Data Acquisition and Management Workshop

Scott Beach, Ph.D.

Director, Survey Research Program

University Center for Social and Urban Research

2/16/06

Part I - Overview

- Design of data collection forms
- Quality control for data collection (Interviewer training)
- Data entry and management
  - Questionnaire editing
  - Coding
  - Data entry and verification
  - Data editing
- Data security/participant confidentiality
- Study Documentation

Data Collection Forms

Important issues:

- Clarity of instructions (self-administered versus interviewer-administered)
- Clarity of layout
- Consistent formatting
- Standardized instruments can dictate design
- Use standardized instruments only as intended, previously published – any changes can alter psychometric properties

Data Collection Forms

- Should be constructed so that:
  - Respondents are motivated to complete both questionnaire & individual questions
  - Respondent burden is minimized
  - Questions and answers are clearly understood (consistently & easily)
  - Respondent understands how to respond (instructions and skip patterns)

Experience of Care and Health Outcomes Survey

Receipt of Treatment

In the last 12 months, did you get counseling, treatment or medication for any reason?

___ Yes, Go to next section ___ No, Go to next question

Are you aware of the counseling and treatment services that are available to you for personal or family problems, mental or emotional illness, and drug and alcohol problems?

___ Yes, Continue ___ No, Go to End of Survey

If you needed counseling or treatment for one of these problems, do you know the procedures for getting that kind of help?

___ Yes ___ No, Go to End of Survey

Experience of Treatment

At any time in the last 12 months, did you need to get counseling or treatment right away?

___ Yes ___ No Skip next question

In the last 12 months, when you needed to get counseling or treatment right away, how often did you see someone as soon as you wanted?

__Never __ Sometimes __ Usually __ Always

PERSONAL OR FAMILY COUNSELING

People can get counseling, treatment or medication for many different reasons, such as:

- For feeling depressed, anxious, or "stressed out"
- Personal problems (like when a loved one dies or when there are problems at work)
- Family problems (like marriage problems or when parents and children have trouble getting along)
- Needing help with drug or alcohol use
- For mental or emotional illness

1. In the last 12 months, did you get counseling, treatment or medication for any of these reasons?

1☐	Yes	➡	If Yes, Go to the Section Entitled "Your Counseling and Treatment in the Last 12 Months "
2☐	No	➡	If No, Go to Question 2

2. Are you aware of the counseling and treatment services that are available to you for personal or family problems, mental or emotional illness, and drug and alcohol problems?

1☐	Yes	➡	If Yes, Go to Question 3
2☐	No	➡	If No, Go to the Section

YOUR COUNSELING AND TREATMENT IN THE LAST 12 MONTHS

For these questions, please include all counseling or treatment you got in the last 12 months except what you got during an overnight stay or from a self-help group.

4. At any time in the last 12 months, did you need to get counseling or treatment <u>right away</u>?

1☐	Yes		
2☐	No	➡	If No, Go to Question 6

5. In the last 12 months, when you needed to get counseling or treatment <u>right away</u>, how often did you see someone as soon as you wanted?

1☐	Never		
2☐	Sometimes		
3☐	Usually		
4☐	Always		

6. In the last 12 months, did you make any <u>appointments</u> for counseling or treatment?

1☐	Yes		
2☐	No	➡	If No, Go to Question 8

Data Collection Forms

- Create an uncluttered, spread out look – lots of white space, fairly large font, fair size margins, don't cram the page, make it easy to read

- Avoid the appearance of burden or length – if it's too cramped, small, or long, people won't want to do it

Data Collection Forms

- Include instructions where they are appropriate in the questionnaire, not only in the beginning – make them brief and clear (e.g., circle all that apply)
- Don't split questions across pages – if a question is too long for one page, restate the question and response categories on new page
- Provide clear skip instructions – use arrows whenever possible

Data Collection Forms

- Leave enough space for a response – space to circle or check the answer without touching an adjoining category or number
- You can use response matrices for several questions that have the same response categories, but carefully – don't overdo it so that a response set is created

Interviewer-Administered Forms

- Include everything that will be read to respondent – instructions, transitions, etc - don't leave this to the interviewer
- Include answers for possible respondent questions about the survey (e.g., sponsor, how will data be used, etc.)
- Include instructions for interviewer (e.g., definitions, what to do for specific responses, etc.) – differentiate them by font, size, capital letters, etc.

