Introduction to Survey Research

Clement Stone

Professor, Research Methodology

Email: cas@pitt.edu
Presentation Outline

• What is survey research and when is it used?

• Stages of survey research

1. Specifying research questions, target population
2. Developing the instrument
3. Choosing an administration method
4. Selecting a sample
5. Analyzing Surveys (Reliability and Validity Evidence)
What is Survey Research and When is it Used?

Survey research is a research method involving the use of questionnaires or surveys to gather information from individuals.

- Surveys involve systematic collection of information using standardized procedures

- Surveys ask people questions designed to measure or identify the status or level of a characteristic

- Respondents self-report on a variety of characteristics:
  - Characteristics of Respondents
  - Behaviors
  - Opinions
  - Feelings
  - Knowledge or Perceived Knowledge
  - Theoretical Constructs - abstract concepts, unobserved variables
What is Survey Research and When is it Used?

Why survey research?

- In contrast to direct observation, surveys are efficient tools for collecting information.
- Surveys can be used when it is not possible to observe behaviors (e.g., self-esteem, drug use).
Sources of Error in Survey Research

Self-report data from surveys are a fallible source of information and subject to many sources of error.

- **Errors of Observation** – due to measurement process
  - Administration Method (e.g., interviewer effect)
  - Respondent (e.g., social-desirable response, careless)
  - Instrument (e.g., wording vague, question order, response format)
Sources of Error (cont.)

• *Errors of Non-Observation* – due to sampling process
  
  - Coverage error - Sample surveyed may not adequately represent all individuals
  
  - Sampling error - Sample may not be large enough to have confidence in population estimates
  
  - Non-response error – Individuals sampled may not respond to all items or elect not to participate.

*Pattern of missing data more important* than amount.

From Social Exchange Theory: to maximize response rate → maximize reward for responding, minimize cost to responding, and maximize relationship between respondent and researcher.
Developing the Instrument

What to measure? Purpose of the research will dictate what to measure and guide specific nature and content.

• Develop a framework or blueprint for what is to be measured - outline the specific content to be measured
  - Specify behaviors, attitudes, feelings, and opinions
  - Two methods of scale development guide measurement of theoretical constructs
    Deductive – based on theory
    Inductive – ask group to define construct

Operationalizing construct – Specify elements that reflect nature and range of each construct
Developing the Instrument

• Develop items based on the framework
  - Given the specified components for a construct or a set of components, one or more items written to measure each component
  - Items written to measure a construct are indicators of the construct

If you want to know what is being measured look at the items.
Example blueprint - Measuring depression

Based on clinical experience – measurement of depression operationalized by specifying 4 different classes of symptoms of depressed patients:

I. Interpersonal relationships (loneliness, likability, communication) (coded I),
II. Sleep and eating habits (coded S),
III. Energy and concentration levels (coded E), and
IV. Feelings about the past and future (coded F)

Items written to reflect the different symptoms
20 Items written to reflect the different symptoms:

1. I was bothered by things that usually don’t bother me. F
2. I did not feel like eating. S
3. I felt that I could not shake off the blues. F
4. I felt that I was just as good as other people. I
5. I had trouble keeping my mind on what I was doing. E
6. I felt depressed. F
7. I felt that everything I did was an effort. E
8. I felt hopeful about my future. F
9. I thought my life was a failure. F
10. I felt fearful. F
11. My sleep was restless. S
12. I was happy. F
13. I talked less than usual. I
15. People were unfriendly. I
16. I enjoyed life. I
17. I had crying spells. F
18. I felt sad. F
19. I felt that people disliked me. I
20. I could not get going. E
Writing Items

Items from other sources *are not copyrighted*

Open-ended versus Closed-ended items

• With open-ended items, respondents use their own words to provide a response – why do you smoke?

  - I smoke to keep my weight down
  - I smoke when socializing,
  - I smoke when drinking, etc.

  Response scale: SA  A  D  SD

• With closed-ended items, respondents map response to a set of options–To what extent do you agree with…
Measurement may be different for open and closed-ended items. Example: People look for different things in a job. Which would you most prefer?

