**FUNDAMENTALS OF SURVEY DATA SET CONSTRUCTION**

**SURVEY RESEARCH LABORATORY**

**GEOFFREY PARKER**

---

**GENERAL INFORMATION**

- Please raise your hand to acknowledge that you can hear me
- Please hold questions until the end of the presentation
- Slides from this webinar are available at [http://www.srl.uic.edu/SEMINARS/seminars.htm](http://www.srl.uic.edu/SEMINARS/seminars.htm)

---

**CONTENTS**

- Sources of Survey Error
- Data Quality
- Coding
- Data Management
- Data File Construction
- Missing Data
- Data Editing
- Disclosure Avoidance

---

**SOURCES OF SURVEY ERROR**

1. Sampling error
2. Coverage error
3. Measurement error
4. Nonresponse error
5. Processing error

---

**DATA COLLECTION STANDARDS**

- Data quality
- Coding
- Documentation
- Data protection
- Confidentiality

---

**DATA QUALITY**

- Data accounting system
- Keep a record of survey responses
- Helps avoid ‘losing’ paper questionnaires
- Also important for electronic data
- Check throughout data processing
DATA QUALITY

- Coding systems & conventions
  - Standardize how unusual situations are coded
  - Document decisions for future users

- Electronic data entry system
  - Programmable range checks
  - Rejects illegal codes
  - Can include error and consistency checks
  - Double data entry
    - Best way to catch and fix keystroke errors
    - Can identify systematic errors

DATA ENTRY SOFTWARE

- Should have data integrity checks
- Should recognize automatic skips & fills
- Available software:
  - SPSS Data Entry
  - EpiData Entry
  - Excel
  - Access
  - CASES

TYPES OF CODING

- Pre-coding
- Post-coding
- Real-time coding
- Back-coding

PRE-CODING

- Primarily used for closed-ended variables
- Assign codes in a systematic way throughout the survey/data set

POST-CODING

- Common for open-ended variables
- Creates new codes for responses
- Should only create codes that answer the question
- Set a minimum threshold for creating a new code
- Computer programs can be used for this if you have a large amount of data
  - ATLAS.ti
  - QDA Miner
REAL-TIME CODING

• Direct computer entry during interviews
• Benefits include being able to probe responses

BACK-CODING

• Usually for ‘other-specify’ responses
• Recode answers into an existing pre-codes
• Can also create new codes

CODE CONSTRUCTION RULES

• Codes should be:
  • Clearly defined
  • Mutually exclusive
  • Exhaustive

BAD CODES – NOT MUTUALLY EXCLUSIVE

“What is your annual Household income, before taxes?”

1. Under $10,000
2. $10,000-$20,000
3. $20,000-$30,000
4. $30,000-$40,000
5. $40,000 or more

BAD CODES – NOT EXHAUSTIVE

“How often do you drink alcoholic beverages?”

1. Less than once a year
2. 2 to 11 times a year
3. 1 to 3 times a month
4. Once or twice a week
5. 3 to 6 times a week

SOME TYPES OF VARIABLE CODES

1. Listing codes
2. Bracket codes
3. Scale codes
4. Binary codes
5. Series codes
Listing codes

“What class are you in?”
1. Freshman
2. Sophomore
3. Junior
4. Senior
5. Graduate
6. Other

Bracket codes

“What is your annual Household income, before taxes?”
1. Under $10,000
2. $10,000–$19,999
3. $20,000–$29,999
4. $30,000–$39,999
5. $40,000 or more

Scale codes

“How would you rate your current knowledge of SPSS?”
1. Excellent
2. Good
3. Fair
4. Poor
5. Very poor

Binary codes

“Are the following amenities included in your rent, or are they not included in your rent?”

<table>
<thead>
<tr>
<th>Included</th>
<th>Not Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Heat</td>
<td>1 2</td>
</tr>
<tr>
<td>b. Water</td>
<td>1 2</td>
</tr>
<tr>
<td>c. Gas</td>
<td>1 2</td>
</tr>
<tr>
<td>d. Electric</td>
<td>1 2</td>
</tr>
</tbody>
</table>

Series codes

“What is the biggest problem facing the United States today?”

