What every science educator should know about psychometrics

Yoav Bergner

HHMI education group meeting 2/28/2013
Psychometrics
measurement of psychological (psychosocial) phenomena

informed by: statistics | psychology | psychophysics
cognitive science | computer science

includes:
educational measurement
math ability
reading ability

personality testing
intelligence testing

Testing is a big part of the story

Scale development requires anticipating the evidence

For the science educator,
psychometrics is not the answer, per se
it may provide insight in framing the question
“Advanced technologies and statistical methods aren’t sufficient. One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them.”

Mislevy, Steinberg and Almond (channeling Messick)
maybe this is you

<table>
<thead>
<tr>
<th>Weighting scheme</th>
<th>Task</th>
<th>Code</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Exams</td>
<td>E</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>Final Exam</td>
<td>FE</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>Problem Sets</td>
<td>PS</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Reading Questions</td>
<td>RQ</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>Concept Questions</td>
<td>CQ</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>In Class Work: Friday</td>
<td>IC</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Problem Solving and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiments</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://web.mit.edu/8.01t/www/coursedocs/overview/grades.htm
Albert Aganov, Kazan University's physics department head, did not believe the professor was intoxicated, according to RIA Novost, saying, “I would have fired her immediately, if I had seen her drunk.” He also told Russian media that the test's length was “not unusual.”

http://www.huffingtonpost.com/2012/07/06/landysh-zaripova-russian-_n_1654529.html
either way, or somewhere in between, this is definitely you

(exremely perceptive, highly evolved)
Face to face, at least, you know many different ways to evaluate a peer or student.

   e.g. Bloom’s 2-sigma effect for expert tutors

So why don’t we all just do the best we can?

Because we worry about *fairness and quality*

   in fact, this is a very old concern...
A false balance is an abomination to the LORD, but a just weight is his delight.

Proverbs
Penny Tour

Measurement & Constructs (or Latent Variables)

Reliability & Validity (psychometric quality)

Graphs (aka paths)

Item Response Theory (with some examples)
Cronbach: a construct is some postulated attribute of people, assumed to be reflected in test performance...[T]he attribute about which we make statements in interpreting a test is a construct.

depression
extraversion
masculinity
scholastic aptitude
chemistry achievement
critical thinking
(new constructs are born all the time)

subject is adept at filling in bubbles
Percy Bridgman, Operation(al)ism (Logic of Modern Physics, 1927)
“the concept is synonymous with a corresponding set of operations”

“the space of astronomy is not a physical space of meter sticks, but is a space of light waves”

Samuel Messick (1995)
“In construct validation the test score is not equated with the construct it attempts to tap, nor is it considered to define the construct, as in strict operationism (Cronbach & Meehl, 1955). Rather, the measure is viewed as just one of an extensible set of indicators of the construct. Convergent empirical relationships reflecting communality among such indicators are taken to imply the operation of the construct to the degree that discriminant evidence discounts the intrusion of alternative constructs as plausible rival hypotheses.
On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

For seven years a committee of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940), the committee chose a classification of scales of measurement.

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties...
Stanley Smith Stevens (Science, 1946):
“[Paraphrasing N. R. Campbell] we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects and events according to rules.”

deprecated by Otis Duncan (1984) as incomplete:
“playing the piano is striking the keys of the instrument according to some pattern”

SS promotes theory of scale types
nominal (classification)
ordinal (e.g. IQ)
interval (e.g. time)
ratio (e.g. mass)

science lives here, but few (if any) educational measurements do
Reliability
(i. quantitative; ii. necessary but not sufficient for validity)

inter-rater reliability
e.g. Cohen’s kappa
how much better than chance
can also use for prediction models

test-retest reliability
different forms reliability

internal consistency, e.g. split-half
e.g. Cronbach’s alpha is a number in [0,1], values closer to 1 are better and > 0.7 is a reasonable criterion

alpha is not a homogeneity or unidimensionality parameter (e.g. it gets larger as test length is increased; moreover, it is possible to demonstrate using a heterogeneous test of m dimensions that alpha is not sensitive to m directly)
Validity
or, what is it all about?

validation is about interpretation or meaning of scores,
it is not a measure of tests in and of themselves

criterion validity
  concurrent validity
  predictive validity
construct validity
content validity

alternately one unified concept with:
content, substantive, structural, generalizability,
external, and consequential aspects

Eric Mazur [FCI]: “How should I answer these questions—according to what you taught me, or how I usually think about these things?”

Cronbach & Meehl, 1955
Messick, 1995
Construct invalidity can come from:

- construct underrepresentation (too narrow)
- construct-irrelevant variance (too broad)
  - construct-irrelevant difficulty (e.g. reading comprehension)
  - construct-irrelevant easiness (e.g. alternative solution)

Criterion validity

Differential Item Functioning (group bias)

an example, not exhaustive
Graphs are a common and useful shorthand for representing probabilistic models with conditional independence used to encode causal structure. NB: because they are shorthand, there is sometimes ambiguity.