Interviewer-Administered Forms

- Include answer categories as part of the question

- Differentiate elements that are not to be read e.g., Use different font or capital letters:

  1 Strongly Agree
  2 Agree
  3 Disagree
  4 Strongly Disagree
  8 DON'T KNOW
  9 REFUSED

- More transition statements may be necessary for verbal administration

Quality Control for Data Collection

- Interviewer training

  - standardized interviewing techniques:
    1. reading questions as written
    2. “training” respondents in standardized interview protocols
    3. Probing non-directively
    4. Recording answers exactly as given
  - study-specific training:
    1. study background
    2. Q by Q / forms
    3. data handling procedures

Quality Control for Data Collection – Interviewer Training

- Lecture/didactic, video, role-play
- Interviewer certification
- Re-training, re-certification to avoid drift

Quality Control for Data Collection – Interviewer Training

- UCSUR – standard 3 day training protocol

1. General training
2. Computer-assisted telephone interviewing (CATI) system
3. Study-specific (Q by Q)

PMBC - Handout

Data Entry and Management: Questionnaire Editing

- Questionnaires reviewed for problems when received. ("Scan editing") For example:
  - partial completions – perhaps discard if too much missing data
  - item missing data or item non-response
  - Mismarked answers; multiple or ambiguous responses
  - incorrectly followed skip instructions
- These must be corrected if possible (perhaps re-contacting the respondent) or noted if not.

Data Entry and Management: Coding

- Translating questionnaire responses into numeric codes for statistical analysis

  – Structured questions usually pre-coded

  – Open-ended text questions require coding

  - verbatim responses classified into one of several numeric codes by trained coders

Data Entry and Management: Coding

- “Field Coding” – Interviewers ask open-ended questions, code responses into pre-defined categories “on the fly” during the interview.
- Use with caution! Generally avoid unless fairly simple questions and relatively few, straightforward categories
- Complicates interviewing task – they already have enough to do!

Data Entry and Management: Coding

- Developing a set of codes
  - Codes or categories can be developed a priori or from the first set of questionnaires
  - Higher level of abstraction (smaller number of codes) versus staying closer to the data (larger number of codes)
  - Use standard missing value codes (e.g., 8, 98 =Don’t know; 9, 99 = Refused)

Data Entry and Management: Coding

- In addition to open-ended responses, other variables also often require coding:
  - occupation / industry / place of work
  - academic field of study
  - home purchases
  - medical / psychiatric diagnoses

- Standard classification systems (SOC; NAICS for occupation; ICD-9-CM; DSM-IV for medical & psychiatric diagnoses)

Data Entry and Management: Coding

- Issues with manual coding (human judgment) – Coder variance / disagreements – independent verification & adjudication
- Also, poor coding schemes
- Automated coding systems – store dictionary of codes in database – checks for matches when text data is entered (unmatched responses need to be manually coded)
- (Stuart Shulman – March 2 workshop!)

Data Entry and Management: Data Entry and Verification

- Data entry can occur from either the original questionnaire or from a coding sheet onto which codes have been transcribed

- Specialized data entry software packages to simplify entry (e.g., SPSS, Epi Info, Keyentry) or create your own (in, e.g., ACCESS, Excel, etc.)

Data Entry and Management: Data Entry and Verification

- Text-based data entry – enter series of numbers in flat text file (increased speed, reduced accuracy?)
- Form-based data entry – mimics paper forms (reduced speed, increased accuracy?)

