Introduction to Survey Research

Clement Stone

Professor, Research Methodology

Email: cas@pitt.edu
Workshop 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 – a conversation with a purpose.
- Respondents self-report on a variety of characteristics:
  Characteristics of Respondents
  Behaviors
  Opinions
  Feelings
  Knowledge or Perceived Knowledge
  Theoretical Constructs - abstract concepts (e.g., depression)
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>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>
</tr>
<tr>
<td>were you happy.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>did you not feel like eating.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did you feel sad.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>did you feel people disliked you</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>could you not get going.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Evaluating Surveys and Summated Scales

• Use of summated scales evaluated by examining:
  • Reliability evidence - refers to consistency of scores obtained by the same person – for example, over time
  • Validity evidence – refers to accuracy of score interpretations – are we measuring what we want to measure and can we predict what we want to predict?

*Can there be reliability without validity?*

*Can there be validity without reliability?*

*Reliability and validity are properties of scores and score interpretations in the population sampled and the particular research application.*
Reliability Evidence

• 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.
  - 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 – measure different sources of error related to making these generalizations

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 reliability → greater confidence in scores
- Different types measure different sources of error
Measuring Stability Over Time or Occasion

- Relevant when focus is on long-range prediction or measuring something relatively stable (e.g., personality)

Question - would results change over time?

Test-Retest Reliability – Obtained by administering a test to a group, waiting a period of time (e.g., 3 weeks), and re-administering the same test to the same group.

- Reliability estimate is the correlation between the scores

- Several factors may affect the estimate of reliability:
  Stability of the behavior being measured
  Length of time between testing
  Possible practice effects
Measuring Equivalence of Items
- Sample some characteristics or behaviors related to a construct but want to generalize results to broader domain

Question – would results change if different item sample?

Alternate Forms Reliability – Group of persons given alternate forms of a test (e.g., 2 equivalent but different measures of depression)
- Reliability estimate is the correlation between the scores
- How alternate forms are constructed – want approximately equivalent forms
- Recall effects mitigated
Measuring Equivalence of Items (continued)

**Split-Halves Reliability** - Test is split into two halves and a group of persons receives each part. Each half is a different sample of items (e.g., odd-even items from depression survey).

- Reliability estimate is the correlation between scores on the two halves
- Don’t need to construct multiple equivalent forms
- How the test is split an issue → different splits change reliability
- Reliability coefficient adjusted using Spearman Brown formula to reflect true length of administered test
Measuring Equivalence of Items (continued)

**Coefficient Alpha (α) Reliability** (Measure of Internal Consistency) - Group of persons given one form of a test

- Reliability estimate reflects extent to which item responses correlate with each other.

Estimate of correlation between current instrument and all possible instruments with the same number of items drawn from the universe of items measuring the trait of interest.

- Do not need to split the test in half or need alternate forms

- More homogeneous the items → larger coefficient alpha

- **CAUTION** - High alpha does not mean items are measuring one dimension or are unidimensional
Estimating Scorer Reliability

- Relevant when items are scored subjectively or based on judgments (e.g., open-ended items)

Example - 2 raters score each individual (scores are 0-6)

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater A</th>
<th>Rater B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

- Correlation = 1 – same student rankings but no agreement

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater A</th>
<th>Rater B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- % exact agreement = 75%
- % agreement within 1 score point = 100%
Validity Evidence

Validity – refers to accuracy of score interpretations and underlying assumptions

• Initial assumptions when using a test or survey are that it measures “what we think we are measuring” or that it can “predict what we think it can predict”.
  - A measure of depression assumed to measure depression
  - A survey of eating and exercise behaviors is assumed to predict risk for cardiovascular disease
  - A variety of evidence can be used to establish support for test score interpretations and underlying assumptions
  - Assumptions indicate the evidence that is required
Validity re-conceptualized as a “unitary” concept with different types of evidence to be examined (Messick, 1992)

- Content evidence – provides content-based evidence that the items measure the construct(s) of interest
  Do the depression items measure all important elements?

- Internal structure evidence - examines relationships among item responses in relation to reported scores
  Does the depression scale measure one dimension?

- External structure evidence - examines relationships of survey results to other variables external to the survey
  Do scores from the depression scale correlate with other measures of depression?
Context for Discussion of Validity Evidence

Pitt School of Nursing uses a general measure of critical thinking thinking (CCTST) to measure CT skills of nurses and determine the change in CT skills for nurses in the program. The test measures core critical thinking skills in 3 domains (Analysis, Inference, and Evaluation). Items are logic type items such as:

“Terry, don’t worry about it. You’ll graduate someday. You’re a college student. Right? And all college students graduate sooner or later.” Assuming all the support statements are true, the conclusion a) could not be false; b) is probably true, but may be false; c) is probably false, but may be true; d) could not be true.
- School of Nursing uses the general measure (CCTST) to measure CT skills of nurses. The test consists of 34 logic type items that measure core critical thinking skills in three domains (Analysis, Inference, and Evaluation). Scores reported for the 3 domains with a total score.