<table>
<thead>
<tr>
<th>Closed-ended question:</th>
<th>Open-ended question:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pays well</td>
<td>1. Pays well</td>
</tr>
<tr>
<td>2. Gives a feeling of accomplishment</td>
<td>2. Feeling of accomplishment</td>
</tr>
<tr>
<td>3. Not much supervision</td>
<td>3. Control of work</td>
</tr>
<tr>
<td>4. Nice people to work with</td>
<td>4. Pleasant environment</td>
</tr>
<tr>
<td>5. Steady and secure</td>
<td>5. Security</td>
</tr>
<tr>
<td>6. Don’t know, no opinion</td>
<td>6. Opportunity for promotion</td>
</tr>
<tr>
<td></td>
<td>7. Short hours</td>
</tr>
<tr>
<td></td>
<td>8. Good working conditions</td>
</tr>
<tr>
<td></td>
<td>9. Good benefits</td>
</tr>
<tr>
<td></td>
<td>10. Satisfaction</td>
</tr>
<tr>
<td></td>
<td>11. Don’t know, no opinion</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>13%</td>
<td>17%</td>
</tr>
<tr>
<td>31%</td>
<td>15%</td>
</tr>
<tr>
<td>12%</td>
<td>5%</td>
</tr>
<tr>
<td>20%</td>
<td>15%</td>
</tr>
<tr>
<td>20%</td>
<td>8%</td>
</tr>
<tr>
<td>4%</td>
<td>1%</td>
</tr>
<tr>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>7%</td>
<td>7%</td>
</tr>
</tbody>
</table>
Common response scales for closed-ended questions

• Rating Scales – on a scale from 1 to 10 with 1 being none to 10 be the most, how would you rate…

• Graphical Scales (intensity) - Mark on the scale below the degree of pain you experienced

|----|----|----|----|----|----| OR |__________________________|
no pain ➔ worst possible pain

• Likert-type Scales:
  None, Very Mild, Mild, Moderate, Severe
  Strongly Disagree, Disagree, Agree, Strongly Agree

- Respondents identify themselves on the continuum or scale for the measured characteristic
Writing Items (cont.)

Other Common response scales for closed-ended questions

• Frequency Scales

  About once a month, 2 or 3 times a month, Few times a week, Almost every day, More than once a day
  Never, Rarely, Sometimes, Usually, Always

- Numerical ranges should reflect anticipated range

You want to discriminate respondents on the characteristic being measured ➔ Want responses across range of response options.
Summated Scales - common practice for a series of ordinally scaled items to be summed to provide an estimate of the characteristic or construct being measured.

• Example – Depression scale with multiple indicators:
  During last week, how often …

<table>
<thead>
<tr>
<th></th>
<th>Rarely</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>were you bothered by things.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>were you happy.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>did you not feel like eating.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>did you feel sad.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>did you feel people disliked you</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>could you not get going.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Summated Scales (cont.)

• Sum up responses to get an estimate of the level of depression for an individual

• Assumes measuring a single dimension and items measuring in same direction (reverse code – how often were you happy?)

• Assumes same meaning attached to response categories across the set of items

• Assumes each item contributing equally to score

• Differences between response options and summed scores assumed equal
Writing Items (cont.)

**FAQ** - *Including Mid-points in Scales?*

Negativity Scale: Equally Positive and Negative
Satisfaction Scale: Neither Dissatisfied nor Satisfied
Agreement Scale: Neutral

- Don’t Know not a mid-point
- Midpoint may be chosen for reasons other than neutrality
- May want to measure number of mid-point responses

Non-response less for items with a midpoint, but respondents gravitate to midpoint
Mid-points in Scales (cont.)

- Study examined the presence and absence of a middle position. Item - Should divorce be easier or more difficult to obtain? *easier*  *more difficult* (stay as is option for one form)

<table>
<thead>
<tr>
<th></th>
<th>Form without <em>stay as is</em></th>
<th>Form with <em>stay as is</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Easier</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>More difficult</td>
<td>45</td>
<td>33</td>
</tr>
<tr>
<td>Stay as is</td>
<td>22 (volunteered)</td>
<td>40</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

-Evidence that tendency to select midpoint related to personality characteristics – correlated with measures of anxiety and social conformity.
Writing Items (cont.)

Other FAQ about response scales

• *Number of response options?* 5-7 options optimal

• “Never” and “Always” categories? Avoid unless their measurement important; EX – How often do you use birth control? (alternatives – “Hardly Ever” and “Almost Always”)

• *Filters:* don’t know, no opinion, not applicable (NA), no basis for judgment, prefer not to answer, can’t recall?

  - Place filters outside the scale continuum:

    | Never | Sometimes | Regularly | Often | NA |
    |-------|-----------|-----------|-------|----|

  - Exercise caution in their use – may be over-selected
Other FAQ about response scales (cont.)

• **Strategies for Response Sets?**

  - **Acquiescence** - responding positively or agreeing

    Occurs disproportionately for some groups (e.g., less educated, impulsive, emotional individuals).