Data Entry and Management: Data Entry and Verification

- Data entry software usually can incorporate data checking and data range rules to minimize incorrect entry (e.g., allowable values from 1 to 4 and 8 and 9; if sex = male, skip question 5)

Data Entry and Management: Data Entry and Verification

- Verification is a second independent entry (double entry) to check for accuracy
- Specialized data entry software incorporates – mismatches are immediately corrected.
- Can reenter all or a randomly selected fraction of the questionnaires.

Data Entry and Management: Data Entry and Verification

- Another data entry approach: scanning, e.g., Teleform (bubble responses) – "Mark Character Recognition" (MCR)
- "Intelligent Character Recognition" (ICR) – turns hand-printed characters into machine-readable form
- "Optical Character Recognition" (OCR) – converts machine-generated characters (e.g., bar codes) into computer-readable form.
- "Voice Recognition Entry" (VRE) – converts voice patterns into machine-readable chracters.
- Efficient, fast, but can be *problems with character recognition, accuracy*

Data Entry and Management: Data Editing

- Checking the data once it is in the dataset (after data entry)
  - Examining the distributions of each variable
    - Range checks; outliers in continuous data
    - Invalid codes
    - Balance edits (percentages should add to 100)
    - Incorrect skip patterns (counts don’t match)
    - Missing data checked
    - Consistency edits (age less than 12 should not be married)

Data Entry and Management: Data Editing

- Must decide on how to proceed when error detected
- Time consuming – can be expensive
- Must be careful not to “fix” the wrong item
- Focus on major issues & variables
- Balance accuracy and costs

Electronic forms of data collection (CATI, CAPI, Web surveys etc.)

- Data entry occurs when question is answered and either interviewer or respondent enters it
- Coding and checking process for most variables must be dealt with in programming the software, so you have to make decisions early (e.g., valid values, out-of-range checks, skip patterns)
- Therefore, less error prone (except for original data entry)
- Still have to do post entry data checking and coding and entry of open-ended questions

- More in Part 2!!

Data Security/Participant Confidentiality

- Protecting data from unauthorized access
- **Paper forms** should use ID code – keep names, addresses, etc. separate
- ID linkage codes stored in locked file cabinets

Data Security/Participant Confidentiality

- **Electronic data** accessible by password-protected login only
- Encryption of data files, e-mail transfers
- Secure study web sites

Data Security/Participant Confidentiality

- **HIPPA** - Health Insurance Portability and Accountability Act
- Restricts use and disclosure (by health providers, health plans, etc.) of identifiable health information (e.g., from medical records) for participant recruitment or retrospective secondary analyses *without prior written authorization by the patient*.
- Research registries or repositories – obtain patient consent during tx
- “Honest Brokers” – organization not affiliated with researchers who de-identifies health info., keeps linkage codes, etc.

Documentation

- Manual of Operations – documents study protocol and data collection procedures

  - Data collection forms and instructions (Q by Q, instructions for handling problems, etc.)

  - Data collection schedule

  - Data transfer protocols

- Administrative memos – sequentially numbered operations memos that document administrative decisions and ongoing instructions for conduct of the study

Sample UCSUR Projects

- Data Management Core for *Pittsburgh Mind Body Center* (handouts)

  - UCSUR provides data management and measurement expertise for the four PMBC R01 projects

- *Frequent Hemodialysis Network (FHN)* study quality of life interviews

  - two randomized trials comparing conventional hemodialysis (3 tx/week) with (a) overnight home dialysis, and (b) in-center daily (brief)

  - UCSUR helped translate self-administered forms to telephone scripts; hire, train, supervise interviewers; conduct 30 min QOL interviews; set up data transfer protocol (receipt of patient contact info & send QOL data to Cleveland Clinic)

Part 2 – Overview
New and Emerging Data Collection Methods

- Computer-assisted data collection
- Recent Innovations in computerized data collection
- Web-based data collection methods
- Real-time data capture methods

