SURVEY DESIGN

By Hui Bian
Office for Faculty Excellence
My contact information

- Email: bianh@ecu.edu
- Phone: 328-5428
- Location: 1001 Joyner Library, room 1006
- Office hours: 8:00am-5:00pm, Monday-Friday
Survey design

- Educational tests and regular surveys
  - Higher stake tests: such as credentialing tests (there is a passing score involved) systematic and thorough methods are required.
  - We focus on lower stake tests/scales measuring social and psychological phenomena.
References


References


Survey design

- Why we need to develop a new survey/instrument/questionnaire?
  - The measurement scale of interest doesn’t exist.
  - Or the existing survey is not enough for your research.
Survey design

- Goal of survey design
  - Practically is to develop reliable, valid, and usable scales.
Survey design

- Relationship of theory and scales
  - Measuring theoretical/hypothetical constructs
  - Measuring measures without any theoretical foundation
Survey design

- Latent variables
  - Something exist and influence people’s behaviors but they are intangible: self-esteem, depression, anxiety, etc.
  - Multiple items may be needed to capture the essence of those variables.
Survey design

- Latent variables
  - We try to develop a collection of items to represent the level of an underlying theoretical variable.
Survey design

- Theory of Planned Behavior (TPB)
Survey design

Factor1

Indica1

error1

Indica2

error2

Indica3

error3

Indica4

error4

Indica5

error5

Indica6

error6

Factor2

Construct

e7

e8
Survey design

Two measurement theories: Classical Test Theory (CTT) vs. Item Response Theory (IRT)

- **CTT**: observed score = true score + error, linkage between the scales and people being measured
- **IRT**: more focus on item performance, no linkage between scales and sample.
Survey design

- CTT: reliability will be enhanced by more items or better items.
- IRT: reliability will be improved by better items. It also concerns about how strongly an item relates to the latent variable.
Reliability

- Scale reliability: the consistency or stability of estimate of scores measured by the survey over time.
Reliability

- Reliability: in another word, scale reliability is the proportion of variance attributable to the true score of latent variable.

  - Reliability = true score/observed score
Reliability

- CTT: observed score = true score + error
- Measurement error: the more error, the less reliable
  - Systematic error: consistently reoccurs on repeated measures of the same instrument.
  - Problems with the underlying construct (measure a different construct: affect validity)
Reliability

- Random error
  - Inconsistent and not predictable
    - Environment factors
    - Administration variations
Reliability

- Internal consistency
  - Homogeneity of items within a scale
  - Items share a common cause (latent variable)
  - Higher inter-item correlations suggest that items are all measuring the same thing.
Reliability

- Measures of internal consistency
  - Cronbach’s alpha (between 0 and 1)
  - Kuder-Richardson formula 20 or KR-20 for dichotomous items
  - Reliability analysis using SPSS (Cronbach’s alpha): data can be dichotomous, ordinal, or interval, but the data should be coded numerically.
Reliability

- Cronbach’s alpha: A high value of alpha is often used as evidence that the items measure an underlying (or latent) construct. However, a high alpha does not imply that the measure is unidimensional.
Reliability

- Alternate-forms reliability: having the same group of people complete two separate versions of a scale.
  - Correlate the scores from two implementations.
Reliability

- Split-half reliability
  - Compare the first half items to the second half
  - Compare the odd-numbered items with the even-numbered items
Reliability

- **Test-retest reliability (temporal stability)**
  - Give one group of items to subjects on two separate occasions.
  - The scores from the first occasion could be correlated with those from the second occasion.
Reliability

- Rule of thumb: strength of correlation
  - .00-.29 weak
  - .30-.49 low
  - .50-.69 moderate
  - .70-.89 strong
  - .90-1.00 very strong

Reliability

- Rule of thumb: Cronbach’s alpha (DeVellis, 2012)
  - Below .60 unacceptable
  - Between .60 and .65 undesirable
  - Between .65 and .70 minimally acceptable
  - Between .70 and .80 respectable
  - Between .80 and .90 very good
  - Above .90, should consider shortening the scale
Reliability

How to control reliability
Validity

Validity is the most fundamental consideration in survey design.

- Validity: the scale truly measures what it is supposed to measure.
- The reliability of the outcomes depends on the soundness of the measures.
Validity

- Handbook of test development: “Effective test development requires a systematic, well-organized approach to ensure sufficient validity evidence to support the proposed inferences from the test scores.”
Validity

- Types of validity (DeVellis, 2012)
  - Content validity
  - Criterion-related validity
  - Construct validity
Standards for test development

- Standards for Educational and Psychological Testing 1999
  - American Educational Research Association (AERA)
  - American Psychological Association (APA)
  - National Council on Measurement in Education (NCME)
Sources of validity evidence

- Evidence based on test content
  - Test content refers to the themes, wording, and format of the items and guidelines for procedures regarding administration.
Sources of validity evidence

- Evidence based on response processes
  - Target subjects.
  - For example: whether the format more favors one subgroup than another group;
  - In another word, something irrelevant to the construct may be differentially influencing performance of different subgroups.
Sources of validity evidence

- Evidence based on internal structure
  - The degree to which the relationships among instrument items and components conform to the construct on which the proposed relationships are based (Confirmatory factor analysis).

