Meeting 7

Evaluability

&

Measurement

Example 1: evaluability

- An enthusiastic researcher traveling in India comes upon an innovative educational program.
- The program takes volunteers from high-end corporations and gets them to supplement school classes and work as role models.
- The researcher decides to evaluate this innovation.
- Good idea or not? Why?

- Is the program systematic and stable?
  - Well structured and described.
- Is there agreement on what the program is doing and trying to achieve?
  - Objectives and criteria for success.
  - Which counter-factual?
- Are there big restrictions on the potential evaluation?
- Is there enough time and money for an evaluation?
- Mature, stable programs are best suited to evaluation.

Example 2: measurement

- You are evaluating a school intervention that aims to improve students' educational outcomes.
- How should/ could you measure this?
  - Class test / home work scores
  - Grades
  - Standardized test scores
  - Knowledge tests *x* months later.
  - Test ability to apply integrate knowledge (e.g.,. Accounting test to test math scores).
- What else do you want to measure?
  - Hint: look at program theory / logic model.
  - E.g., effect of education on social attitudes / aspirations.

Example 2, continued

- I decide to use a test bank whose questions a specialist assures me measures ability.
- Does this test really measure academic achievement? I.e., is it a valid measure?
  - Does it make sense as a measure? (face validity)
  - Does it cover key concepts included in academic achievement? (content validity)
  - Do students with very different test scores have different levels of knowledge? (Discriminant/concurrent validity)
  - Do students who do very well / very badly on the test line up other measures like literacy, etc. (convergent validity)
  - Compare test scores among students with same grades. (convergent validity)
  - Do students with low test scores later do badly in relates tests/activities (predictive validity).

Example 2, continued

- But I know that students' test performance is noisy. What should I do to ensure that this test nonetheless works?
  - Test the same student on different days with different randomly selected questions (parallel form reliability).
  - Split the class in half and compare test scores from similar questions. (split half reliability)
  - Compare scores on individual questions within students. (internal consistency / intra-rater reliability)

  →Will determine how often and with how many questions I test students.

Measurement

- How you codify or quantify your hypotheses into data points
  - Develop measurable research questions/ hypotheses
  - Conceptualize, operationalize, devise measurement strategies
  - Demonstrate validity and reliability (how good)

Measurement Process

- Going from program goals to measurement
  - Goals: often vague, broad, hard to measure
    - Turn into research questions/directional hypotheses
    - One program goal may → multiple research questions
- Operationalize into definitions or variables
  - Be specific, narrow and concrete (make operational)
  - Refer to time frame (over what time period?)
  - Refer to counterfactual (compared to what?)
- Develop measures of variables

Example 3: Moving to Opportunity

- Designed to investigate the impact of living in bad neighbourhoods on outcomes
- Gave some residents of public housing projects chance to move out
- Two treatments:
  - Voucher for private rental housing (Section 8)
  - Voucher for private rental housing restricted for use in 'good' neighbourhoods (Section 8 with restrictions)
  - + Control
- No-one forced to move so imperfect compliance – 60% and 40% did use it

How would measure the impact?

- Goals to measures:
  - Breakdown goals into broad categories:
    - Economics: improved work life / opportunities.
    - Education: improved education for children.
    - Psychological: increased happiness.
    - Health: improved health.
- Operationalize into variables:
    - Economics: e.g., wages, hours worked.
    - Education: e.g., grades, continuation.
    - Psychological: e.g., increased "life satisfaction".
    - Health: improved weight, blood pressure, cholesterol.
- Develop measures:
    - Economics: wages in main job, income in last year, hours work typical week.
    - Education: completed high school? Grades on standardized tests.
    - Psychological: Life wellbeing questions over the last week, month, year.
    - Health: Measure changes in biometrics, blood results over 1-2 years.

