stratification, *The Brown Journal of World Affairs*, 14 (2): 127-138.

Zunzunegui, M.V., Alvarado, B.E., Beland, F., & Vissandjee, B. (2008). Explaining health differences between men and women in later life: A cross-city comparison in Latin America and the Caribbean. *Social Science and Medicine*, 68: 235-242.

Helpful Links

Harvard Program on the Demography and Aging:
http://www.hsph.harvard.edu/pgda/lasi.html

International Institute of Population Science (IIPS Mumbai):
http://www.iipsindia.org/

RAND Survey Meta Data Repository:
https://mmicdata.rand.org/megametadata/