Comps Study Outline

# PRE 822/921: CTT 1&2

Validity

* Constructs
* Argument-based approach – evidence accumulation
* Establish purpose/intended use, then collect appropriate evidence
* Content
  + Define performance domain
  + Select qualified experts
  + Provide structure frameworks (test specs) and have experts match items to the performance domain
  + Experts evaluate whether the test’s content meets specifications
* Criterion
  + Identify criterion behavior and way to measure it
  + Select appropriate sample of representative examinees
  + Administer test and evaluate relationship between test scores and criterion scores
    - Convergent: predictive and/or concurrent validity evidence from related measures
    - Divergent/discriminant: predictive and/or concurrent validity evidence that is not expected to correlate
  + Multi-trait multi-method matrix can be assembled to express convergent/divergent evidence
* Construct
  + Factor analysis typically way of assembling construct validity evidence – support factor structure
  + Encompasses all types of validity evidence according to Messick
* Relevance/utility, value implications, consequential (debates whether to include/consider)
* Messick’s progressive matrix and its iterations
* Unitary concept
* Validity is about scores, not tests
* Evidence-Centered Design (ECD)
  + Claims are supported by observable evidence
  + Assessment opportunities are based on observable evidence, and their characteristics are task models
  + Task models can generate assessment tasks and items
* Threats to Validity
  + Construct underrepresentation
  + Construct-irrelevant variance

Steps in Scale Development

1. Define construct and its grain size/number of factors
2. Generate the item pool (via ECD)
3. Determine measurement format (scoring & scaling via CTT/IRT/FA/other)
   1. CTT: unweighted sum of item scores
   2. FA: empirically weighted sum of item scores
   3. IRT: latent variable inferred by pattern of item responses
   4. Thurstone scaling: item responses weighted by judges; precursor to FA and IRT, in which weights are empirically determined
   5. Others: Likert scaling, visual analog scale
4. Review item pool to ensure item relevance and check for threats to validity
5. Consider inclusion of validation items (e.g., social desirability scale)
6. Administer items to pilot sample resembling population
7. Evaluate items – descriptive statistics, reliability
8. Produce final scale

Sampling

* Must adequately represent the population – generalizability
* Random sampling needed to control for known and unknown population effects
* Steps for sampling:
  1. Define general universe (e.g., U.S. teachers)
  2. Identify working universe (e.g., members of NEA)
  3. Choose sampling unit (e.g., one teacher)
  4. Find sampling frame (e.g., list of NEA members’ e-mails)
* Sampling strategies: random, systematic, stratified random, cluster, judgment, convenience
* After sampling, random assignment needed to control for confounds
* Larger samples 🡪 consistency w/ true population, smaller sampling error
* Standard error of proportion used to evaluate sampling error:
* Central Limit Theorem: sample means for N > 30 are normally distributed, so SE(P) can be used to create confidence intervals – “margin of error”
* To determine the sample size needed to achieve a given margin of error: , where is the desired standard error of the proportion and the margin of error = 1.96\*SE

Strategies for eliciting honest answers to threatening questions

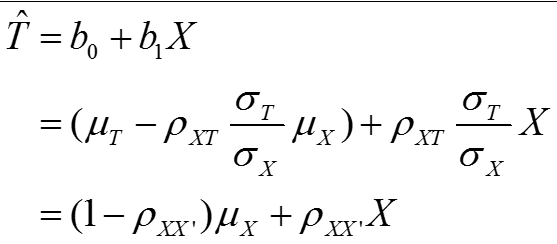
Classical Test Theory (CTT)

* CTT: observed score = true score + error
* True score = mean of observed scores from infinite administration of the test
* Error assumed to have true mean of 0 and variance distributed normally
* Reliability = reproducibility of scores across administrations
* CTT reliability depends on parallel forms being theoretically possible
* Assumptions
  + Mean of error () = 0
  + Correlation of true score & error () = 0
  + Correlation of errors () = 0
* Parallel forms: two or more tests measuring the same content where and
  + Tau (τ) equivalence: True scores are equal (), but error variances may be unequal. Errors still uncorrelated.
  + Essential tau equivalence: True scores differ by an additive constant ( ), and error variances may be unequal. Errors still uncorrelated.
  + Congenericity (or congeneric forms): True scores differ by a positive linear function ( ), where d > 0, and error variances may be unequal. Errors still uncorrelated. This is the assumption used in equating multiple forms.
* Strengths
  + Weak assumptions, wide applicability
  + Simple model which does not require model-data fit because definition makes it tautological
  + Useful test summary statistics: SEM, reliability, validity
  + Doesn’t require large samples
* Important Outcomes
  + SEM is function of observed scores
  + True scores are function of observed scores & reliability
  + Reliability requires idea of parallel forms
  + Spearman-Brown prophecy

