Nonresponse and Data Quality Control

Survey Research Design and Analysis

Soledad Artiz Prillaman
Oxford University

May 2, 2018

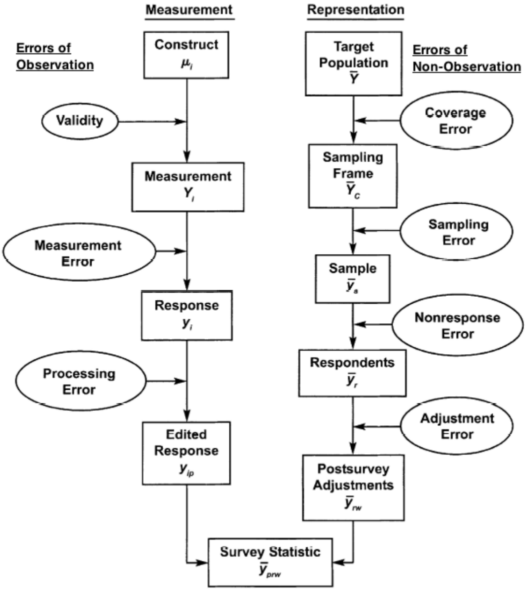
OUTLINE

Nonresponse

Item Nonresponse

Survey Interviewing

Quality Control


Measurement
Representation
Errors of Observation
Errors of Non-Observation
Construct μ_i
Validity
Measurement Y_i
Measurement Error
Response y_i
Processing Error
Edited Response y_{ip}
Target Population $\bar{Y}$
Coverage Error
Sampling Frame $\bar{Y}_C$
Sampling Error
Sample $\bar{y}_s$
Nonresponse Error
Respondents $\bar{y}_r$
Adjustment Error
Postsurvey Adjustments $\bar{y}_{rw}$
Survey Statistic $\bar{y}_{prw}$

REPRESENTATION

Target Population: the set of units to be studied.

Sampling Frame: the set (lists and procedures) used to identify the elements of the target population.Identifies the set of target population members that has a chance to be selected into the survey sample.

Sample: the set of elements from the sampling frame selected to participate in the survey.

Respondents: the individual elements from the sample for which data are successfully measured.

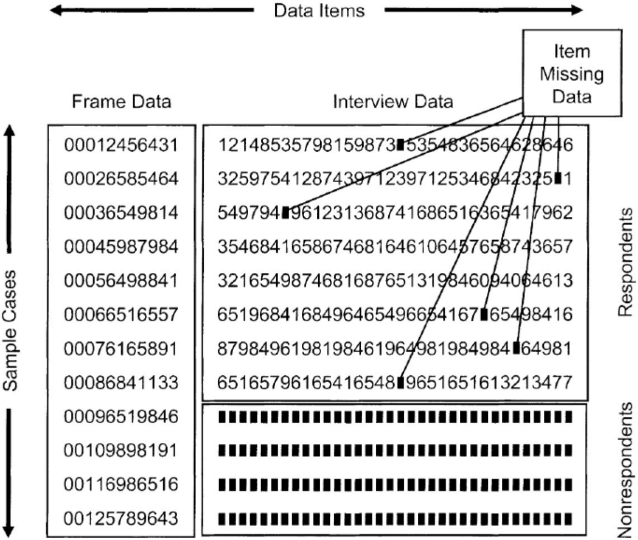
Postsurvey Adjustments: adjustments to individual data made because of concerns around the sample.Ex: weighting, imputation, etc.

NONRESPONSE

Two types:

- Unit nonresponse
  - Noncontacts
  - Refusals
  - Inability to participates
- Item nonresponse

NONRESPONSE



Frame Data	Interview Data
00012456431	12148535798159873■5354836564628646
00026585464	32597541287439712397125346842325■1
00036549814	549794■961231368741686516365417962
00045987984	3546841658674681646106457658743657
00056498841	3216549874681687651319846094064613
00066516557	6519684168496465496654167■65498416
00076165891	879849619819846196498198498 4■64981
00086841133	65165796165416548■9651651613213477
00096519846	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
00109898191	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
00116986516	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
00125789643	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■

NONRESPONSE

Nonresponse is a missing data problem!

