Developing a Measurement Strategy

Steps of the Measurement Process

1. Decide what to measure
   - Based on the variables specified in your hypotheses
2. Identify ways of measuring each variable
   - Self-report, behavioral, physiological, implicit measures
   - Level of measurement can be nominal, ordinal, interval, or ratio (continuous/discrete)

Advantages of Self-Report Measures

- Behavioral Self-Reports, assess how people:
  - acted in the past (retrospective reports)
  - currently act (prospective reports)
  - believe they would act (hypothetical reports)
- Most direct way to get some kinds of information
  - Moods, beliefs, thought processes
  - Behaviors that cannot be directly observed
- Are easy to collect
  - Can be administered in groups or on-line
  - Research assistants need little training
  - Are inexpensive

Disadvantages of Self-Reports

- People may remember the past incorrectly
- People do not necessarily know what affects their behavior
  - People may not accurately know their inner states, beliefs, etc.
- Depend on respondents' verbal skills
  - Some populations have low skills

Examples of Behavioral Measures

- Frequency of response
- Rate of response
- Response latency
- Accuracy of performance
- Persistence of behavior
- Total number of different behaviors performed

Behavioral Measures

- Approximate behavior by measuring commitment to perform a behavior
  - Example: Do participants agree to spend two hours with new international students on campus?
    - Do not actually have to meet with new students
    - How hard do animals work to gain access to an enrichment item or a sheltered area?
    - How many bar presses does a rat make to obtain a drug infusion?

Experience Sampling

- Participants are asked to record behaviors as they are performed or soon after
  - Researchers can control timing of recording by use of "beepers"

Advantages of Behavioral Measures

- Shows what people actually do, as opposed to what they say they do
  - Reality vs. perception
- Can be used without people's awareness
- Allows study of behaviors that are automatic
  - Bypasses any tendency to "edit" responses
- People may find "behaving" more engaging than self-report

Disadvantages of Behavioral Measures

- Does not indicate *why* people engaged in the behavior
- Behaviors can be highly situation specific
  - Results may not generalize to other situations
- Need a systematic way to classify observed behaviors
  - Observers must be trained to accurately observe and record behaviors
- Can be expensive

Physiological Measures

Measures used to assess people's biological responses to stimuli, such as

- EEG
- fMRI
- blood pressure
- heart rate
- electrical conductivity of the skin
- sexual arousal

Advantages of Physiological Measures

- Provide most direct means of quantifying biological responses
- Provide highly precise measures
  - Error of estimate is known and can be taken into account
- Assess behaviors that are usually not under people's voluntary control

Disadvantages of Physiological Measures

- Depend on sophisticated electronic and mechanical equipment
  - Equipment is expensive to buy and maintain
  - Equipment operators must be trained to both use and maintain equipment
- A lot of information is collected
  - Some is extraneous
  - Can be hard to separate "signal" from "noise"

Disadvantages of Physiological Measures

- Devices can be intrusive and constrain people's freedom of movement
- Research participants can find task long and boring
  - May result in low quality data

Disadvantages of Physiological Measures

- Difficult to establish construct validity
- To do so, must show that
  - a particular physiological state results in a particular physiological response (convergent validity)
  - the physiological factor varies only in response to changes in the psychological state (discriminant validity)

Implicit Measures

- Explicit measures assess responses that research participants can think about and consciously control
- In contrast, implicit measures assess responses people make without thinking
  - Therefore cannot easily be edited

Implicit Association Task (IAT)

- You may believe that women and men should be equally associated with science, but your automatic associations could show that you (like many others) associate men with science more than you associate women with science.
- 1.
  - Press "X" if image represents males or non-science
  - Press "M" if image represents females or science
- 2.
  - Press X if image is female or non-science
  - Press M if image is male or science
- Faster responses for 2. (male science/female non-science) than for 1. (male non-science/female science) indicate a stronger association of males than of females with science

Advantages of Implicit Measures

- Participants do not have time to edit responses
  - Therefore, IAT is assumed to tap people's true beliefs and feelings
  - Addresses problem of people's inability to access their own motivations, beliefs, and feelings
- Can aid in developing operational definitions for theory testing
  - Can test propositions that distinguish between conscious and unconscious processes

Limitations of Implicit Measures

- Conclusions must be drawn carefully
  - IAT indicates relative but not absolute preferences
- Measures require that participants concentrate on the task
- Requires appropriate equipment and software
- Stimuli used might influence the nature of the concept being studied
  - Responses may represent reactions to concept (female) *or* characteristics of the stimulus used to represent category (high pitched voice, smaller stature)

Manifest Variables Vs. Hypothetical Constructs

- Hypothetical constructs cannot be directly observed
- So….
- Manifest variables: Directly observable variables, such as
  - physical characteristics of people and objects
  - behaviors
  - physiological responses
  - answers to questions

Assumptions

- The manifest variable represents the hypothetical construct of interest
- The hypothetical construct is causing the presence and strength of the manifest variable
- Can infer the strength of the hypothetical construct from the strength of the manifest variable