Computerized Survey Data Collection

- Use of computer technology for survey data collection has become commonplace
- General names:

1. *CASIC* – Computer Assisted Survey Information Collection
2. *CADAC* – Computer Assisted Data Collection
3. *CAI* – Computer Assisted Interviewing

Computerized Survey Data Collection

- Varies by *Mode of data collection*:

- *Face-to-Face*:

  CAPI – Computer Assisted Personal Interviewing

- Interviewer conducts survey with PC, usually in home. Questions appear in proper order (skips automatic). Enters respondent answers directly into computer. Data taken back to central location for extraction and processing.

Computerized Survey Data Collection

- Sample UCSUR CAPI project:

  - Interviews in the homes of low-income, single-mother, current and former welfare recipients – focus on economic circumstances, psychological function, parenting practices, etc.

  - UCSUR programmed CAPI instrument, trained interviewer, and provides data management

- Certain portions of interview are CASI (computer-assisted self-interviewing) – interviewer gives laptop to respondent who fills it out herself

Computerized Survey Data Collection

- *Telephone*:

  CATI – Computer Assisted Telephone Interviewing

- Survey programmed with special software.

- Interviewer conducts survey with PC, with questions appearing in proper order (skips automatic). Enters respondent answers directly into computer.

- Centralized telephone facility (e.g., UCSUR)

Computerized Survey Data Collection

- UCSUR does many CATI studies

- Random digit dialing (RDD) – general population – surveys and participant recruitment
- Also list-based phone surveys (program participants, students)
- National, statewide, and local surveys
- Recruitment targeting demographic sub-groups who meet basic eligibility criteria (more representative, efficient than mass mailings, advertisements, etc.)

Collection Computerized Survey Data

- *Self-Administered*:

  CASI – Computer Assisted Self-Interviewing

- Respondent uses PC to complete survey, with questions appearing in proper order (skips automatic). Enters own answers directly into computer.

- Interviewer / researcher often present

Collection Computerized Survey Data - Recent Innovations

- *A-CASI* – Audio Computer Assisted Self-Interviewing
- Respondent listens to questions through headphones as they appear on screen. (Increased privacy, reduced literacy demands.)
- Mixed, CAPI + A-CASI applications (NSAM, NSFG, NSDUH, NHANES)
- Research: Acceptable; increases reports of illegal, sensitive behaviors above CASI or paper SAQ (most work with teens, younger adults)
- Feasibility / acceptability questions with older respondents

Collection Computerized Survey Data - Recent Innovations

- *IVR* – Interactive Voice Response
- Telephone methodology – computer voice administers questions, respondent keys answers using numbers on phone
- Call-in versus interviewer handoff approaches (Mixed CATI + IVR approach)
- Research: Mixed CAT-IVR more reports of sensitive behaviors than CATI, but issues:
  - break-offs / hang-ups after transfer
  - lack of touchtone phones (elderly?), cordless phones w/keys on handset – awkward responding

Collection Computerized Survey Data - Recent Innovations

- A-CASI & IVR example
- Elder mistreatment survey proposal
- Compare CATI vs. CATI+IVR vs. CAPI vs. CAPI+A-CASI to elicit elder reports of potentially abusive behavior

Collection Computerized Survey Data - Recent Innovations

- Coming Soon….

- Digital sound for administering and responding to surveys
- Visual enhancements (e.g., graphics, photos, interactive media) to web and CASI surveys
- Multimedia communication and surveys (videos as stimuli)
- Touch interfaces for response

- Reference: Couper MP (2005). Technology trends in survey data collection. *Social Science Computer Review, 23,* 486-501.

