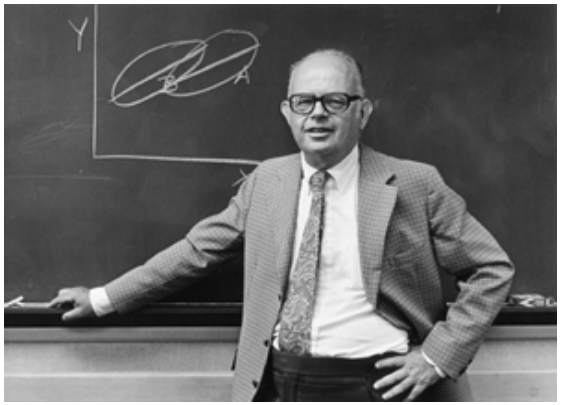
STATSU webinar: Psychometrics for survey construction

RME—Research, Measurement, and Evaluation

June 30, 2020

Welcome! Psychometrics for survey construction



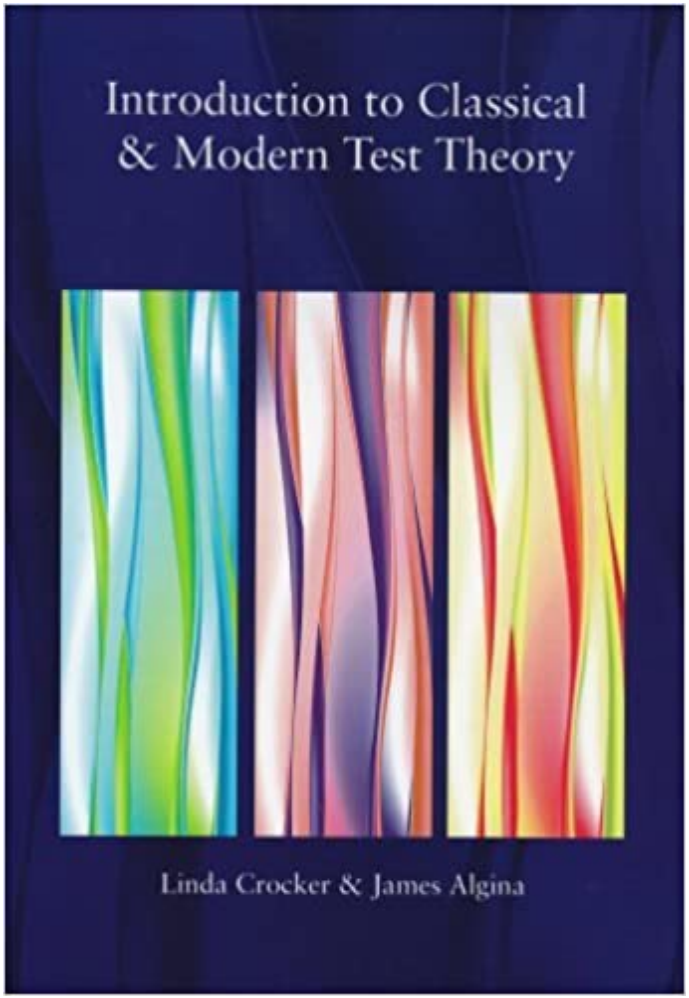
"It is an embarrassment to me that the formula (for coefficient alpha) became conventionally known as Cronbach's α."