Step Back

- Move back and forth between measurement/ operationalization and theory
  - Real World
  - Theoretical World (idea of real world)
  - Operational World (representation of real world)
- Translation → Operationalization
  - Translating a construct (idea or theory) into its manifestation
  - Take idea and describe as a series of operations or procedures
  - Instead of idea in your mind, becomes public entity (critique-able)

Social science research uses

> "imperfect indicators of theoretical concepts to discover imperfect associations"

Measurement quality

- Criteria for Measurement Quality
  - Precision: Fineness of distinctions between attributes composing a variable or measure
    - 3 item scale vs 10 item scale
    - Greater precision usually better than less precision
    - But consider meaningfulness and possibility of precision
      - Can you use a 100 point scale for amount of agreement?
  - Accuracy: Correctness
    - E.g., weighing scale for weight.
  - Reliability: Measure yields the same result repeatedly (consistency)
    - If moving target, can't be useful
  - Validity: Extent to which a measure adequately reflects the real meaning of the construct under consideration
    - Does it measure what it is supposed to measure?

Mundane: Level of Measurement

- Nominal scale
  - Categories that are exhaustive and mutually exclusive (labels)
    - How do you feel? Good/bad.
- Ordinal scale
  - Logical order to variables
    - Rate your happiness on a scale of 1 to 7.
  - But magnitude of difference between items unclear
    - Rankings, scales of agreement
- Interval scale
  - Know exact difference between objects on the scale
    - Rate your happiness on a scale of 0 to 10.
    - Continuous variables, data
    - Unit of measurement is standardized and replicable

Types of Reliability

- Consistency over time
  - Test-retest: individual responds similarly to same instrument repeatedly over time.
- Internal consistency
  - Parallel form reliability (creating parallel forms of a measure)
    - Divide large set of assessment items into half randomly
      - How related are two halves?
  - Split half
    - Split sample randomly in half and see if you get same results.
  - Within instrument
    - How internally consistent the items comprising the measure are
      - Do people who do well in 1 section, do well on others?
      - Items hang together?
    - Average inter-item, item-total correlations

Types of Reliability

- For qualitative assessments:
  - Inter-rater reliability
    - Agreement between raters (human judgment)
  - Intra-rater reliability
    - Consistency in rating within same rater.

Validity

- Validity: “best available approximation to the truth of a given proposition, inference, or conclusion”

Four Types of Validity

- Statistical Conclusion Validity
  - Is there a statistically valid relationship between the measure and concept?
- Internal Validity
  - Assuming there is a relationship, is it causal?
- External Validity
  - Assuming that there is a causal relationship between the measures and the constructs of cause and effect, can we generalize these effects to other people, places, times, situations?
- Construct Validity
  - Assuming that there is a causal relationship, can we claim that our measures of the variables represented the construct and that the construct represented the reality?
    - Did our measure of academic achievement capture academic achievement? Is our idea of academic achievement accurate (reflected in real world)?

Construct Validity 1: Translation Validity

- Translation Validity
  - Whether operationalization is a good reflection of the construct
    - Definitional: Assumes you have a good, detailed definition of the construct and that you can then check your operationalization against it
  - *Face Validity*
    - You look at operationalization and see whether "on its face" it seems like a good translation of the construct
    - Pretty weak to demonstrate construct validity
      - But it's a good place to start
  - *Content Validity*
    - Check operationalization against the relevant content domain for the construct
      - Make sure your measure covers the relevant aspects of a construct
      - E.g., if the concept of educational attainment includes reading and writing, test should cover both.