<table>
<thead>
<tr>
<th>Domestic Issues</th>
<th>Foreign Affairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Unemployment</td>
<td>11. Terrorism</td>
</tr>
<tr>
<td>2. Inflation</td>
<td>12. Iraq war</td>
</tr>
<tr>
<td>3. Taxes</td>
<td>13. Trade</td>
</tr>
<tr>
<td>5. Moral value</td>
<td></td>
</tr>
</tbody>
</table>

DATA DOCUMENTATION

• Copies of questionnaires and training materials
• Coding guidelines
• Sampling plan
• Make it clear which respondents were asked each question
• Link all variables to specific items on the questionnaire
DATA PROTECTION

• Data should always be backed-up
• Control data access via passwords, directory permissions, etc.
• Data accounting system to distinguish data set versions
• Be aware that storage media will deteriorate

CONFIDENTIALITY

• Control access to your data
• Beware of direct and indirect identifiers

PLANNING FOR DATA REDUCTION

• File structure
• Naming conventions for variables
• Data integrity checks
• Data editing if errors are discovered
• How to handle missing data
• Codebook format
• New variable construction
• Decision documentation
EMPHASIZE DATA SET INTEGRITY

- Use a data-entry program
- Conduct quality-control checks
- Consider 100% double-entry
- Separate coding & data entry tasks
- Have complex coding done by specialists
- Use computer programs for complex coding & recoding if possible

BASICS OF DATA FILE CONSTRUCTION

- Variable names and labels
- Value labels
- Data record formats

METHODS FOR CONSTRUCTION OF VARIABLE NAMES

1. One up numbers
   V1 to Vn
2. Question numbers
   Q1, Q2a, Q2b ... Qn
3. Mnemonic names
   Num_kids, age_grp
4. Prefix, root, suffix systems
   Pre_qs1, post_qs1

VARIABLE NAMES

Should provide:
- Question number
- Clear definition
- Whether variable is constructed from other variables
- If imputed values included

VARIABLE LABELS

Always take the time to include these
Employ a consistent format
Label missing values as well:
- Refusals (Ref)
- Don't know (DK)
- Not applicable (NA)
- No Coded Response Applicable (NCRA)

VALUE LABELS
DATA RECORD FORMATS

- Cross-sectional surveys
- Longitudinal surveys

COMMON TYPES OF MISSING DATA

- Refused/Not answered
- Don’t know
- Not applicable/System Missing
- Processing error

MISSING DATA RETRIEVAL

- Should be done whenever missing data is:
  - Accidental and can be easily recovered with minimal cost and/or effort
  - A serious problem, or essential to your research, regardless of difficulty
CODING MISSING DATA

• In general, best to use values that exceed the maximum legal value for each variable
• If missing data have been imputed, this needs to be indicated. Approaches to doing this:
  • Include an imputation flag
  • Include imputed & non-imputed values

CODING SLOPPY SELF-ADMINISTERED DATA

• Multiple responses checked or circled
• Check marks between responses
• New responses written in the margins
• Imprecise responses (i.e., “3 or 4 times”)
• Items left blank
• Not following skip patterns
• Not taking survey seriously

EFFECTS OF DATA EDITING ON DATA QUALITY

Bauer & Johnson (2000):

• Compared effects of five consistency-check editing procedures on tobacco use estimates among public school students
• Various editing rules produced estimates ranging from 25.6 to 29.7%

EDITING RULES Examined

<table>
<thead>
<tr>
<th>Approach</th>
<th>Estimate</th>
<th>N missing*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do-nothing</td>
<td>29.5%</td>
<td>81</td>
</tr>
<tr>
<td>Gatekeeper</td>
<td>28.7</td>
<td>33</td>
</tr>
<tr>
<td>Limited</td>
<td>27.4</td>
<td>572</td>
</tr>
<tr>
<td>Global</td>
<td>25.6</td>
<td>1,374</td>
</tr>
<tr>
<td>Preponderance</td>
<td>29.7</td>
<td>58</td>
</tr>
</tbody>
</table>

*Total n = 10,563

OTHER CONCLUSIONS

• Extensive exclusionary rules (i.e., the global approach), may bias results toward the null
• Males and African Americans were more likely to have responses excluded because of inconsistencies
• When comparing across years, make sure same edit procedures are used

DISCLOSURE AVOIDANCE

Types of identifying information that need to be considered:

• Direct identifiers (name, address, SS#)
• Indirect identifiers (combination of variables used to identify individuals)
  • Detailed geography
  • Organizational affiliation
  • Educational institution & year of graduation
  • Exact occupation or job title
  • Detailed income
  • Exact dates of events
  • GIS (Geographic Information System) indicators
HOW TO HANDLE INDIRECT IDENTIFIERS

- Removal – eliminate variable from the dataset entirely
- Bracketing – combining the categories of the variable
- Top-coding – restricting the upper range of the variable
- Collapsing and/or combining variables – merging the concepts embodied in two or more variables by creating a new summary variable
- Disturbing – adding random variation or stochastic error to variable

INSTITUTIONAL REVIEW BOARDS

- Very concerned with disclosure avoidance issues
- Confidentiality
- Privacy
- Should be viewed as an ally in this endeavor
- Will point out potential risks you may have overlooked

THANK YOU!

- Please complete our feedback survey, coming your way within a day or two.
- A recording of this webinar and a bibliography should be available in a week or two at http://www.srl.uic.edu/seminars.htm
- You can e-mail me at geoffp@uic.edu