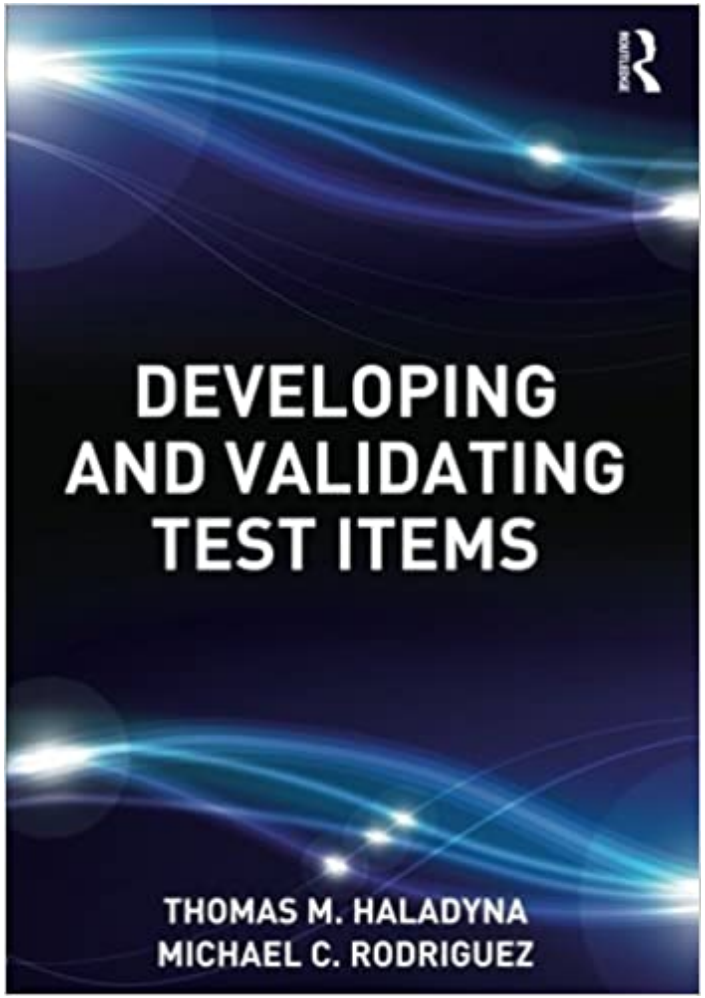
-Cronbach (2004)

Cronbach & Shavelson (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement, 64(3) 391-418

Outline

Resources


Introduction to Classical
& Modern Test Theory
Linda Crocker & James Algina


DEVELOPING
AND VALIDATING
TEST ITEMS
THOMAS M. HALADYNA
MICHAEL C. RODRIGUEZ

Introduction

Purpose: measure weight for person.

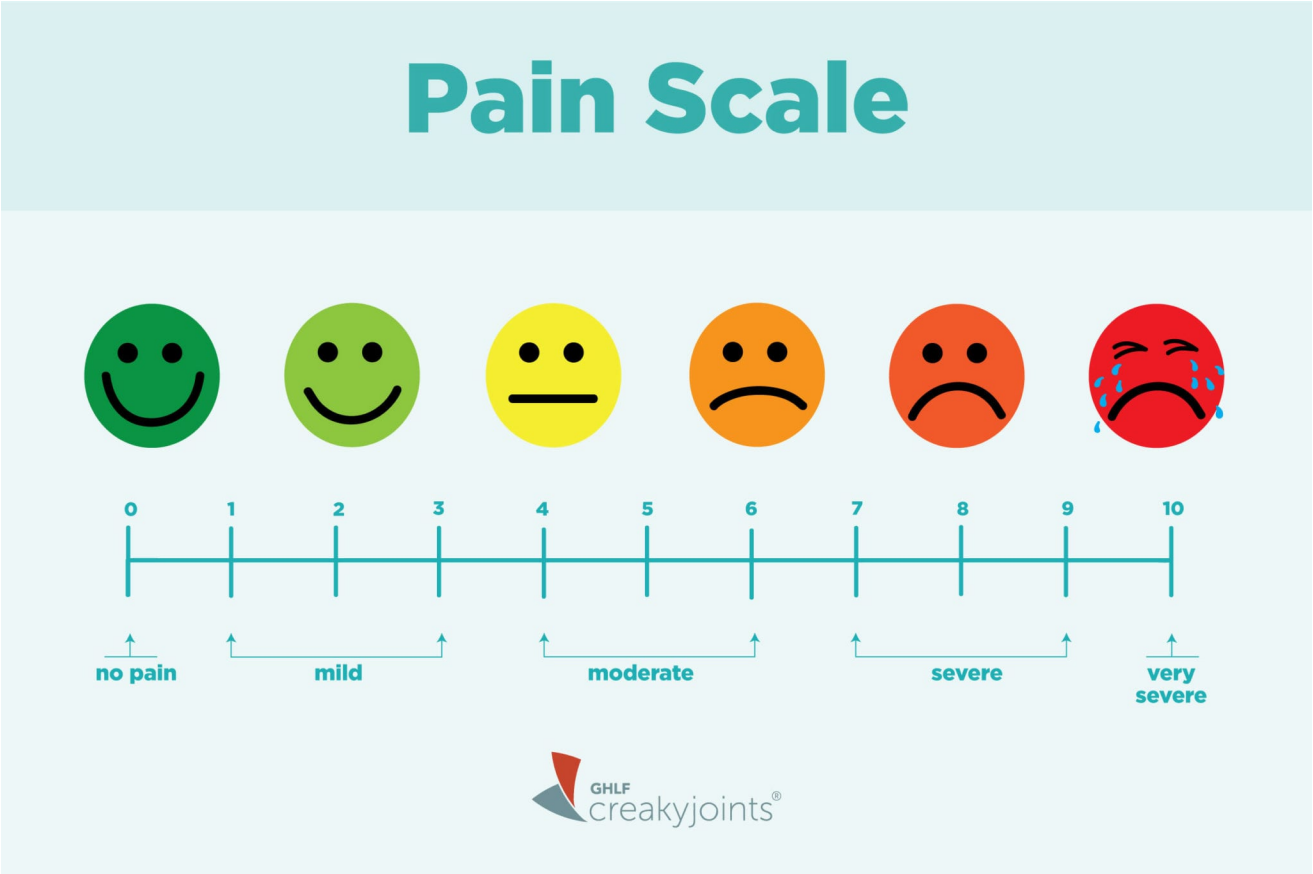
Variable: body mass measured in weight (lbs).