Nonresponse Error

Nonresponse Error: when the respondents do not fully and accurately represent the sample.

Function of:

- Noncontacts
- Refusals
- Inability to participates

Nonresponse Bias

	Size	Total	Mean	Variance
Respondents				
Nonrespondents				
Sample	N	t	$\bar{y}$	S^2

Nonresponse Bias

	Size	Total	Mean	Variance
Respondents	N_R			
Nonrespondents	N_M			
Sample	N	t	$\bar{y}$	S^2

Nonresponse Bias

	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\bar{y}_R$	S_R^2
Nonrespondents	N_M			
Sample	N	t	$\bar{y}$	S^2

Nonresponse Bias

	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\bar{y}_R$	S_R^2
Nonrespondents	N_M	t_M	$\bar{y}_M$	S_M^2
Sample	N	t	$\bar{y}$	S^2

Nonresponse Bias

$$\bar{y} = \frac{N_R}{N}\bar{y}_R + \frac{N_M}{N}\bar{y}_M$$

$$\text{Bias} = E[\bar{y}_R] - \bar{y} \approx \frac{N_M}{N}(\bar{y}_R - \bar{y}_M)$$

When will the bias be small? (1) $\bar{y}_M \approx \bar{y}_R$ and (2) $\frac{N_M}{N}$ is small

CAUSES OF NONRESPONSE

- Sensitive survey content
- Time/context of data collection
- Interviewers
- Data-collection method
    - In person tends to get better response rates
- Questionnaire design (wording, visuals)
- Respondent burden/length of survey
    - Can split survey and run with subsamples
- Survey introduction
- Incentives
- Follow-up!

Dealing with Nonresponse

1. Prevent it!
2. Representative subsample of nonrespondents
3. Predict values for nonrespondents
4. Ignore it (don't do this!)

Dealing with Nonresponse

Some nonresponse is inevitable. The most important thing you can do is:

collect as much information about nonrespondents and reasons for nonresponse as possible!

OUTLINE

Nonresponse

Item Nonresponse

Survey Interviewing

Quality Control

Item Nonresponse

Item Nonresponse: when a response to a single question is missing.

- Frequent in sensitive questions

Nonresponse Bias: When the likelihood of responding is correlated with variables of interest.

- Nonresponse is not inherently problematic.
- Will only affect statistics that involve the particular variable.

Item Nonresponse: An Example

ccexpend	age	income	homeowner
124.98	38	4.52	1
🟥	33	2.42	0
15.00	34	4.50	1
🟥	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Item Nonresponse: An Example

Load the data in R:

```
## First set your working directory
setwd("c:/[your working directory]")

# Load our packages
library(Zelig)
library(Amelia) # We will use this for multiple
    imputation

# Load in credit card data with missing values
cc.missing <- read.csv("ccarddata_missing.csv")
```

Load the data in Stata:

```
** First set your working directory
cd "c:/[your working directory]"

** Load in credit card data with missing values
import delimited "ccarddata_missing.csv"
```

Item Nonresponse

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds;} \\ 0 & \text{if unit } i \text{ does not respond.} \end{cases}$$

$$\phi_i = P(R_i = 1)$$

- X_i - vector of data known about respondents **and** nonrespondents
- y_i - vector of data known about only respondents

Understanding Item Nonresponse

How was this missingness generated?