Reliability

* Necessary but not sufficient condition for validity
* Reliability Index = correlation between true and observed scores ()
* Reliability coefficient = correlation between observed scores or parallel forms ()

Reliability coefficient is the ratio of true score variance to observed score variance

* Reliability coefficient = (reliability index)2   
  So as error decreases, the reliability coefficient increases.
* Ways to get reliability:
  + Over time: test-retest reliability
  + Between raters: interrater reliability
  + Across forms: parallel forms reliability
  + Single form: internal consistency reliability (the inter-correlations among test items)
* Coefficient (Cronbach’s) Alpha  
  This is actually the lower bound of reliability due to non-parallel items.
* Standard Error of Measurement – average standard deviation of observed scores around examinee true scores (NOT conditional on true score)
* Confidence Intervals assume normally distributed errors (Central Limit Theorem):
* Can use the standard error of measurement to build a confidence interval around any observed score
* is analogous to R2 in regression, & the standard error of measurement is analogous to the standard error of the estimate
* Spearman-Brown Prophecy provides a way to find the length of test needed to obtain the desired level of reliability
  + is the desired reliability C, reached by lengthening test X by a factor of k (assuming parallel components are added)
  + k is the factor of increase (or decrease) needed for a test with reliability to reach reliability
* Attenuation: estimating the correlation of true scores between two less-than-perfectly-reliable tests, X, and Y
  + Correlation can’t exceed either test’s reliability index
* Estimates of true scores through simple regression: Kelly Estimates
  + 
  + Kelly estimates reduce the random variability in test scores by regressing estimates of true scores toward the mean of the observed scores, but they don’t change the overall mean or the rank order
  + Sometimes used to adjust for unequal observed score means in regression, DIF, and equating studies
* Decision reliability: decision consistency around one or more cut scores
  + The extent to which classification decisions are the same from parallel forms or multiple administrations of one form.
  + Assessed using a contingency table, and decision consistency = 1- the proportion of consistent decisions
  + Cohen’s Kappa (κ) is used to adjust for decision consistency due to chance agreement



* + Affected by test reliability, cut score location across score distributions, similarity of score distributions
* Other reliability estimates (historical methods)
  + Split-half correlation: correlation between two halves of the test
    - Issue: many ways to split a test in half
    - People have used Spearman-Brown prophecy to adjust for shortness of two halves  
      Using the formula in this way assumes the two created halves are parallel; violation of this assumption means the estimate will be positively biased
  + Rulon’s method: the idea that the differences between scores from two halves should be small, and smaller differences = greater consistency
    - Calculate difference scores: D = A – B
    - Estimate total variance and variance of difference scores
    - Estimate of consistency:
  + Kuder-Richardson shortcut methods for estimating reliability when all items are dichotomous: KR-20 and KR-21
    - KR-20 is same as Cronbach’s alpha but replaces item variance with simpler Bernoulli expression:
    - KR-21 adds the assumption that all items are equally difficult
* Reporting reliability data
  + Multiple reliability studies are necessary to account for different sources of measurement error; internal consistency is necessary but does not indicate stability
  + Standard error of measurement and score bands for different CIs should accompany reliability estimates
  + Reliability and SE estimates should also be reported for subscores
  + Procedures and samples should be described sufficiently to facilitate comparisons with the reader’s situation
  + If a test is normally used for a population of examinees, reliability estimates and SEMs should be reported separately for other specialized populations
  + If SEMs are being estimated via a model other than CTT, this should be specified – otherwise, readers will assume the CTT version

Conditional Standard Errors of Measurement

* SEM calculation:
* The group-level SEM is the average of conditional SEMs; group-level reliability is the average of conditional reliabilities