- Evidence based on relationships to other variables
  - Relationships of test scores to variable external to the test.
Content validity

- Definition of the construct being examined
  - Adequacy of the content
  - Relevance of the content
  - Concerns: how to know the selected items are representative and how to know the items capture the aspects of the construct.
It is critical to establish accurate and comprehensive content for an instrument. Selection of content is based on sound theories and empirical evidence or previous research. A content analysis is recommended. It is the process of analyzing the structure and content of the instrument. Two stages: development stage and appraisal stage.
Content-related validity evidence

- Instrument specification
  - Content of the instrument
  - Number of items
  - The item formats
  - The desired psychometric properties of the items
  - Items and section arrangement (layout)
  - Time of completing survey
  - Directions to the subjects
  - Procedure of administering survey
Content-related validity evidence

- Content evaluation (Guion, 1977)
  - The content domain must be with a generally accepted meaning.
  - The content domain must be defined unambiguously.
  - The content domain must be relevant to the purpose of measurement.
  - Qualified judges must agree that the domain has been adequately sampled.
  - The response content must be reliably observed and evaluated.
Content-related validity evidence

- Content evaluation
  - Clarity of statements
  - Relevance
  - Coherence
  - Representativeness
Criterion-related validity

- An scale is required to have empirical association with a criterion or gold standard.
- Collect data from new developed instrument and from criterion.
Criterion-related validity

- Evidence based on relationships to other variables.
  - External variables can be measures of criteria, other tests measuring the same/related/different constructs.
Criterion-related validity

- Convergent and discriminant evidence
  - Convergent: relationship of test scores and similar constructs.
  - Discriminant: relationship of test scores and different constructs.
Construct validity

- In order to demonstrate construct validity, we should provide evidence that the scale measures what it is supposed to measure.
- Construct validation requires the compilation of multiple sources of evidence.
  - Content validity
  - Item performance
  - Criterion-related validity
Construct validity

- Validity studies should address both the internal structure of the test and external relations of the test to other variables.
  - Internal structure: subdomains or subconstruct
  - External relations: relationships between test measures and other constructs or variables.
Evidence against validity

- Construct-irrelevant variance: the degree to which test scores are affected by processes that are extraneous to the construct.
Evidence against validity

- Construct-irrelevant variance
  - Systematic error
  - May increase or decrease test scores

\[ y = t + e_1 + e_2 \]

- $y$ is the observed score.
- $t$ is the true score.
- $e_1$ is random error (affect reliability).
- $e_2$ is systematic error (affect validity).
Evidence against validity

- Construct underrepresentation: the degree to which a test fails to capture important aspects of the construct.
  - It is about fidelity.
  - It is about the dimensions of studied content.
Survey design

- Validation is the process of developing valid instrument and assembling validity evidence to support the statement that the instrument is valid.
- Validation is on-going process and validity evolves during this process.
Survey design

- 12 steps of test development
  - Overall plan
  - Content definition
  - Test specifications
  - Item development
  - Test design and assembly
  - Test production
Survey design

- 12 steps of test development
  - Test administration
  - Scoring test responses
  - Passing scores
  - Reporting test results
  - Item banking
  - Test technical report
Steps for instrument development

1. Determine what you want to measure
2. Generate an item pool
3. Determine the format for items
4. Expert review of initial item pool
5. Add social desirability items
6. Pilot testing and item analysis
7. Administer instrument to a larger sample
8. Evaluate the items
9. Revise instrument

DeVellis (2012); Fishman & Galguera (2003); Pett, Lackey, & Sullivan (2003)
Determine what you want to measure

- What will the instrument measure?
- Will the instrument measure the construct broadly or specifically, for example: self-efficacy or self-efficacy of avoiding drinking?
- Do all the items tap the same construct or different one?
- Use sound theories as a guide.
- Related to content validity issues.
Generate an item pool

- It is also related to content validity
- Choose items that truly reflect underlying construct.
- Borrow or modify items from already existed instruments (they are valid and reliable).
- Redundancy: more items at this stage than in the final scale. A 10-item scale might evolve from a 40-item pool.
Generate an item pool