Construct Validity 2: Criterion-Related Validity

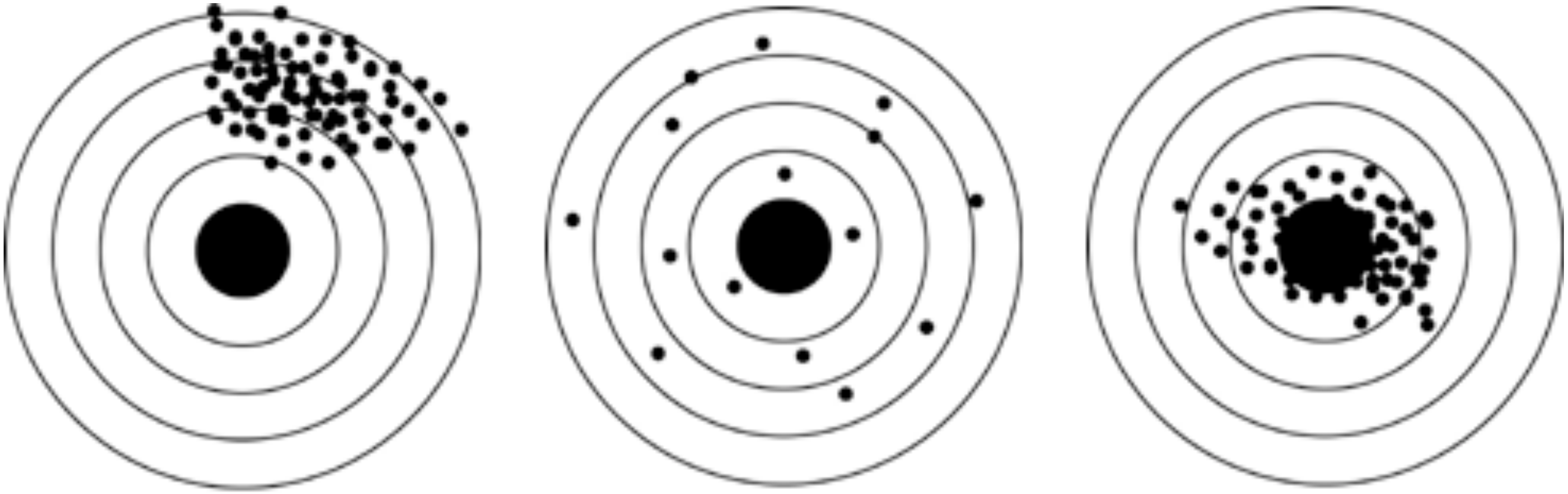
- Criterion-Related Validity
  - Whether operationalization behaves the way it should given your theory of the construct and how it works
  - Relational: Assumes that operationalization should function in predictable ways in relation to other operationalizations

1. *Discriminant Validity* (Concurrent)
   - Operationalization’s ability to distinguish between groups that it should theoretically be able to distinguish between
     - Skills test can distinguish those completed training and those who didn’t
     - Standardized test can distinguish literate vs illiterate
2. *Convergent Validity* (Divergent)
   - Degree to which operationalization is similar to or converges with other operationalizations that it should theoretically be similar to
     - Grades and standardized tests should be related, if not…?

Construct Validity 2: Criterion-related validity (cont'd)

- Criterion-related validity (cont'd)
  3. *Predictive Validity*
     - Operationalization's ability to predict something that it should theoretically be able to predict
       – E.g., high school test scores should be able to predict whether you go on to college and if so how you do.
       – If test scores measure education, should be related to how you use medical care (based on other research).