Method: digital scale.

Latent variables need to be defined for quantifying on a scale that covers the full continuum of its range.

Introduction

Define construct of interest (Operational Definiton), measurement model (T/F, MC, Likert/rating, constructed response), and scoring (answer key/ essay rubric); Mindful of cultural differences.



Levels of measurement (Stevens, 1946)

1. Nominal: categorical data; numerical values only indicate labels (1 = orange, 0 = green)
2. Ordinal: Number assigned indicates a rank. Rank difference is unequal (winner of race is assigned 1 for first place; does not indicate gap between second or subsequent ranks)
3. Interval: numerical values indicate both ranking and distances between values reflect the differences on construct—proportionate intervals. It is an arbitrary scale where a zero does not indicate an absence of construct.
4. Ratio (sometimes grouped as Interval/Ratio): rank-orded, units are proportionate intervals, and an absolute zero point—there is a total absence in quanitity.

Other theories of measurement beside Stevens. For example, Coombs (1950) talks about a level between nominal and ordinal, plus a level between ordinal and interval.

Summary: Levels of measurement (Stevens, 1946)

	Nominal	Ordinal	Interval	Ratio
Calculation: Mode	✓	✓	✓	✓
Ranked values		✓	✓	✓
Calculation: Median		✓	✓	✓
Calculation: Mean			✓	✓
Evaluate numerical differences			✓	✓
Zero = total absence of variable?				✓

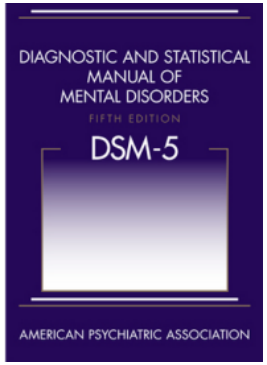
Types of tests

1. Ability: Aptitude or proficiency (often grouped with achievement tests).
2. Achievement: Content mastery after instructional period.
3. Personality: Stable ways people think, act, and feel; traits or types.
4. Preference: Attitudes or Interest.

Power v. Speeded tests: Unspeeded if at least 80% of the test takers complete all items and all examinees complete at least 75% of the items (ETS, 1974).
....what about respondents that rush to finish?