Web/Internet Surveys

- Recent development
- Not much research yet
- Initial optimism that web/internet surveys are faster, better, cheaper, easier to conduct has not necessarily proved true

Web/Internet Surveys

- Many types
- Non-Probability sample methods – polls as entertainment; unrestricted self-selected surveys; volunteer opt-in panels (e.g., Harris Interactive)
- Probability sample methods – List-based samples; mixed-mode designs with a web option; pre-recruited panels of internet users (e.g., recruited via RDD phone survey); pre-recruited panels of full population (provide internet access, Web TV – Knowledge Networks))

Web/Internet Surveys

- Typical web *survey* procedures:
- Respondent sent e-mail invitation with embedded link to survey.
- Unique username/password also provided.
- Secure transmission (e.g., SSL) procedures to ensure confidentiality
- Can send follow-up reminders/additional e-mails (similar to traditional mail surveys).
- Can also send written letter with URL provided (low response rates).

Web/Internet Surveys

- Shares advantages of CATI, etc. – automated skips, range checks, etc.
- Little known about how visual aspects of design (# questions per screen, response formats, use of graphics, matrix question formats, progress indicators, and error/warning messages) affect response
- User-end technical issues

Web/Internet Surveys

- UCSUR examples:

- Library workforce
- Undergraduate student experience with writing education
- Political culture survey using national web panel
- Frequent Hemodyalysis Network (clinical – web-based form)

Web/Internet Surveys

- Still in infancy as a technique.
- Severe coverage problems with general populations.
- Non-response issues and concerns.
- Little known about measurement error.
- Currently, best (i.e., faster, cheaper, easier) with well-defined, list-based populations for which e-mail addresses are available.

Web Resource

- Web-based survey methods:

  http://www.websm.org/

(Source for web survey methods papers)

Real-Time Data Capture

- Retrospective self-reports and memory distortion (reconstruction versus actual retrieval; effects of current state)

- Use of instantaneous reports of immediate (or very recent) experience **("ecological momentary assessment" - EMA)** in the natural environment

- Averaging of randomly selected momentary reports of pain over a one week period (vs. asking to recall how much pain experienced in past week)

Real-Time Data Capture

- Diary methods (paper & pencil)
- Computerized diary methods (palm pilots)
- Respondents signaled/prompted at various times to provide reports (e.g., pain level; drinking; smoking; mood; social interactions)

Real-Time Data Capture – Sample Applications

- Coping with daily stressful events – random reports of stress levels; stressful events reported via electronic diary shortly after occurrence
- Assessing physiological processes in the natural environment – sampling at various times of day ("diurnal cycle") to get more accurate measures of salivary cortisol (stress marker) level variation
- Variability of asthma symptoms – EMA assessment of activities, location, social contacts, mood and stress, medication, etc. along with asthma symptoms, and peak expiratory flow in real time
- Momentary assessment of pain in chronic pain patients – momentary reports of immediate pain taken several times a day, along with environmental correlates

Real-Time Data Capture

- Issues to consider:

1. *Sampling plan* – random, fixed intervals? Only when events occur? Number of reports per time period? Current vs. recent experience? # questions per report – keep short to reduce burden
2. *Procedures* – signaling (alarm); response interface (paper vs. PDA) – usability, burden
3. *Respondent training* – informal vs. formal; allow practice period and feedback?

Real-Time Data Capture

- Potential barriers/problems:

- Compliance (especially with paper diaries)
- Technology issues
- Respondent vision, hearing, dexterity
- Respondent burden
- Reactivity/bias due to knowing you are being measured
- Amount and complexity of resulting data – advanced data management & statistical techniques required
- Cost

Real-Time Data Capture

- Example (UCSUR)

- Pilot study of Arthritis sufferers and their spouses using palm pilots to collect data on patient pain and mood (spouse perceptions of patients) 3 or 4 times/day (Lynn Martire)
- Fixed sampling interval
- Importance of training
- Simple stylus responding
- High resolution, loud alarms (vision & hearing)
- Keeping PDA battery charged - compliance

Web Resource

- Real-time data capture

http://dccps.cancer.gov/hprb/real-time/

(Summary of National Cancer Institute conference on real-time data capture held in Sept 2003. Includes presentations)