Example

- Hypothetical construct: Self-esteem
- Manifest variables: scores on measures like Rosenberg's self report self-esteem scale, State self-esteem, eye contact, predicted level of success
  - Assumption is that behaviors (e.g., risk-taking, business success) predicted by self-report scores, eye-contact etc. are really being influenced by underlying self-esteem
  - i.e., Self-esteem causes success/risk-taking
  - Self report scores themselves do not cause success/risk-taking

Measures of Hypothetical Constructs

Psychometric tests:
- provide information about a particular individual's score on a construct
- normed with large, representative samples
- indicate where that individual stands on the construct relative to other people

Research measures:
- provide information about the mean scores of groups of people
- indicate the relationships between constructs
- can be used to compare groups of people

Psychometric tests can be used as research measures but research measures cannot be used as psychometric tests

Reliability

The degree of consistency of a measure
- A perfectly reliable measure gives the same result every time it is applied to the same person or thing, barring changes in the variable being measured

Validity

The degree of accuracy of a measure
- A perfectly valid measure assesses
  - the construct it is supposed to assess
  - all aspects of the construct
  - only that construct

Measurement Error

Link between reliability and validity

Components of an observed score that one did not want to assess, but assessed anyway because of the imperfections of the measuring instrument

- Random error:
  - Fluctuates each time a measurement is made
    - causes observed score to fluctuate
    - leads to instability of measurement
    - lowers reliability estimates
- Systematic error:
  - Present every time a measurement is made
    - affects extent to which the observed score is an accurate indicator of the true score

Relationship between Reliability and Validity

- Measure has to be reliable in order to be valid
- However, measure can be reliable but not valid (e.g., clock that always runs fast)

Assessing Reliability

Consistency across time

- Assessed by test-retest reliability
  - Correlation computed between scores assessed at Time 1 and Time 2
  - Scores need not be exactly the same but should fall in same rank order

Assessing Reliability

Consistency across forms

- Assessed by alternate forms reliability
  - Correlation computed on scores on two different forms of the same measure

Assessing Reliability

Internal consistency: Degree to which responses to the items on a measure are similar

- Assessed by split-half reliability
  - Correlation computed between mean scores on one half of the items and mean scores on the other half of the items
- Also assessed by Chronbach's alpha
  - Alpha is a function of the mean correlation of all the items with one another
  - Can be interpreted like a correlation coefficient

Limitations of Chronbach's Alpha

- A high alpha does not mean a measure is unidimensional
  - Scale dimensionality is assessed by factor analysis
- If scale items are too similar, alpha will be artificially high
- Alpha increases as number of items on scale increases
  - Does not necessarily mean long scales are better

Standards for Reliability

Higher reliability coefficients/correlations indicate better reliability

- Minimum internal consistency *alpha* = 0.70
- Minimum test-retest reliability $r = 0.50$

Consistency Across Raters

Assessed by inter-rater reliability

- Correlation computed on two raters' ratings (of same behavior)
- Cohen's kappa: Used when raters put behavior into categories
  - Takes into account the likelihood that agreements would occur by chance

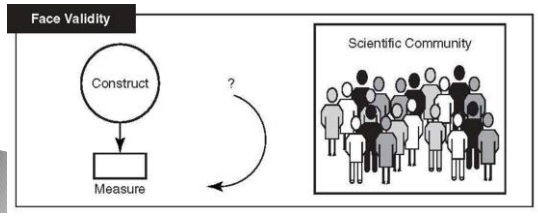
		Rater 1 (C.P.)			
		L	R	N	
Rater 2 (S.B.)	L	51	0	4	55
	R	2	51	3	56
	N	2	0	139	141
		55	51	146	241

$$\text{kappa} = \frac{(p_o - p_e)}{(1 - p_e)}$$

where p_o is the observed proportion of agreement and p_e is the proportion expected by chance.

Validity

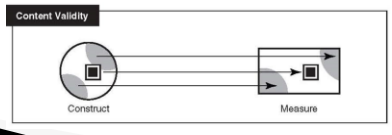
- Is inferred from available evidence
- Is expressed by degree
  - high
  - moderate
  - low
- No one type of validity necessarily indicates better evidence of validity than the other types



Content Validity

Extent to which the content of a measure adequately assesses all aspects of the construct being measured

- Content must be
  - relevant (it assesses only the trait of interest and little else)
  - representative (it covers all topics it is supposed to)



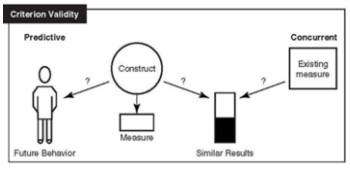
~~External Validity~~

- **Criterion validity**

Extent to which scores on a measure correlate with relevant factors or criteria outside, or external to, the measure

- That is, scores are shown to be conceptually relevant to the construct being measured
- Chosen on basis of the theory underlying that construct
- Concurrent validity
  - Correlate with other existing measures

~~Substantive validity~~



- **Criterion validity**
  - Predictive validity

Are research results using a measure consistent with hypotheses derived from the theory of the construct being measured?