Latent/hidden

<table>
<thead>
<tr>
<th>Observed</th>
<th>Continuous</th>
<th>Discrete</th>
</tr>
</thead>
</table>

e.g. skill

e.g. score

Correlated, but conditionally independent
Factor Analysis

$X_1 \cdots X_k$

$Y_1 \cdots Y_m$
Test model

it's possible for these paths to represent the same probabilities or independent probabilities (more on this later)
Dynamic Bayesian Network (e.g. Hidden Markov Model)

for modeling a changing state, e.g. learning
Classical Test Theory vs Item Response Theory

true score

\[ X = T + E \]

reliability

\[ \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \]

\[ P(x^{(1)}|\theta, \xi) \quad Q = 1 - P \]

\[ L(\theta, x_1^{(0)}, x_2^{(1)}) \propto P(x_1^{(0)}, x_2^{(1)}|\theta, \xi_1, \xi_2) \]

\[ = P(x_1^{(0)}|\theta, \xi_1)P(x_2^{(1)}|\theta, \xi_2) \]

\[ = Q(x_1^{(1)}|\theta, \xi_1)P(x_2^{(1)}|\theta, \xi_2) \]

\[ \hat{\theta} = \arg\max_\theta L(\theta) \quad \text{skill (ability)} \]

\[ I(\theta) = \sum P_i(\theta)Q_i(\theta) \quad \text{test information} \]

student-test vs. student-item

boils down to this difference in interaction granularity
Classical Test Theory
Item Response Theory

\[ \theta \]

- \( b_1 \)
- \( b_2 \)
- \( b_3 \)
- \( b_4 \)
- \( b_5 \)

Q1, Q2, Q3, Q4, Q5
Item Response Theory

*coming to a discipline-based education research journal near you!*

e.g. physics education

C. N. Cardamone et al., in *PERC Proceedings* (2011)

Perhaps IRT appeals to scientists’ notion of a best instrument for the job when items are essentially hierarchical (cf. Guttman scale), but that’s not the only option.
IRT was designed as an improved solution to testing problems

The goal is an ability score for the examinee

independently of which questions are selected from an item pool

(useful for high-stakes tests and also CAT)

“to describe the items by item parameters and the examinees

by examinee parameters in such a way that we can predict

probabilistically the response of any examinee to any item, even

if similar examinees have never taken similar items before.”

- F. Lord

More accurate, and with fewer items, than raw scores.

Measures items as well as students, and on the same scale.

IRT analysis reveals both “faulty” and highly discriminating items

But: there are many IRT models/methods, and details are fussy.
Some typical Item Response Curves

- Q25
  - $d = -1.44$
  - $a = 0.676$
  - $X^2 = 1.03$

- Q26
  - $d = -0.561$
  - $a = 1.93$
  - $X^2 = 0.664$

Mechanics Baseline Test (MIT)
Some typical Item Response Curves

Mechanics Baseline Test (MIT)
And some pathological ones

ambiguous questions, not model fail!
misreading + common misconception → correct response

(Cardamone et al. PERC 2011)
2-dimensional 2PL

Item
Characteristic
Surfaces

source:
Reckase, McKinley (1991)
What exactly is meant by dimensionality of skill?

*Dimensionality = number of parameters used to model a student’s skill (over a domain)*

Use of one final grade *implies* “unidimensionality”

A *multidimensional* basis can mean many things:
- topics (energy conservation or rigid body motion, etc.)
- conceptual vs. procedural knowledge
- language/reading ability vs. math proficiency
- problem types (graphing, algebraic numeric, analytical, etc.)
- something else...

We use a technique called *collaborative filtering* to look for the “best” number of parameters to use in terms of predicting the correct/incorrect response data in a held-out (cross-validation) set.

(We are not specifying any basis.)
MBT at MIT
2005-2009
Pre & Post
N=4700

dimensionality?
it depends
Parameter space projection of MBT items using (3140) model

Q1 & Q2 cluster

Q4 & Q22 may be outliers

Q20 is the only work-energy question
Step 1. Determine clearly what it is you want to measure
Step 2: Generate an item pool
Step 3: Determine the format for measurement
Step 4: Have the initial item pool reviewed by experts
Step 5: Consider inclusion of validated items
Step 6: Administer items to a development sample
Step 7: Evaluate the items (item performance, factor analysis, alpha)
Step 8: Optimize scale length
“Advanced technologies and statistical methods aren’t sufficient. One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them.”

Mislevy, Steinberg and Almond (channeling Messick)

Thank you!