Assumption – General test measures CT problem solving skills important to nursing practice

Assumption – Students’ CT skills impacted by curriculum

Assumption – There are three somewhat unique dimensions to the test that are also moderately related

- The assumptions indicate the evidence that is required
Content Validity Evidence

• Content related evidence – examination of the *item content* in relation to characteristics or constructs being measured

• Content validity evidence obtained by consulting experts in the measured domains to determine:
  - extent items reflect elements that define the construct
  - importance of items to intended interpretations
  - if enough items to support generalizing to the domain
  - if items may be measuring facets beyond the construct
  - if other methods may be more appropriate - for example, observational vs. self-report measurements
Content Validity Evidence for the CCTST

• Nursing faculty from accredited programs surveyed
  
  - Using definitions for the skills, respondents asked to indicate degree to which they believe each of the skills were essential to practice nursing competently (scale: 1=Not essential to 4=Absolutely Essential)
  
  - Respondents also asked to indicate the most and least important skills to practice nursing competently
  
  - Respondents also instructed to think about the way in which a critical thinking skill was being measured by an item and rate degree to which they thought the skill was measured in a way that reflects CT skills used by nurses
Table 1. Respondents’ Beliefs that Skills are Essential

CT Skills (partial list): Mean (Scale 1-4)

Categorize or organize information\(^\text{A}\) \hspace{1cm} 3.7 \hspace{1cm} \text{M}
Decode the significance of information\(^\text{A}\) \hspace{1cm} 3.7 \hspace{1cm} \text{M}
Detect arguments\(^\text{A}\) \hspace{1cm} 3.1 \hspace{1cm} \text{L}
Analyze arguments\(^\text{A}\) \hspace{1cm} 3.3 \hspace{1cm} \text{L}
Assess claims or judge information\(^\text{E}\) \hspace{1cm} 3.7 \hspace{1cm} \text{M}
Evaluate arguments\(^\text{E}\) \hspace{1cm} 3.4 \hspace{1cm} \text{L}
Present arguments\(^\text{E}\) \hspace{1cm} 3.4 \hspace{1cm} \text{L}
Draw conclusions\(^\text{I}\) \hspace{1cm} 3.7 \hspace{1cm} \text{M}
Examine ideas, explanations, results, biases\(^\text{I}\) \hspace{1cm} 3.7 \hspace{1cm} \text{M}

- 12 of 16 skills viewed as “Absolutely Essential” by 50%
- ~All skills viewed “Essential” or “Absolutely Essential”
### Table 2. Evaluation of Way Items Measure Skills

<table>
<thead>
<tr>
<th>Critical Thinking Item:</th>
<th>Mean (Scale 1-4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Item 1</td>
<td>2.5</td>
</tr>
<tr>
<td>Evaluation Item 2</td>
<td>3.1</td>
</tr>
<tr>
<td>Analysis Item 1</td>
<td>2.9</td>
</tr>
<tr>
<td>Analysis Item 2</td>
<td>2.5</td>
</tr>
<tr>
<td>Inference Item 1</td>
<td>2.6</td>
</tr>
<tr>
<td>Inference item 2</td>
<td>3.2</td>
</tr>
<tr>
<td>Inference Item 3</td>
<td>2.7</td>
</tr>
</tbody>
</table>

- Most items viewed not even “Essential” (range 2.5 – 3.2)
- Conclusion: Table 1 supports inferences based on skills identified by the test developers, but less support for the way in which CT for nurses actually measured (Table 2)
Conceptual framework for construct implies a structure
- 1 dimension (1 score); >1 distinct dimensions (>1 score)

Internal structure examines consistency in relationships among item responses relative to the framework
- Obtain evidence items within dimensions are more related to one another than items across dimensions
- Obtain evidence subscale scores are not strongly related
- Obtain evidence that subscales are moderately related if total score and subscale scores reported

Factor analysis methods often used – item factor loadings examined in relation to hypothesized subscales
Internal Structure Validity Evidence for the CCTST

• Examined the inter-correlations among the subscale scores

<table>
<thead>
<tr>
<th></th>
<th>analysis</th>
<th>eval</th>
<th>infer</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis</td>
<td>Pearson Correlation</td>
<td>1</td>
<td>.282**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>159</td>
<td>159</td>
<td>159</td>
</tr>
</tbody>
</table>

- Correlations low to moderate between Analysis, Inference, and Evaluation subscale scores
External Structure Validity Evidence

• Validity evidence also derived by examining relationship of survey results (scores) with other variables or criteria
  - Measurement of an attribute implies predictable relationships to results from other measures or criteria

Convergent Evidence
Example: Scores on the general CCTST test should be related to scores from other aptitude measures – SAT

Divergent Evidence
Example: Scores on the general CCTST test should not be strongly related to verbal ability measures
External Structure Validity Evidence (cont.)

Evidence of Pretest/Posttest Changes
Example: Scores on the CCTST should increase as students proceed through the program

Evidence from Contrasted Groups
Example: Compare scores on the CCTST for the nursing students with scores on the CCTST for art students.

Predictive Validity Evidence
Example: Examine relationships between scores on the CCTST for nurses and GPA for the same nurses (other criterion measure of critical thinking skills).
External Structure Validity Evidence for the CCTST

• Scores on the CCTST for nurses were correlated with their SATV and SATM scores, GPA, and a GPA based on clinical courses (CGPA) that fostered the development of critical thinking and problem solving skills.

- CCTST, SATM $\rightarrow$ convergent external evidence
- CCTST, GPA & CGPA $\rightarrow$ predictive evidence
- CCTST, SATV $\rightarrow$ divergent external evidence

<table>
<thead>
<tr>
<th>CCTST Score</th>
<th>SATV</th>
<th>SATM</th>
<th>QPA</th>
<th>CQPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>.39*</td>
<td>.47*</td>
<td>.30*</td>
<td>.29*</td>
<td></td>
</tr>
</tbody>
</table>
External Structure Validity Evidence for the CCTST (cont.)

• CCTST scores compared for group who passed NCLEX-RN exam (1st time) vs. group requiring >1 attempt to pass
  - Mean for the group who passed on the first attempt was slightly higher, but no significant difference

• Conclusion: some empirical evidence to support using general CT test to measure nursing students’ CT
Don’t Forget about Reliability and Validity Evidence

Questions?