Inferences

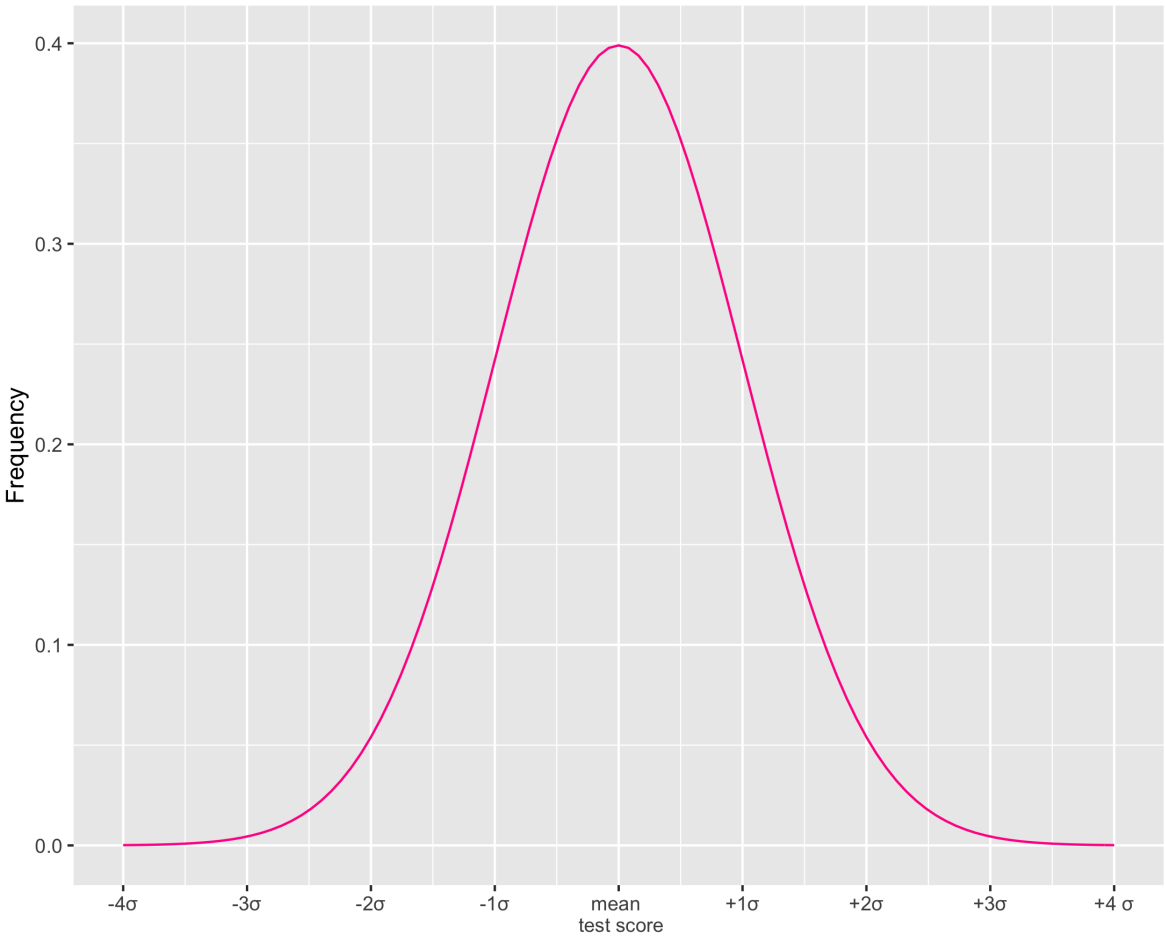
1. Rating
   norm-referenced: relative to examinees
   criterion-referenced: relative to benchmark
2. Selection
3. Placement
4. Diagnosis
5. Outcome evaluation