- **Missing completely at random (MCAR)**: missingness purely random; unrelated to variables in data or any unobserved variables
  - $\mathrm{Cov}(\phi_i, y_i) = \mathrm{Cov}(\phi_i, X_i) = 0$
  - $\mathrm{E}[Y_R] = \mathrm{E}[Y_M]$; i.e. unbiased estimates
- **Missing at random (MAR)**: missingness related to *observed* data
  - $\mathrm{Cov}(\phi_i, X_i) \neq 0$ but $\mathrm{Cov}(\phi_i, y_i | X_i) = 0$
  - $\mathrm{E}[Y_R | X] = \mathrm{E}[Y_M | X]$; i.e. conditionally unbiased estimates
- **Nonignorable/ Not missing at random (NMAR)**: missingness related to *unobserved* data
  - $\mathrm{Cov}(\phi_i, y_i) \neq 0$
  - Generally biased

Understanding Item Nonresponse in Our Data

How can we characterize the missingness in our credit card data?

We can break the data into two groups: observations with missing values and observations without missing values.

Then we can summarize our data for each of these groups *separately*. What we want is balance - for our missing data group to look the same as our non-missing data group.

What would we expect if we believe our missingness is **MCAR? MAR? NMAR?**

BALANCE TABLE IN R

```
# create a variable that tells us which observations
# have a missing value for ccexpend
cc.missing[is.na(cc.missing$ccexpend),]$missing <-1

# For each covariate we use the command t.test
age.ttest <- t.test(cc.missing[cc.missing$missing==1, "
    age"],cc.missing[cc.missing$missing==0,"age"])
income.ttest <- t.test(cc.missing[cc.missing$missing==1,
     "income"],cc.missing[cc.missing$missing==0,"income"
    ])
homeowner.ttest <- t.test(cc.missing[cc.missing$missing
    ==1, "homeowner"],cc.missing[cc.missing$missing==0,"
    homeowner"])

# Let's create a vector of these test statistics
t.stats <- c(age.ttest$statistic, income.ttest$statistic
    , homeowner.ttest$statistic)
t.stats
```

Balance Table in Stata

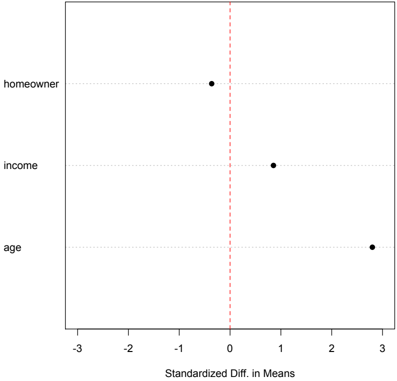
```
* create a variable that tells us which observations
* have a missing value for ccexpend
generate missing = 0
replace missing = 1 if ccexpend == .

* For each covariate we use the command t.test
ttest age, by(missing)
ttest income, by(missing)
ttest homeowner, by(missing)
```

Understanding Item Nonresponse in Our Data

	$\bar{X}_{\text{missing}}$	$\bar{X}_{\text{non-missing}}$	$\bar{X}_{\text{missing}} - \bar{X}_{\text{non-missing}}$	t-stat
`age`	33.68	29.13	4.54	2.80
`income`	3.62	3.27	0.34	0.85
`homeowner`	0.35	0.39	-0.04	-0.36

Item Nonresponse in Our Data

We can also look at the missingness mechanism graphically:



Does it look like we have balance across the covariates?

DEALING WITH ITEM NONRESPONSE

We deal with item nonresponse by imputing the missing values.

A few common ways to deal with missingness:

- Complete case analysis (listwise deletion)
- Mean imputation
- Regression imputation
- Multiple imputation

Complete Case Analysis

ccexpend	age	income	homeowner
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R:

```
# R automatically row-deletes observations with missing
   data
lm(ccexpend ˜ income + homeowner + age, data=cc.missing)
Zelig(ccexpend ˜ income + homeowner + age,
    data=cc.missing, model="normal")
```

Complete Case Analysis

ccexpend	age	income	homeowner
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In Stata:

```
* Stata automatically row-deletes observations with
   missing data
reg ccexpend income homeowner age
```

Complete Case Analysis

What are the consequences of complete case analysis if we have:

MCAR? Unbiased inference but fewer observations

MAR? Unbiased inference *if* missingness covariates included (**ignorable**) but fewer observations

NMAR Possibly biased inference and fewer observations

Main Concern: Likely to induce bias unless MCAR.