    Include mix of negatively worded items with positively worded items (e.g., I am happy, I am sad).

  - **Deliberate misreporting** – providing answers that are more socially acceptable or exaggerating a problem

Motivate respondent

If sensitive content – increase perceived anonymity

Use a scale to measure propensity for faking
**Other sources of error in items**

*Misinterpretation*

Be clear, concise, and use appropriate wording
Poor - How often do you deliberately refrain from lighting up a cigarette to keep your smoking rate down?

Measure one issue at a time per question
Poor - Please rate your satisfaction with the amount and kind of care you received

Avoid modifiers in questions
Poor - I frequently worry about my health
   never sometimes often always

Avoid vague references
Poor: Have you ever smoked?
   How many drinks did you have last week?
Other sources of error in items (cont.)

Flawed reporting (e.g., respondents overestimate occurrence of low frequency events, underestimate high frequency events)

Question order effects – order in which questions measuring similar content appear may affect results.

Example: One question: “In general, how would you describe your marriage? Would you say that you are very happy, pretty happy, or not too happy?” Second question: “In general, how would you say things are these days? Would you say that you are very happy, pretty happy, or not too happy?”

Study found .32 correlation between responses when general question posed first and .67 when marital question posed first.

Logistical problems (e.g., instructions not clear)
Other sources of error in items (cont.)

Response option effects

- Negative pole (end of scale) may be viewed as a more extreme position than the positive pole (end of scale)

- Respondents also differ in willingness to select extreme options (e.g., Extremely Important)

- Respondents make assumptions about set of options with frequency scales – assume end points reflect more extreme behavior and middle points reflect “normal” behavior

- Labels have different meanings to different individuals and under different contexts. Scales obtain an objective frequency relative to the respondent’s subjective standard for labels

Example: “Often” different meaning: How often do you suffer headaches? vs. How often do you suffer chest pains?
**Response option effects (cont.)**

Example of differences in interpreting labels – In a post-interview, individuals were asked to define what “very often”, “pretty often”, and “not too often” meant in terms of number of days / month they were excited or bored.

<table>
<thead>
<tr>
<th></th>
<th>Excited</th>
<th>Bored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not too often</td>
<td>6.7</td>
<td>4.2</td>
</tr>
<tr>
<td>Pretty often</td>
<td>13.0</td>
<td>Means</td>
</tr>
<tr>
<td>Very often</td>
<td>17.7</td>
<td>17.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Not too often</th>
<th>8.6</th>
<th>5.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretty often</td>
<td>12.1</td>
<td>SD</td>
<td>10.6</td>
</tr>
<tr>
<td>Very often</td>
<td>15.0</td>
<td></td>
<td>13.1</td>
</tr>
</tbody>
</table>
Choosing an Administration Method

What is the best way to collect the information?

• Two types - self administered written survey or interview
  - Interviews conducted face-to-face or over the telephone
  Adv: Clarify and probe responses, captive audience, build rapport → higher response rates
  Disadv: Cost, potential interviewer effects, lower perceived anonymity, smaller samples
  - Written surveys delivered face-to-face, mail and internet
  Adv: More items, complex items can be administered, perceived anonymity greater, larger N, low cost
  Disadv: Lower response rates, harder to motivate, no opportunity to clarify and probe responses
Choosing an Administration Method (cont.)

• Method should be consistent with target group population

• Increasing response rates for mail and telephone surveys
  - Pre-contact – letter or call in advance of survey
  - Use deadlines and follow-up notices or calls
  - Use self-addressed stamped reply envelopes
  - Identify benefits to participation
  - Consider incentives – size of award not really relevant ($2 bill, lottery ticket, entry for drawing)

Remember, to maximize response rate → maximize reward for responding, minimize cost to responding, and maximize relationship between respondent and researcher.
• **Internet-based Surveys** – our mode of communication is changing, why not our methods for delivering surveys
  - Survey e-mailed, returned via e-mail, fax, or snail mail
  - Survey forms completed directly on web-pages

Adv: low cost, fast, can incorporate graphics, audio, video etc., handles skip rules, eliminates data entry

Disadv: internet vs. P&P versions may not be comparable, potential for coverage error, computer literacy issues, bandwidth limits, acceptance not predictable, less control over response process
• Consider using multiple administration modes

- One mode for initial administration – another mode(s) for reminders/follow-up.
  Start with least expensive approach – mail
  For those that don’t respond, then try next least expensive mode – telephone
  To get remainder of sample, use more expensive mode – face-to-face interview

- One mode for one wave, another mode for additional waves in a panel study.
  Use face-to-face in 1st wave – motivate respondents
  Later waves use written questionnaires (e.g., mail)
• Consider using multiple administration modes (cont.)
  - One mode for some respondents (mail), another mode for other respondents (web survey).
  - One mode for recruitment (telephone – motivate respondents), another mode for data collection (mail)
  - One mode for one subset of items (interview), another mode for other items (written - sensitive items)
Selecting a Sample

Who gets sampled?