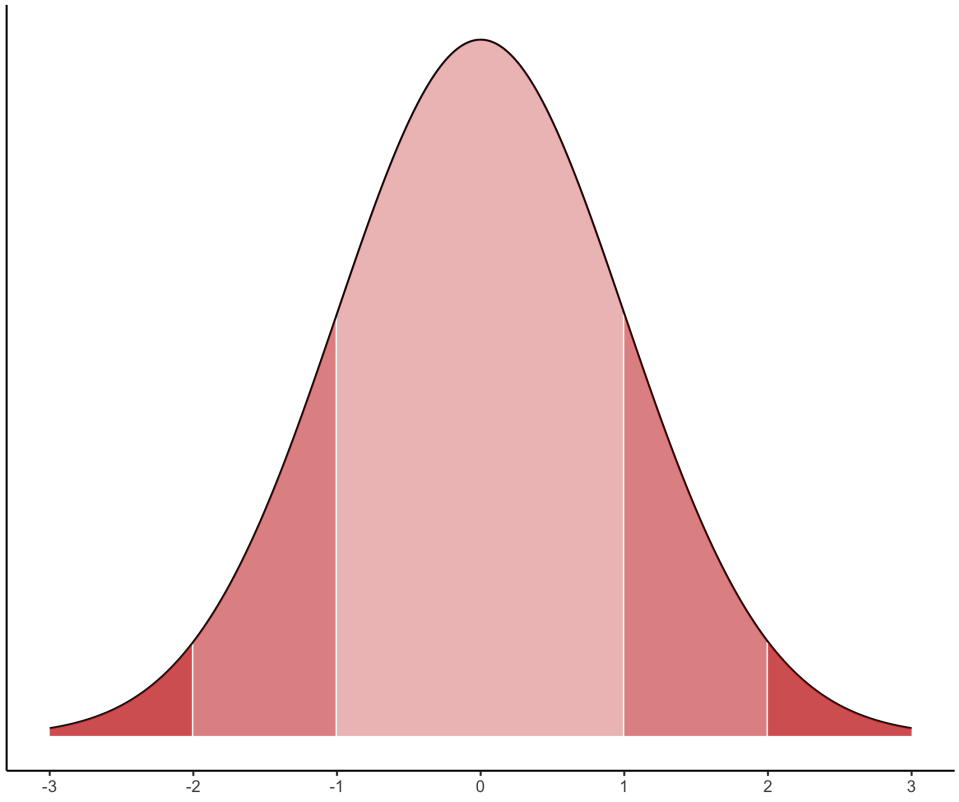
Person responsible for selecting test must be able to evaluate a test regarding: norms, reliability, validity, and test administration variables that impact scores.

Standarized scores: z-score

$$z = \frac{X - \bar{X}}{SD}$$



Standarized scores: empirical rule 68-95-99.7



Approximately:

$$4\sigma < 0.14\% < 3\sigma$$
$$3\sigma < 2.14\% < 2\sigma$$
$$2\sigma < 13.59\% < 1\sigma$$
$$1\sigma < 34.13\% \leq 0$$

Standardized score

Negative scores aren't ideal to report.

Report on another scale by generically transforming z-score:

$$Score = (new\ SD)(\text{z-score}) + (new\ mean)$$

A t-score (M = 50, SD = 10) is one possible standardization.

Reliable/Valid test introduction

A test must be valid and reliable to make inferences about respondents.... "estimation precision + accuracy."

Reliable: Reproducible and consistent. A pre-requiste for validity.

Valid: Does the score measure what you indended? When measuring depression, are you actually measuring someone's self-esteem or current mood?

4 different scales, weights consecutively measured on each:

Validity

1. Construct validity: Are the items assessing the construct of interest?
   Factor Analysis to determine traits or factors from intercorrelations among tests or items on test.
   Multitrait-Multimethod Matrix to assess convergent/discriminant validation.
2. Content validation: extrapolating the items on test to similar items; "item bank." Table of test specification.
3. Criterion-realted validity: What extent do the items measure the topic of interest? Criterion contamination correlation as evidence.
   Concurrent or predictive validity are types of CRV.

Verifying constructs: Crocker & Algina, 2008

1. *Content analysis*. With this method, open-ended questions are posed to subjects about the construct of interest, and their responses are sorted into topical categories. Those topics that occur predominantly are taken as major components of the construct. For example, Jersild (1952) published results of a content analysis of compositions by children describing themselves, and the resulting categories served as the basis for generating items for two widely used inventories designed to measure children's self-concepts (Gordon, 1967; Piers and Harris, 1964).
2. *Review of research*. Those behaviors that have been most frequently studied by others are used to define the construct of interest. The test developer may use an eclectic approach or select the work of one particular theorist in specifying behavioral categories to be represented by test items.
3. *Critical incidents*. A list of behaviors is identified that characterizes extremes of the performance continuum for the construct of interest. This method is usually attributed to Flanagan (1954), who asked job supervisors to describe situations in which an employee had functioned with outstanding effectiveness or ineffectiveness and thereby generated a list of "critical behaviors" to use for rating job performance.
4. *Direct observations*. The test developer identifies the behaviors by direct observation. For example, a vocational counselor, developing an inventory to assess job-related stress in high-risk occupations, might find that actual observations of such workers on the job would help identify situations that are potential sources of emotional stress.
5. *Expert judgment*. The test developer obtains input from one or more individuals who have first-hand experience with the construct. Written questionnaires or personal interviews are used to collect information. As an illustration, a personnel psychologist who wants to develop a checklist for rating performance of staff nurses in a large hospital can survey a group of nursing supervisors to identify the types of performance that should be included.
6. *Instruction objectives*. Experts in a subject are asked to review instructional materials and develop a set of instructional objectives when an achievement test is being developed. An instructional objective specifies an observable behavior that students should be able to exhibit after completion of a course of instruction. Such objectives communicate to the item writer both the specific content on which items should focus and the nature of the tasks the examinees should be able to perform. A thorough description of the process of developing objectives is provided in most introductory texts in measurement (e.g., Brown, 1983; Mehrens and Lehmann, 1984; Popham, 1981).