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\bar{y}$	33	2.42	0
15.00	34	4.50	1
$\bar{y}$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R:

```
cc.meanimpute <- cc.missing
cc.meanimpute[cc.meanimpute$missing==1,]$ccexpend <-
    mean(cc.missing$ccexpend, na.rm=TRUE)
mean.imputation <- zelig(ccexpend ~ income + homeowner +
     age, data=cc.meanimpute, model="normal")
```

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\bar{y}$	33	2.42	0
15.00	34	4.50	1
$\bar{y}$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In Stata:

```
generate ccexpend_meanimp = ccexpend
sum ccexpend
local ccmean = r(mean)
di ‘ccmean’
replace ccexpend_meanimp = ‘ccmean’ if missing == 1
regress ccexpend_meanimp income homeowner age
```

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
209.45	33	2.42	0
15.00	34	4.50	1
209.45	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Mean Imputation

What are the consequences of regression imputation if we have:

MCAR? Unbiased inference

MAR? Possibly biased inference

NMAR? Possibly biased inference

Main Concern: Distorted distribution of imputed variable, unbiased mean but underestimated standard deviation, attenuated correlations. Better for univariate than multivariate statistics.

REGRESSION IMPUTATION

We can model missingness!

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\hat{y}_2$	33	2.42	0
15.00	34	4.50	1
$\hat{y}_4$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R, we can predict missing values:

```
cc.regimpute <- cc.missing
predict.model <- lm(ccexpend ~ income + homeowner + age,
     data=cc.missing)
cc.regimpute[cc.regimpute$missing==1,]$ccexpend <-
    predict(predict.model,missing.data.frame)
reg.imputation <- zelig(ccexpend ~ income + homeowner +
    age, data=cc.regimpute, model="normal")
```

Regression Imputation

We can model missingness!

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\hat{y}_2$	33	2.42	0
15.00	34	4.50	1
$\hat{y}_4$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In Stata, we can predict missing values:

```
generate ccexpend_regimp = ccexpend
regress ccexpend income homeowner age
predict ccexpend_predict, xb
replace ccexpend_regimp = ccexpend_predict if missing ==
     1
regress ccexpend_regimp income homeowner age
```

Regression Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
94.41	33	2.42	0
15.00	34	4.50	1
109.15	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

REGRESSION IMPUTATION

What are the consequences of regression imputation if we have:

MCAR? Unbiased inference

MAR? Unbiased inference

NMAR? Possibly biased inference

Main concern: Does not account for uncertainty around fitted value, i.e. residual for imputed observations will always be 0.

Multiple Imputation

We can model missingness!

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\hat{y}_{2,1}; \hat{y}_{2,2}; \ldots; \hat{y}_{2,m}$	33	2.42	0
15.00	34	4.50	1
$\hat{y}_{4,1}; \hat{y}_{4,2}; \ldots; \hat{y}_{4,m}$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Include in your model all X's that you think affect missingness! (And all X's in your model)

Multiple Imputation

Steps to multiple imputation:

1. Impute m values for each missing element
2. Create m completed data sets
3. Run your statistical model on each imputed data set
4. Calculate the point estimate *across* imputed data sets:

$$\bar{\hat{\beta}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j$$

5. Calculate the standard error for this estimate *across* imputed data sets:

$$S[\bar{\hat{\beta}}] = \sqrt{mean(SE_k^2) + Var[\hat{\beta}_j] \times (1 + \frac{1}{m})}$$

Multiple Imputation in R: Amelia

```
library(Amelia)

set.seed(1234)
a.out <- amelia(x = cc.missing, m = 5)
names(a.out)
```

`a.out` is a *list* of 5 imputed datasets, each of which can be accessed using `a.out$imputations[[i]]`.

