Agenda

- Context for measurement – what (all) do we use tests for?
- Historical perspective
  - Names to know
  - How measurement has changed over time
    - Why we are where we are today

Purposes of Testing

- What do we use tests for?
  - "We" meaning society, not just us personally.
- Describe individual people.
- Describe relationships among variables.
- Make predictions about future behavior.
- Make decisions about individuals.
  - Selecting, classifying.
- Quantify differences.
  - Before and after, etc…

Why does it matter?

- What happens if we don't measure well?
  - What are the consequences?
  - For who?
- Is measurement hard?
  - Justifications for poor measurement…

Hard & Soft Measurement

- DeVellis quotes Duncan (1984):
  - "All measurement… is social measurement."
  - What does he mean? Is this true?
- Even "hard" physical measurements – length, temperature, mass – are socially constructed.
  - Shared standards – consensus.
  - Someone had to propose a way to measure even these obvious physical things.
- Is psychological measurement different?

In The Beginning

- Pretty much as far back as we have records of people, we have records of people measuring people.
- In the Bible: Judges 7, the first personnel selection test!
- Ancient China: testing for civil servants, ~2200 BC.
- Socrates: questioning, *testing* knowledge.
- And lots of concern with the quality of physical measurement.

Measurement & Error

- 1660s – Isaac Newton and others begin to use averages rather than single observations in their calculations.
  - "Measure twice, cut once"
  - Acknowledging that a single observation had the potential to be wrong.
  - Principle of *aggregation* is one we would do well to remember today (cf. Epstein, 1979 in personality research).
- This acknowledgement of measurement error in the physical sciences led to the development of statistics.

Psychological Measurement

- Late 1800s:
  - Substantial developments in mathematics.
  - Interest in studying human mental processes.
- These interests came together in Sir Francis Galton.
  - Focused on intelligence / ability – how multiple kinds of fundamental abilities were related (and inherited).
  - Karl Pearson developed the correlation coefficient to support Galton's work.
  - Notion of "regression to the mean" also comes from Galton's work.

Psychophysical Measurement

- Wilhelm Wundt
  - Operationalizing and measuring fundamental human abilities.
- S. S. Stevens
  - Levels of measurement: nominal, ordinal, interval, & ratio.
  - Argued that (some) psychological variables *could* be measured on a ratio scale (e.g., volume).

Measurement & Math

- Charles Spearman (1900s)
  - Built on Pearson's statistical work – developed the *common factor model.*
  - Different measures may be tapping the same fundamental ability (intelligence).
  - Tests are correlated to the degree they measure something in common.
- L.L. Thurstone (1930s)
  - Built on Spearman's work, but from a different perspective.
  - Tests might measure not just one, but *multiple* factors.

Measuring Intelligence

- Alfred Binet & Victor Henri
  - Intelligence in the context of school – identifying French schoolchildren with developmental delays.
  - Can you pass items that most other people your age pass?
    - "Mental age" -> IQ
  - First test to include detailed instructions for standard adminstration.
- Adopted & elaborated by Louis Terman @ Stanford:
  - The Stanford-Binet Tests of Intelligence

Measuring Intelligence

- David Wechsler (1930s-1940s)
  - Stanford-Binet tests not really appropriate for assessing *adult* intelligence.

"To ask the average adult to say as many words as he can think of in three minutes, or to make a sentence of the words *to asked paper my teacher correct I my*, and assume that he will be either interested or impressed, is expecting too much."

- (Wechsler, 1944, p.17)

  - Using measurement techniques that were not appropriate for the population led to error!
  - Viewed intelligence as multidimensional.

Practical Problems

- World War I: US Army wanted to use intelligence tests to screen and place new recruits.
- But existing intelligence tests were individually administered ($!) and long.
- Solution: Army Alpha & Beta Tests
  - First multiple-choice intelligence tests, suitable for group administration & objective scoring.
- Needed to measure personality efficiently too:
  - Woodworth's Personal Data Sheet
  - Strong Vocational Interest Inventory

Developing Factor Analysis

- Guttman (1940s -50s): factor analysis doesn't replace theory.
  - FA can only be generalized to more tests/items "of the same kind" – you can't get out what you don't put in.
  - You need to have some idea about what you are measuring before you begin.
- Lawley (1940s – 50s) – developed *confirmatory* factor analysis.
  - Not computationally practical until the 1970s - Jöreskog.

Item Response Theory

- Lazarsfeld (1950s): factor models don't really work properly for binary items.
  - *Phi-gamma law* was a better model.
- Frederick Lord (1950s-1980s) – developed Lazarsfeld's ideas into Item Response Theory.
  - Factor models yield approximate answers to measurement questions – how much error is there, what is a person's true score, etc.
  - Often this level of precision is just fine… but sometimes we need more.
  - IRT = more sophisticated, more precise ways to answer the same questions.

Civil Rights & Fairness

- 1964 – Civil Rights Act
  - Prohibits discrimination in **workplace** and **educational** decisions.
  - These decisions tend to be based on tests, so…
  - Raised concerns about whether tests measured all people equally well.
- First concerns were item "bias" and test "bias."
  - Developed into concern about culture and measurement equivalence.

A Shocking Oversimplification of the History of Measurement

	What	How
1890s	basic human abilities	correlating psychophysical tests
1900s	complex human abilities	standardized intelligence tests
1910s	making testing practical	multiple choice
1920s	structure of ability	factor analysis
1930s	personality & interests	
1940s	structure of personality	
1950s		item response models
1960s	group differences & bias	differential validity
1970s		
1980s	culture and equivalence	measurement equivalence
1990s		DIF; polytomous IRT; latent class models
2000s		

Common Threads

- Developments in measurement are sparked by both theoretical and practical concerns.
  - "How can we be confident that the inference we want to make from these item responses is correct?"
- New developments often start in the ability domain and then are generalized to other content areas.
- Application of developments in measurement theory often waits on computer technology.
  - LISREL, BILOG… now R and Mplus. ☺

Questions?

Lab Friday: How to Read Math

For next time:
Constructs
Read: R & M Sections 1.1 – 1.7
1st Reading Response on Canvas