Reliability

$$\text{Score} = \text{True Score} + \text{Measurement Error}$$
$$X_i = T_i + E_i$$

1. Congeneric items: Each item measures a single construct, but the items have varying weights in calculating the composite score. Depression screening example: suicidal thoughts v. waking up several hours sooner than normal.
2. Tau-equivalent items: Each item is equally weighted for composite score. Depression screening example: waking up several hours sooner than normal and interrupted sleep.
3. Parallel items: All items account for the same variance in score. Essentially a respondent interprets this as the same question asked more than once.
   Elicited response intention $\neq$ question interpretation

Link: Classical Test Theory and the Measurement of Reliability

Reliability: Parallel forms

1. cover same number of items
2. same content
3. same level of difficulty

$$S_x^2 = S_t^2 + S_e^2 \qquad\qquad r_{xx} = \frac{S_t^2}{S_x^2}$$

variance of observed scores = S_x^2
variance of true scores = S_t^2
variance of error scores = S_e^2
r_{xx} is the proportion of observed/total variance that is S_t^2
1 - r_{xx} = proportion of error variance
True variance = $(r_{xx})S_x^2$
Error variance = $(1 - r_{xx})S_x^2$

Reliability: split-half

1. Divide test in half and treat as parallel forms
2. Measure of internal consistency
3. Spearman Brown formula used to get full length of test

If we assume that each item is an alternate form of every other item then we calculate coefficient-α. If the items are scored dichotmously then this special case of coefficient-α is Kuder-Richardson Formula 20 (KR-20).

Coefficient α: alpha....incorrectly "Cronbach's alpha"

Assumptions: Scale is unidimensional, continuous, normally distributed, and items are tau-equivalent. Measurement error for items don't covary.

$$coefficient\ \alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{i=1}^{J} s_j^2}{s_X^2}\right)$$

For i items on a scale with J items, s_j^2 is the variance for an item, and s_X^2 is the variance for all items on the scale.

Possible interpretations are the average of all potential split-half correlations, or a lower bound estimate of test variance due to item communalities for tau-equivalent items.

Reliability: Sources of error

Reliability	Error Variance
Test-retest	time sampling
Alternate forms	content (and time if delayed testing)
Split-half	content sampling
α/KR20	content sampling and heterogeneity
Rater	Interrater differences

Coefficient ω: omega total

Reliability of test measured by ratio of variance explained by items over total scale variance. If tau-equivalence assumption is satisfied, coefficent-α will be estimated as an equivalent.

$$coefficient\ \omega = \frac{\left(\sum_{i=1}^{J} \lambda_i\right)^2}{\left(\sum_{i=1}^{J} \lambda_i\right)^2 + \sum_{i=1}^{J} \theta_{ii} + 2\sum_{i=2}^{J}\sum_{k=1}^{i} \theta_{ik}}$$

λ is the coefficent or "factor loading" for item i on a scale with J number of questions. θ_{ii} is the error variance for item i, and k is defined as an item on the scale so a covariance between items i and k can be calculated.

Coefficient ω_R, ω_H, and H

Revelle's ω, hierarchical ω, and maximal reliability (Hancock & Mueller, 2001; Bentler, 2007)

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S. Lee (Ed.), Handbook of computing and statistics with applications: Vol. 1. Handbook of latent variable and related models.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), Structural equation modeling: Present and future—A festschrift in honor of Karl Jöreskog.