Multiple Imputation in R: Amelia

Now, we just want to estimate a basic regression model with our imputed data, but we have 5 datasets!

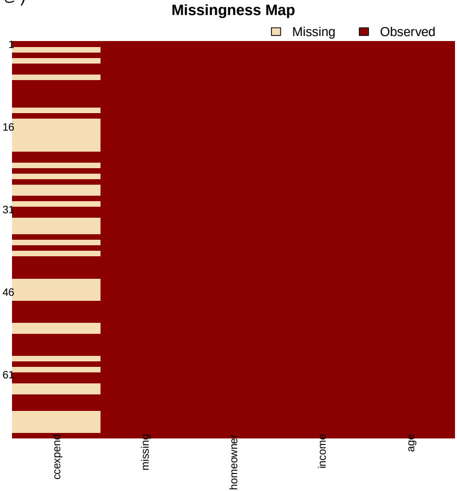
```
mult.imputation <- zelig(ccexpend ~ income + homeowner + age,
    data=a.out$imputations, model="normal")
summary(mult.imputation)

            Estimate Std.Error z value Pr(>|z|)
(Intercept)   109.34     99.88    1.09   0.2737
income         42.90     22.23    1.93   0.0536
homeowner     136.49     50.98    2.68   0.0074
age            -3.23      4.66   -0.69   0.4883
```

`Zelig` will automatically combine the results of the different models, but if a model you are using isn't programmed in `Zelig`, it isn't hard to combine your estimates.

Multiple Imputation: Diagnostics

The missingness map gives an overall sense of the shape and extent of the missingness.

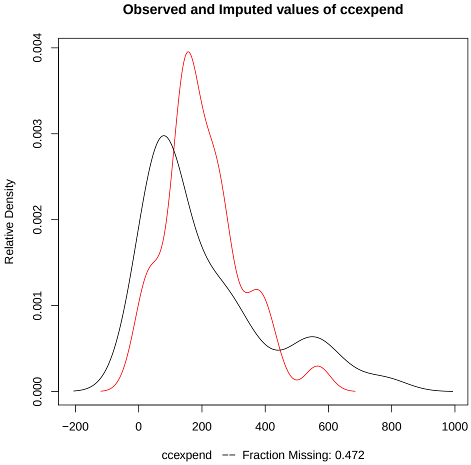
```
missmap(a.out)
```



Multiple Imputation: Diagnostics

Plotting the Amelia object contrasts empirical and imputed densities.

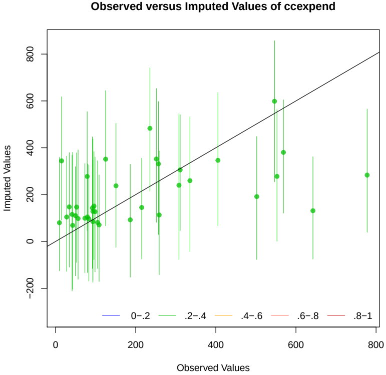
```
plot(a.out)
```



Multiple Imputation: Diagnostics

Overimputation for a specific variable tests the imputation model by imagining that each observation is missing and generating some imputations to check performance.

```
overimpute(a.out, var = "ccexpend")
```



Considerations: Transformations

Our model assumes multivariate normal data. This suggests some issues:

1. Ordinal variables: imputation is faster and more informative if ordinal variables are permitted to be continuous, but if undesirable use `ords` argument to constrain.
2. Nominal (unordered) variables: amelia automatically converts into factors and imputes accordingly if a variable is passed to the `noms` argument.
3. Various transformations (logarithmic, root) are pre-programmed to make skewed distributions more normal.