Validity and Reliability Compared



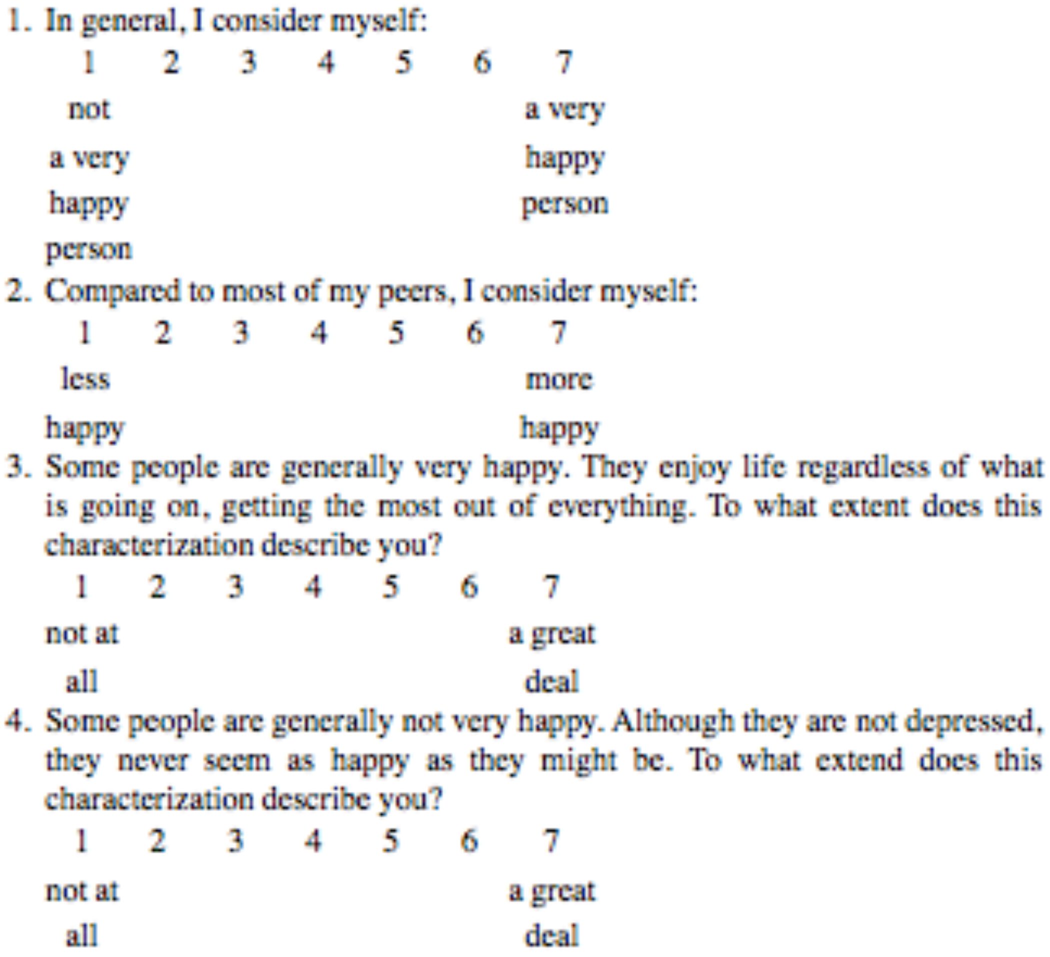
Reliable but Not Valid

Valid but Not Reliable

Valid and Reliable

Example: measuring happiness

Lyubomirksy and Lepper



1. In general, I consider myself:

1	2	3	4	5	6	7
not						a very
a very						happy
happy						person
person						

2. Compared to most of my peers, I consider myself:

1	2	3	4	5	6	7
less						more
happy						happy

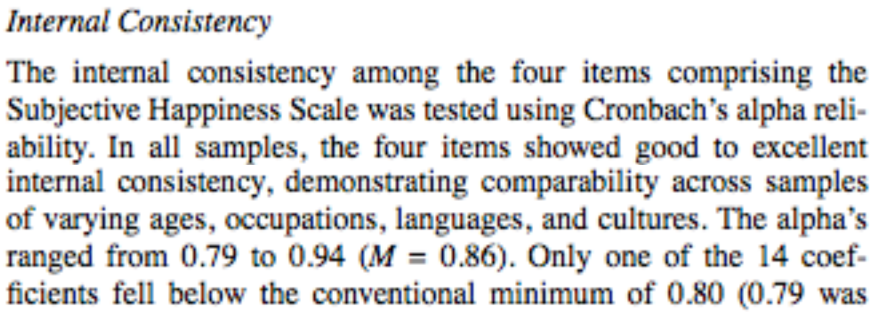
3. Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?

1	2	3	4	5	6	7
not at						a great
all						deal

4. Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extend does this characterization describe you?