Overview of test construction

1. Behavioral Objectives
2. Table of Specifications
3. Item Writing
4. Item analysis

Behavioral Objectives: Bloom's Taxonomies

Ranked from simple to complex:

1. Knowledge
2. Comprehension
3. Application
4. Analysis
5. Synthesis
6. Evaluation

Taxonomy: Knowledge

Knowledge: Memorization and recall of facts.

Instructional verb examples:
define, identify, cite, label, list, match, select.

Example question:
"List all of Bloom's taxonomies ranked from simple to complex."

Taxonomy: Comprehension

Comprehension: Understand content then interpret.

Instructional verb examples:
Explain, extend, interpret, summarize, give examples.

Example question:
"Correctly explain Freud's stages of development."

Taxonomy: Application

Application: Application of knowledge to novel tasks.

Instructional verb examples:
Demonstrate, compute, relate, solve, use, produce.

Example question:
"Given algebra equations not previously discussed in class lectures, correctly compute the solution."

Taxonomy: Analysis

Analysis: Break down new information on the topic then differentiate between subcomponents

Instructional verb examples:
Estimate, distinguish, analyze, determine.

Example question:
"Given two conflicting publications, distinguish the pros and cons of each article."

Taxonomy: Synthesis

Synthesis: Develop a pattern from pieces of information.

Instructional verb examples:
Develop, design, compose, propose, write.

Example question:
"Design an experiment for a research question."

Taxonomy: Evaluation

Evaluation: Inspect others ideas then gauge the effort and the value.

Instructional verb examples:
Evaluate, judge, assess, compare, contrast, justify.

Example question:
"Judge a peer-reviewed article's value according to current literature and research standards."

Table of Specifications

Topic	Comprehension 30%	Synthesis 30%	Evaluation 40%	Points
10% Content A	3	3	4	10
30% Content B	9	9	12	30
20% Content C	6	6	8	20
40% Content D	12	12	16	40
Total points	30	30	40	100

Topic percentages should reflect importance or time dedicated on instruction in class.

Points per cell = Row percent $\times$ column percent $\times$ total points

$[row = 3, column = 1] = (0.2)(0.3)(100) = 6$ points

MC: Stem & options

Stem: Select the objective for Piaget's first stage of developtment:

Options:
A. Abstract concepts
B. Operational thought
C. Object Permanence

Guidelines

Test questions should include: verb for cognitive domain, and success criterion if not explicitly defined (e.g., in correct order, with 80% percent accuracy, results reported formatted APA style...).

Score should reflect explicit question, not implicit expectations or directions outside of the test (i.e., a syllabus). Do not deduct points for implicit understandings or stylistic preference.

Example Question: "Conduct a *t*-test on the provided dataset and report results in APA format"

AVOID: Deducting points for respondents not reporting results for testing statistical assumptions of a *t*-test, or providing background information paragraph on the dataset, or rounding to 3 decimal places instead of 4 decimals. Peer-reviewed journals don't challenge coefficient-α assumptions....

Multiple Choice Items

1. State stem in clear, simple language
2. Avoid negatations
3. Intended answer is correct and clearly best
4. Alternative answers are grammatically consistent
5. Vary position of correct answers
6. Avoid "all of the above" and "none of the above" or "other"
7. 3 MC response options sufficient
8. Options should be independent
9. No-nonsense/humor: all options plausible

Introduction: EFA

Exploratory factor analysis (EFA): Determine number of latent variables that explains covariance among observed, manifest, variables.

Continuous latent variabales also referred to as factors and the maifest variables are referred to as indicators (can be non-continuous level of measurement).

EFA: informs about factors for tentative items on underlying scale, and which items have "loadings" on the scale.

The main goals are to discover the number factors on a scale and salience of factor loadings (small loadings and large error variances candidates for removal).