Multiple Imputation in Stata

```
* Declare the data to be multiple imputation data
mi set mlong

* mi register imputed specifies the data to be imputed
* mi register regular specifies the variables to be used
     as predictors in the imputation model
mi register imputed ccexpend
mi register regular age income homeowner

* Now we impute using mi impute regress
mi impute regress ccexpend age income homeowner, add(5)
    rseed(1234)
```

Multiple Imputation in Stata

Now, we just want to estimate a basic regression model with our imputed data, but we have 5 datasets!
MI will combine them for us:

```
mi estimate: regress ccexpend income homeowner age
```

THINGS TO REMEMBER

1. Set the seed!
2. Include any variable in the analysis model in your imputation model.
3. Don’t impute things that don’t make sense.
4. Check diagnostics (and think carefully about applicability).
5. Remember transformations, polynomials and data structure.

OTHER REASONS FOR MISSINGESS

- Attrition due to social/natural processes
- Skip pattern in survey
- Don’t know responses

HANDLING DON'T KNOWS

What can we do with don't know responses?

- Discard them from the analysis (pretend it is missing data)
- Include it as a category if categorical variable
- Imputation
  - Include indicator for don't know

No matter what, check to see if don't knows are "missing at random", i.e. how they correlate with other data! If it is MCAR or MAR can we just discard?

Handling Skip Patterns

What can we do with missing data from skip patterns?

- Discard them from the analysis
- Model data separately
- Impute data altogether
  - Include indicator for skip pattern
- Impute data separately

Outline

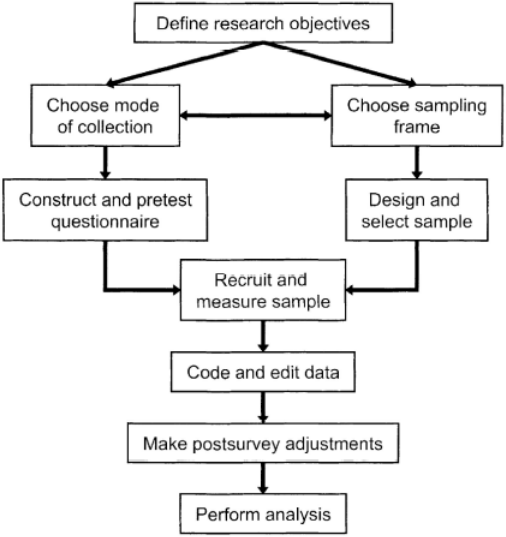
Nonresponse

Item Nonresponse

Survey Interviewing

Quality Control

PROCESS OF A SURVEY


Define research objectives
Choose mode of collection
Choose sampling frame
Construct and pretest questionnaire
Design and select sample
Recruit and measure sample
Code and edit data
Make postsurvey adjustments
Perform analysis

“Recruit and Measure Sample”

1. Field Team Selection
2. Interviewer Training
3. Fielding of Survey
4. Quality Controls

Our data is only as good as our interviewers!

ROLE OF THE INTERVIEWER

1. Build sample frames (listings)
2. Sample respondents from frames
3. Elicit cooperation of respondents
4. Help respondents perform/complete the survey
5. Manage the question-and-answer process
6. Record the answers
7. Edit answers and transmit data

INTERVIEWER EFFECTS

Interviewer Bias: systematic deviations from the true values of responses that result from characteristics and behaviors of the interviewer.

1. Social desirability bias
2. Interviewer characteristics
3. Interviewer experience

Interviewer Variance: the component of overall variability in survey statistics associated with the interviewer.

- Respondent assignment is not random across interviewers
- Interviewers influence the answers

Strategies to Reduce Interviewer Effects

1. Encourage rapport with professionalism
2. Read questions exactly as worded
3. Explain survey procedures and question-and-answer processes to respondent (train the respondent)
4. Nondirective probing
5. Record answers without interpreting, paraphrasing, or inferring

Probing

Ways that respondents may answer inaccurately:

1. Not answer the question asked or answer another question
2. Answer is to vague
3. Give only one answer when question allows for multiple

What should we do?