1	2	3	4	5	6	7
not at						a great
all						deal

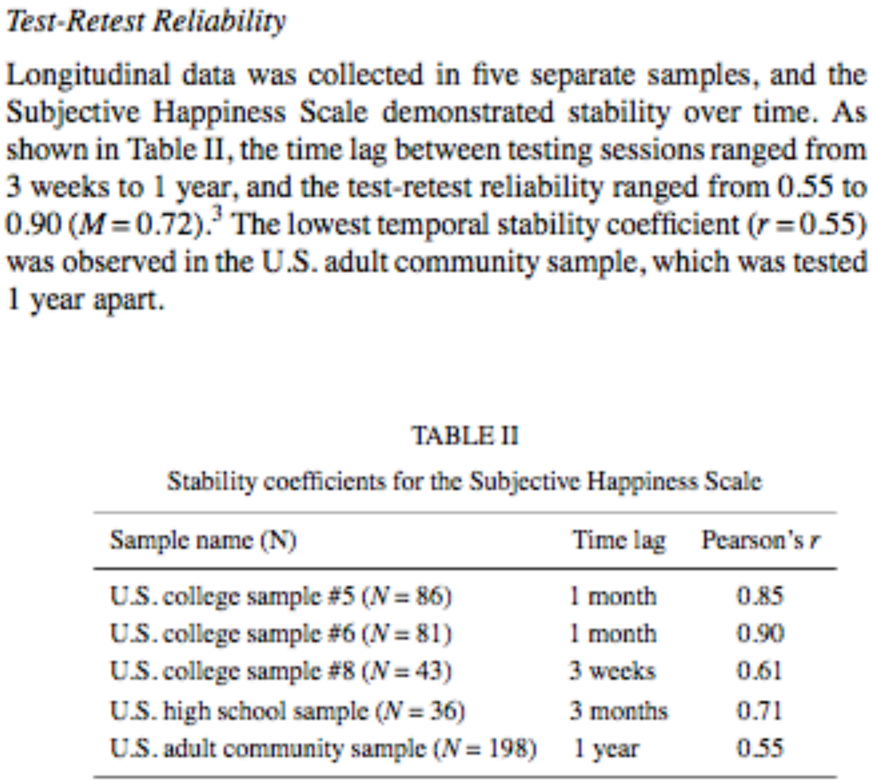
Reliability: Internal Consistency



Internal Consistency

The internal consistency among the four items comprising the Subjective Happiness Scale was tested using Cronbach's alpha reliability. In all samples, the four items showed good to excellent internal consistency, demonstrating comparability across samples of varying ages, occupations, languages, and cultures. The alpha's ranged from 0.79 to 0.94 (M = 0.86). Only one of the 14 coefficients fell below the conventional minimum of 0.80 (0.79 was

Reliability (cont’ d)