• Specify a target population, obtain a list of members from the population, identify a subset from the population (sample), and survey this sample. Objective of sampling is to estimate population parameter.

- Eligibility criteria should be clearly defined

*Inclusion criteria* - characteristics of respondents eligible to participate

*Exclusion criteria* – characteristics that rule out people

*Study population* - subset of people from the *target population* who satisfy inclusion, exclusion criteria
Selecting a Sample (cont.)

How do we identify the sample from the population?

• Methods classified as probability sampling or non-probability sampling methods

*Probability sampling* involves systematic methods for assigning probabilities to members of the population and basing the sample on these probabilities.

- Sample is likely to be representative of population
- Make valid inferences from sample to population

*Non-probability sampling* methods do not consider all members of the population.
Selecting a Sample (cont.)

What if you can’t use probability sampling methods?

• Randomization not always practical or ethical
• Common non-probability sampling methods → lead to observational studies

Convenience sampling - use samples that are readily available or volunteers that meet general requirements

Quota Sampling – convenience sample but respondents selected based on proportions of subgroups needed to represent proportions in population
How can observational studies be used to evaluate the impact of a program or intervention?

• Random assignment controls for observed and unobserved covariates between treatment and control groups.

• To account for group differences between the treatment and potential control group members in observational studies, *propensity scoring methods* have been proposed.

*Propensity scoring in healthcare research drawing interest.*

Idea is that when a member of treatment group is matched (using many relevant characteristics) with a potential control group member, both considered to have same probability of being in treatment condition.
Selecting a Sample

*How many individuals do you sample?*

- No simple rule of thumb
  - Estimating sample size depends on the error in estimating population parameters from samples
  - Collect lots of data?
  - Collect enough data to have a specified level of confidence in our parameter estimates?

- If probability sampling methods used, can calculate how many individuals from the population to sample so that a specified degree of confidence or error is achieved.
- Web tool: [stat.uiowa.edu/~rlenth/Power/index.html](http://stat.uiowa.edu/~rlenth/Power/index.html)
Evaluating Surveys and Summated Scales

- A basic assumption of a test or survey is that the sample of behavior that is collected generalizes beyond the sample and to other occasions. Sources of error related to making these generalizations can be estimated.
- Generalize from the subset of administered items to all possible subsets of items for a given construct.
- Generalize to all possible administration occasions (e.g., different days)
- Generalize to all possible scorers or raters (e.g., open-ended items or personality evaluations)
Reliability Evidence

*Three Types of Reliability Coefficients* - estimate the consistency of scores obtained by the same person

1. Test-retest reliability – measures error due to administration occasion (time)

2. Alternate Forms, Split Half, Internal Consistency reliability – measures error due to sampling of items

3. Scorer reliability – measures error due to raters

- Greater the reliability, the greater confidence in obtained scores
Evaluating Surveys and Summated Scales (cont.)

• Another assumption when using a test or survey is that it measures “what we think we are measuring” or that it can “predict what we think it can predict”. Evidence for these assumptions should be obtained.

- A measure of depression is assumed to measure depression
- A survey of eating and exercise behaviors is assumed to predict risk for cardiovascular disease

Validity Evidence – refers to accuracy of score interpretations and underlying assumptions

• Variety of evidence can be used to establish the validity of the interpretation and assumptions
Validity Evidence Example: A general critical thinking (CT) test is used to measure critical thinking skills of nurses. Scores used to evaluate the extent to which a nursing program enhances the CT skills of nursing students.

Assumption: Test measures CT skills relevant to nursing.
Assumption: Nursing students CT skills enhanced by exposure to nursing curriculum.

- Ask experts whether items on the CT test measure important nursing skills (content-based evidence).
- Examine relationship between scores from the CT test and NCLEX board scores (convergent evidence).
- Examine relationship between scores from the CT test and GPA from clinical courses (predictive evidence).
“What we observe is not nature itself but nature exposed to our method of questioning” Heisenberg, 1958

Questions?