Response Recording

Should interviewers read out all responses?

Advantages?

- Less interviewer discretion and so likely less error

Disadvantages?

- Respondents grow tired
- “Field coding” can solicit natural responses

FIELD TEAM STRUCTURE

- Interviewers
- Supervisors
- Monitors
- Field Manager

- Backcheckers
- Auditors

Recruiting Interviewers

- Hire a survey firm
- See if research organizations in the area maintain lists of experienced interviewers
- See if other academics in the area have lists of interviewers they have worked with
- Recruit via local methods (from Universities; newspapers; etc)

Interviewer Training

Of critical importance!

Allocate more time than expected.

Need to train on:

- Questions and response coding
- Responses to range of scenarios
- Extra training for sensitive or complex questions; survey experiments
- Mode of data collection (phones/tablets, paper, etc)
- Consent protocols
- Sampling protocols

Outline

Nonresponse

Item Nonresponse

Survey Interviewing

Quality Control

THREATS TO DATA QUALITY

Separate from concerns of measurement error/sample error/coverage error, the process of collecting data can often create threats to data quality. This is particularly true when interviewers are involved.

1. Fabrication of all or part of an interview
2. Deliberate misreporting of process data (ex. refusals)
3. Deliberate miscoding a response to avoid further follow-up questions
4. Deliberate interviewing of a nonsampled person to reduce effor
5. Miscoding of responses due to poor understanding of survey protocols

DATA QUALITY

How can we ensure that the data that we get is of high quality?

We establish a set of stringent data quality protocols:

- Observational methods
- Recontact methods
- Data analysis methods

Which of these methods you choose to ensure data quality depends on the survey itself and mode of data collection, but often survey involve a combination of all three!

Observational Methods

Observational Methods to Ensure Data Quality: a third party hears and/or sees the interview take place.

- Accompaniments - supervisor/monitor accompanies the interviewer for all or part of their survey
- Random Field Checks - field manager/project associate/researcher randomly visits the field team to observe interviews and overall sampling and recruitment
- Audio Audits - part of each interview is recorded and a separate auditor listens and validates interview

Recontact Methods

Recontact Methods to Ensure Data Quality: another staff member speaks with a respondent after the interview is reported to verify the interview was completed according to the specified protocol.

- Supervisor Follow-up - supervisor or other staff member immediately debriefs with respondent after interview

- Backchecks - a different interviewer resurveys a respondent so that data can be cross-checked with original data
  - Backcheck questionnaire is short version of full questionnaire
  - Select questions to validate easy measures, measures with expected variance, and difficult measures
  - Regularly change questions in the backcheck questionnaire to ensure interviewers cannot predict where they must solicit high quality data

Data Analytic Methods

Data Analytic Methods to Ensure Data Quality: examination of completed data records results from the interviews.

- Pre-submission Data Scrutiny - monitors/field manager check interview forms each day as it is submitted, looking for:
  - screening
  - respondent sample
  - completeness of data
  - visual patterns
- Post-submission Data Scrutiny - researchers analyze submitted data, looking at:
  - high frequency checks
  - inconsistent answers
  - completeness of data
  - overuse of particular response items (ex. Don't know)
  - inaccurate coding
- Geolocating Surveys - record GPS coordinates during interviews and validate with sample frame maps

Data Reconciliation

If concerns with the data emerge:

- Have a conversation with or remove interviewers for whom data quality issues arose
- Retrain some or all interviewers
- Establish tighter monitoring of particular interviewers
- Rethink or redesign survey protocols

Best Practices

- 10% Accompaniments
- 10% Audits
- At least 10% of all surveys backchecked

Crowd-source Knowledge

For those who have run or participated in surveys, what have been challenges that you have experienced in implementation and how have you overcome them?

QUESTIONS

Questions?