- Writing new items (Focus group)
  - Wording: clear and inoffensive
  - Avoid lengthy items
  - Consideration of reading difficulty level
  - Avoid items that convey two or more ideas
  - Be careful of positively and negatively worded items
Determine the format for items

- Items include two parts: a stem and a series of response options.
- Number of response options
  - A scale should discriminate differences in the underlying attributes
  - Respondent’s ability to discriminate meaningfully between options
Determine the format for items

- Number of response options
  - Equivocation: you have neutral as a response option
- Types of response format
  - Likert scale
  - Binary options
  - Selected-response format (multiple choice format)
Develop the form of an instrument

- Components of instrument
  - Format (font, font size)
  - Layout (how many pages)
- Instructions to the subjects
- Wording of the items
- Response options
- Number of items

58
Expert review of initial item pool

- Purpose of expert review is to maximize the content validity.
- Panel of experts are people who are knowledgeable in the content area.
- Item evaluation
  - How relevant is each item to what you intend to measure?
  - Items’ clarity and conciseness
  - Missing content
- Final decision to accept or reject expert recommendations
  - It is developers’ responsibility
Social desirability items

- It is the tendency of subjects to respond to test items in such a way as to present themselves in socially acceptable terms in order to gain the approval of others.
- Individual items are influenced by social desirability.
- 10-item measures by Strahan and Gerbasi (1972)
Questions before pilot test

- Do those selected items cover the content completely?
- How many items should there be?
- How many subjects do we need to pilot test this instrument?
Pilot test

- Also called tryout or field test
- Purposes
  - Identify, remove or revise bad items
  - Any problems related to item contents and formats
  - Data collection procedure
  - Prevents researchers from spending valuable resources on a study using not valid or reliable instrument
  - Determine the amount of time to complete the instrument
Pilot test

- Sample size: one tenth the size of the sample for the major study.
- People who participate in the pilot test can not be in the final study.
Item analysis

Item analysis:
- it is about item performance.
- Reliability and validity concerns at item level
- As means of detecting flawed items
- Help select items to be included in the test or identify items that need to be revised.
- Item selection needs to consider content, process, and item format in addition to item statistics.
Item analysis

- **Item response theory (IRT)**
  - Focuses on individual items and their characteristics.
  - Reliability is enhanced not by redundancy but by indentifying better items.
  - IRT items are designed to tap different degrees or levels of the attribute.
  - The goal of IRT is to establish item characteristics independent of who completes them.
Item analysis

- IRT concentrates on three aspects of an item’s performance.
  - Item difficulty: how hard the item is.
  - Item discrimination: its capacity to discriminate.
    - A less discriminating item has a larger region of ambiguity.
  - Guessing or false positive
Item difficulty

- Knowing the difficulty of the items can avoid making a test so hard or so easy.
- The optimal distribution of difficulty is normal distribution.
- For dichotomous variable: (correct/wrong)
  - The rate of wrong answers: 90 students out of 100 get correct answers, item difficulty = 10%.
## Item difficulty

<table>
<thead>
<tr>
<th>Items</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>a59_9</td>
<td>2.04</td>
</tr>
<tr>
<td>a59_10</td>
<td>1.77</td>
</tr>
<tr>
<td>a59_11</td>
<td>1.93</td>
</tr>
<tr>
<td>a59_12</td>
<td>1.95</td>
</tr>
<tr>
<td>a59_13</td>
<td>1.60</td>
</tr>
<tr>
<td>a59_14</td>
<td>1.58</td>
</tr>
<tr>
<td>a59_15</td>
<td>1.61</td>
</tr>
<tr>
<td>a59_16</td>
<td>1.87</td>
</tr>
<tr>
<td>a59_17</td>
<td>2.75</td>
</tr>
<tr>
<td>a59_30</td>
<td>1.63</td>
</tr>
</tbody>
</table>

Four-point scale: 1 = Strongly agree, 2 = Agree, 3 = Disagree, 4 = Strongly disagree.

Less difficult → More difficult

Strongly agree → Strongly disagree
Item difficulty

Difficulty distribution on a four-point scale
Item reliability

- Inter-item consistency: scale reliability if item deleted
  - Deletion of one item can increase overall reliability. Then that item is poor item.
  - We can obtain that statistic from Reliability Analysis (SPSS)
## Item reliability