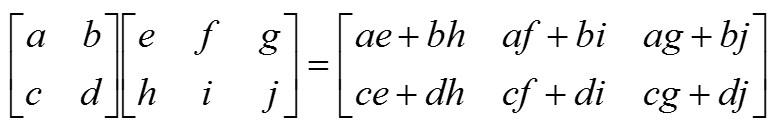
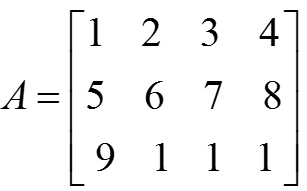
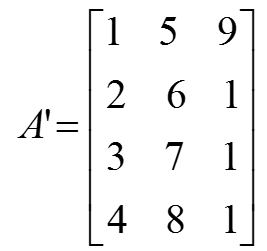
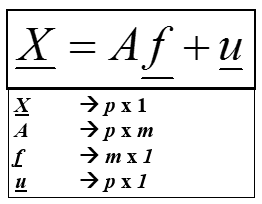
Item Analysis

* Item-level means & variances + covariance/correlation matrix determine test characteristics
* With dichotomous items, p-values are way of expressing difficulty (though larger p-value = easier item)
* Item variance
  + for any item:
    - Item variance is maximized when the item’s mean is at the midpoint of the possible score range
  + for a dichotomous item:
    - Item variance is maximized when p = 0.5
* Item correlations contribute to unidimensionality/higher reliability
  + Correlation for a pair of items *j* and *k*
    - Population:
    - Sample:
  + Correlation for a pair of items is maximized when means are equal
* Item discrimination = item is effective at separating examinees with lower and higher true scores
  + Point-biserial correlation = *rj* = correlation between the item *j* and the total test score *X* =
  + Items that positively covary are positively discriminating
* Test-level mean & variance are composites calculated from item statistics
  + Mean = sum of items means
  + Variance = sum of item variances + (2 ∙ the sum of the n(n-1)/2 covariances); in other words, the sum of the variance-covariance matrix
* To increase test variance, choose items w/ large individual variance and high intercorrelation (more variance = more discrimination)

Evaluating Dimensionality

* Principal Components Analysis: analyses correlation matrix to determine whether one or more components are present
* Exploratory Factor Analysis: specified # of dimensions, then all possible relationships estimated.
  + Also uses correlation matrix, but 1s on diagonals have been replaced by estimates of communality (sum of squared factor loadings across factors, which is theoretically analogous to variable’s reliability)
  + Two approaches: Principal Axis Factoring (PAF) & Maximum Likelihood (ML); PAF tries to extract max orthogonal variance w/ each factor, and ML estimates loadings that maximize the probability of sampling the observed correlation matrix from a population
* Confirmatory Factor Analysis: specified # of dimensions and expected relationship; other relationships constrained to be 0
* Factor loadings = correlation w/ factor; rule of thumb: loadings > |0.3| are meaningful
* Factor retention rules:
  + Eigenvalues (λ) > 1
  + Ratio of adjacent eigenvalues
  + Scree test
  + Bartlett’s χ2 test of sphericity (tests H0 that pop. correlation matrix is an identity matrix; should be non-significant, but has Type I error w/ large samples)
  + Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (determines items’ shared variance after accounting for partial correlations; should be as close to 1 as possible)
* Factor rotations:
  + Orthogonal = uncorrelated; varimax, quartimax, equamax; maximizes variability of loadings
  + Oblique = allowed to correlate; direct oblimin, promax; better fit, difficult interpretation
* Factor scores – can be estimated via regression: standardized coefficients are produced and multiplied by observed variables to produce subjects’ factor scores

Matrix algebra

* Addition: all corresponding elements are added
* Multiplication: # columns in one much match # rows of another  
  
* Transpose: reversing rows & columns  
     
  AA’ and A’A are always degined and are always square and symmetric
* Inverse
  + scalar: **a**-1 = 1/**a**
  + matrix: **AA**-1 = **A**-1**A** = **I**, where **I** is the identity matrix; inverse is only defined when **A** is a square matrix with a determinant != 0
* Use in regression:
  + Model:
  + Solve for *b*
* Eigenvalues (λ) & eigenvectors
  + Eigenvalues are a coefficient to the identity matrix that “captures” the variance in a correlation matrix
  + The corresponding eigenvector’s elements are coefficients of the best linear combination of observed variables that account for the greatest amount of common variance – in EFA & PCA, there is one eigenvalue and corresponding eigenvector for each factor/component
  + The vector of *p* observed variables (*X*) is decomposed into a *p* x *m* matrix of factor loadings (*A*) multiplied by a vector of *m* factors (*f*) plus a vector of *p* random errors (*u*)  
    
  + *X* is a linear combination of *f* and *u*, its V/C matrix (ΣXX) is a function of the V/C matrix for *f* (Φ), factor loadings (*A*), which are the “weights,” and the error V/C matrix (Θ).