- People who score differently on a construct should respond differently to situational variables
  - E.g., higher in empathy should display more helping
- Scores on measure behave same way that actual behaviors/constructs should behave in given situation
  - E.g., hostility scores increase with crowding

Structural Validity

Extent to which the dimensionality of a measure reflects the dimensionality of the construct it is measuring

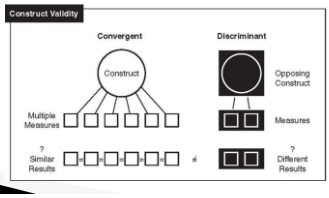
- Unidimensional constructs should show single dimension
- Multidimensional construct should show multiple dimensions

Construct Validity

Convergent Validity

The extent to which evidence comes together, or converges, to indicate the degree of validity of a measure

- This evidence addresses whether the measure is assessing what it is designed to assess
- Different measures of same construct produce same results



Construct Validity cont.

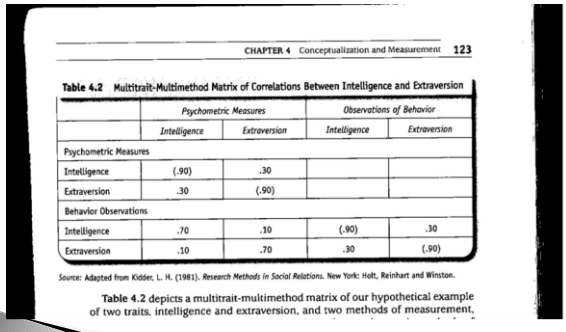
Discriminant Validity

Divergent Validity

Evidence that a measure is not measuring something it is not supposed to

- Assessed by correlating scores on a measure with scores on irrelevant constructs
  - Low correlation is evidence of good discriminant validity

Multitrait-multi-method matrix



CHAPTER 4 Conceptualization and Measurement 123

Table 4.2 Multitrait-Multimethod Matrix of Correlations Between Intelligence and Extraversion

	Psychometric Measures		*Observations of Behavior*	
	Intelligence	*Extraversion*	*Intelligence*	*Extraversion*
Psychometric Measures				
Intelligence	(.90)	.30		
Extraversion	.30	(.90)		
Behavior Observations				
Intelligence	.70	.10	(.90)	.30
Extraversion	.10	.70	.30	(.90)

Source: Adapted from Kidder, L. H. (1981). *Research Methods in Social Relations.* New York: Holt, Reinhart and Winston.

Table 4.2 depicts a multitrait-multimethod matrix of our hypothetical example of two traits, intelligence and extraversion, and two methods of measurement,

Identify Which Type of Validity is Described in Each of the Following

- The Fear of Negative Evaluation Scale (FNE) has items such as "I rarely worry about seeming foolish to others" and "I am afraid that people will find fault with me"
- People high in FNE experienced anxiety in an evaluative setting; people low in FNE do not

Identifying Validity, cont.

- Exam One for a course in Evolutionary Psychology covered three chapters in the textbook. The exam contained eight items from each chapter
- Scores on the Counterbalanced F-Scale (a measure of authoritarianism) correlates with other measures of authoritarianism
- A measure of activity on the playground correlates with teacher's rankings of child activity levels

Identifying Validity, cont.

- Scores on the Fear of Fat Scale are unrelated to measures of political conservatism
- A researcher wanted to show that their IQ test was not a measure of spatial ability. They had students complete their test and a standard object rotation test. Scores on the two measures were uncorrelated

Identifying Validity, cont.

- A researcher developed the Dating and Assertion Questionnaire. Scores on the measure predicted who had the most successful long-term relationship later in life
- Two researchers developed an Attitudes About Make-up Scale. Results based on samples of older women, younger women, and men showed similar evidence of reliability and validity

Differential Validity

A condition that exists when a measure is more valid for assessing a construct for members of one group than for members of another group

e.g., Openness from Big 5 in U.S. versus Chinese samples

Check whether

- group means on measure differ when the theory of the construct predicts no such differences
- different correlations emerge with measures of the same construct or related constructs for different groups

More about Research Measures

Developed research measures

- have undergone validation research
- Thus, information is available on their reliability, validity, and other characteristics

Ad-hoc research measures

- are created for use in a particular study
- are not typically assessed for reliability and validity
- can't easily be assed for how accurately they assess the construct
- should be avoided when possible

Evaluating Measures

Consider which theoretical background the developers used

- Did they try to develop a comprehensive measure that covers many aspects of the construct?
- Did they focus on a specific aspect of the construct?
- Is their strategy appropriate for your purposes?

Evaluating Measures

Review the quality of the development

- Were appropriate samples used in tests for reliability and validity?
- What norms were used to compare sample statistics to the larger norming sample?
- Are norms available for different population groups?
  - Do your results mirror those results?

Evaluating Measures

Look for measures that are unlikely to elicit response bias

- The tendency for a person to respond to a measure for reasons other than the response being a reflection of the construct being assessed by the content of the measure

Comparing Measures

- Good measures are comprised of a large number of characteristics
  - Realistically, no one measure can be high on all of the possible dimensions
- In choosing a measure, consider which characteristics are most essential for *your* research
  - Those characteristics should be given more weight in evaluating your measure