### Item-Total Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>a57a Specific goal setting: get moderate exercise</td>
<td>12.09</td>
<td>6.498</td>
<td>.411</td>
<td>.784</td>
</tr>
<tr>
<td>a57b Specific goal setting: get more vigorous exercise</td>
<td>11.96</td>
<td>6.525</td>
<td>.355</td>
<td>.792</td>
</tr>
<tr>
<td>a57c Specific goal setting: reduce/avoid drinking alcohol</td>
<td>11.97</td>
<td>5.824</td>
<td>.670</td>
<td>.751</td>
</tr>
<tr>
<td>a57d Specific goal setting: reduce/avoid cigarette smoking</td>
<td>12.02</td>
<td>5.863</td>
<td>.670</td>
<td>.752</td>
</tr>
<tr>
<td>a57e Specific goal setting: reduce/avoid marijuana use</td>
<td>11.96</td>
<td>5.897</td>
<td>.635</td>
<td>.756</td>
</tr>
<tr>
<td>a57f Specific goal setting: eat healthier food</td>
<td>12.15</td>
<td>6.600</td>
<td>.400</td>
<td>.785</td>
</tr>
<tr>
<td>a57g Specific goal setting: relax or reduce stress</td>
<td>12.15</td>
<td>6.649</td>
<td>.377</td>
<td>.788</td>
</tr>
<tr>
<td>a57h Specific goal setting: get more sleep</td>
<td>12.16</td>
<td>6.994</td>
<td>.221</td>
<td>.804</td>
</tr>
<tr>
<td>a57i Specific goal setting: reduce/avoid driving after drinking/use drug</td>
<td>11.99</td>
<td>5.838</td>
<td>.668</td>
<td>.752</td>
</tr>
<tr>
<td>a57j Specific goal setting: lose or gain weight</td>
<td>12.12</td>
<td>6.862</td>
<td>.258</td>
<td>.801</td>
</tr>
</tbody>
</table>

### Reliability Statistics

<table>
<thead>
<tr>
<th>Cronbach's Alpha</th>
<th>N of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>.795</td>
<td>10</td>
</tr>
</tbody>
</table>
Item discrimination

- Item validity
  - Correlation of each item’s response with the total test score minus the score for the item in question.
  - Corrected item-total correlation.
**Item discrimination**

Higher value of corrected item-total correlation is desired.

<table>
<thead>
<tr>
<th>Item: Specific goal</th>
<th>Scale Mean if Item Deleted</th>
<th>Scale Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>a57a</td>
<td>12.09</td>
<td>6.498</td>
<td>.411</td>
<td>.784</td>
</tr>
<tr>
<td>a57b</td>
<td>11.96</td>
<td>6.525</td>
<td>.355</td>
<td>.792</td>
</tr>
<tr>
<td>a57c</td>
<td>11.97</td>
<td>5.824</td>
<td>.670</td>
<td>.751</td>
</tr>
<tr>
<td>a57d</td>
<td>12.02</td>
<td>5.863</td>
<td>.670</td>
<td>.752</td>
</tr>
<tr>
<td>a57e</td>
<td>11.96</td>
<td>5.897</td>
<td>.635</td>
<td>.756</td>
</tr>
<tr>
<td>a57f</td>
<td>12.15</td>
<td>6.600</td>
<td>.400</td>
<td>.785</td>
</tr>
<tr>
<td>a57g</td>
<td>12.15</td>
<td>6.649</td>
<td>.377</td>
<td>.788</td>
</tr>
<tr>
<td>a57h</td>
<td>12.16</td>
<td>6.994</td>
<td>.221</td>
<td>.804</td>
</tr>
<tr>
<td>a57i</td>
<td>11.99</td>
<td>5.836</td>
<td>.668</td>
<td>.752</td>
</tr>
<tr>
<td>a57j</td>
<td>12.12</td>
<td>6.862</td>
<td>.256</td>
<td>.801</td>
</tr>
</tbody>
</table>
Item discrimination

- Item validity
  - A bell-shaped distribution with its mean as high as possible.
  - Higher correlation for an item means people with higher total scores are also getting higher item score.
  - Items with low correlation need further examination.
Administer instrument to a larger sample

- Sample size: no golden rules
  - 10-15 subjects/item
  - 300 cases is adequate
  - 50 very poor
  - 100 poor
  - 200 fair
  - 300 good
  - 500 very good
  - 1000 or more excellent
Administer instrument to a larger sample

- Administration threats to validity
  - Construct underrepresentation
  - Construct irrelevant variance
- Efforts to avoid those threats
  - Standardization
  - Administrator training
Evaluate items and revise instrument

- Item analysis: item performance
- Factor analysis

- Purposes
  - determine how many latent variables underlie a set of items.
  - label identified factors to help understand the meaning of underlying latent variables.
Evaluate items and revise instrument

- Factor analysis
  - Exploratory factor analysis: to explore the structure of a construct.
  - Confirmatory factor analysis: confirm the structure obtained from exploratory factor analysis.
Optimize scale length

- Effects of dropping items
  - Reliability
  - Construct underrepresentation
  - Construct irrelevant variance
THANK YOU