Test-Retest Reliability

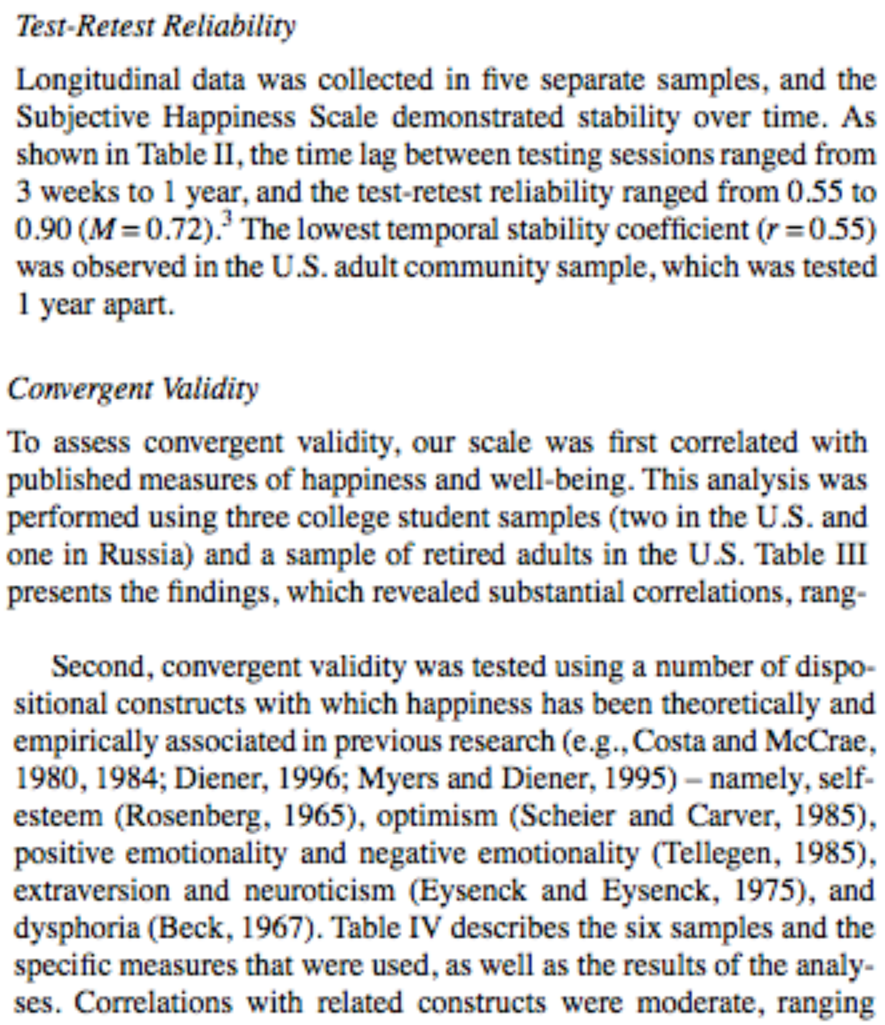
Longitudinal data was collected in five separate samples, and the Subjective Happiness Scale demonstrated stability over time. As shown in Table II, the time lag between testing sessions ranged from 3 weeks to 1 year, and the test-retest reliability ranged from 0.55 to 0.90 ($M = 0.72$).[3] The lowest temporal stability coefficient ($r = 0.55$) was observed in the U.S. adult community sample, which was tested 1 year apart.

TABLE II

Stability coefficients for the Subjective Happiness Scale

Sample name (N)	Time lag	Pearson's r
U.S. college sample #5 ($N = 86$)	1 month	0.85
U.S. college sample #6 ($N = 81$)	1 month	0.90
U.S. college sample #8 ($N = 43$)	3 weeks	0.61
U.S. high school sample ($N = 36$)	3 months	0.71
U.S. adult community sample ($N = 198$)	1 year	0.55

Reliability (cont’d) and Validity



Test-Retest Reliability

Longitudinal data was collected in five separate samples, and the Subjective Happiness Scale demonstrated stability over time. As shown in Table II, the time lag between testing sessions ranged from 3 weeks to 1 year, and the test-retest reliability ranged from 0.55 to 0.90 ($M = 0.72$).[3] The lowest temporal stability coefficient ($r = 0.55$) was observed in the U.S. adult community sample, which was tested 1 year apart.

Convergent Validity

To assess convergent validity, our scale was first correlated with published measures of happiness and well-being. This analysis was performed using three college student samples (two in the U.S. and one in Russia) and a sample of retired adults in the U.S. Table III presents the findings, which revealed substantial correlations, rang-

Second, convergent validity was tested using a number of dispositional constructs with which happiness has been theoretically and empirically associated in previous research (e.g., Costa and McCrae, 1980, 1984; Diener, 1996; Myers and Diener, 1995) – namely, self-esteem (Rosenberg, 1965), optimism (Scheier and Carver, 1985), positive emotionality and negative emotionality (Tellegen, 1985), extraversion and neuroticism (Eysenck and Eysenck, 1975), and dysphoria (Beck, 1967). Table IV describes the six samples and the specific measures that were used, as well as the results of the analyses. Correlations with related constructs were moderate, ranging