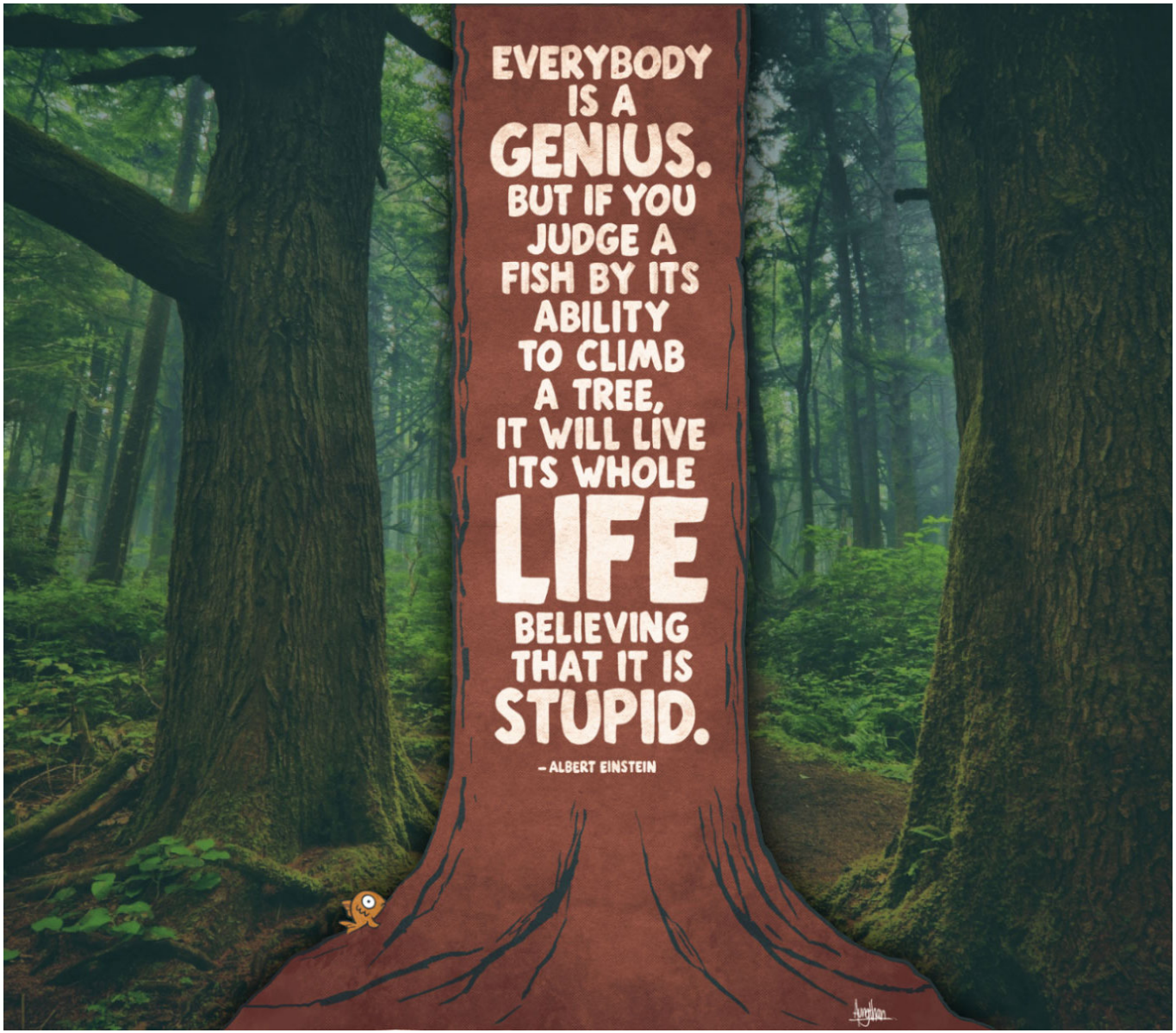
Fit Statistics

Joint criteria is recommended
e.g., CFI $\geq$ .96 & SRMR $\leq$ .10

1. Loglikelihood: compares nested models (deviance)
   H_0 Value: null model $p < .05$
2. Information Criteria: non-nested models
   AIC, BIC, SBIC: *smallest value*
3. RMSEA and SRMR
   value $< .05$
4. CFI/TLI: (comparative fit index & Tucker-Lewis Index[1])
   value $> .95$
   literature can support values above .90 as acceptable
5. χ^2 test for baseline
   Baseline is a null model $p < .05$

[1] NNFI is another name for TLI

Thank you!


EVERYBODY
IS A
GENIUS.
BUT IF YOU
JUDGE A
FISH BY ITS
ABILITY
TO CLIMB
A TREE,
IT WILL LIVE
ITS WHOLE
LIFE
BELIEVING
THAT IT IS
STUPID.
– ALBERT EINSTEIN