MAP ACADEMY

Methodology, Analytics & Psychometrics

Designing Better Questionnaires and Measures: Psychometric Review

Leslie R. Hawley, Ph.D.

Michelle Howell Smith, Ph.D. & Ann Arthur, MS

Presented on 4/3/15

CYFS

University of Nebraska Lincoln

Nebraska Academy for Methodology, Analytics and Psychometrics

- Wide range of services for funded research projects
- Expertise in
  - Statistics & Modeling
  - Applied Psychometrics
  - Program Evaluation
  - Mixed Methods
  - Prevention Science

Three Part Series

Designing Better Questionnaires and Measures

1. Initial considerations and construct operationalization*
2. Constructing and Testing the Instrument*
3. Psychometric Review

*available at:

http://mapacademy.unl.edu/presentations/methodology-application-series/2014-2015/

Focus of the Series

- Evaluation of **non-cognitive** measures (questionnaires/surveys) for use in educational, psychological, and social science research
  - Non-cognitive measures
    - Attitudes, opinions, perceptions
- Concepts generalize to other applications
  - Cognitive measures
    - ACT, SAT, GRE

Presentation Overview

- Introduction
  - Definitions related to psychometric review
  - General framework for review process
- Reliability Evidence
- Validity Evidence
- Final Thoughts

DEFINITIONS

Definitions

- Measurement
  - Systematic process of assigning numbers as a way of representing a characteristic/property (Raykov & Marcoulides, 2011)

- How would you measure 5 feet of fabric?

- How would you measure self-efficacy?

Definitions

- Unlike the length of fabric, psychological characteristics cannot be measured directly using a ruler or some other tool

- Instead, researchers have to develop measures and questionnaires to indirectly measure latent constructs such as self-efficacy

- Constructs
  - Unobserved, latent characteristics given meaning through the combination of measurable attributes, skills, or traits
    - Ex: Depression, IQ, Conflict, Self-Efficacy, Motivation
  - Operalization of constructs is guided by theory

Definitions

- There is always a degree of error in our measures because latent constructs are not observed directly
  - Error may be due to aspects related to the participant, setting, and/or instrument
- Due to these potential sources of error, researchers need to evaluate the reliability and validity of the scores from measures used to evaluate latent constructs
- Reliability
  - "Consistency of a measurement procedure" (John & Benet-Martinez, 2000, p. 342)
    - Consistency is not enough → need to evaluate accuracy as well
- Validity
  - How well an instrument measures what it claims to measure

FRAMEWORK

Framework

- Anyone using or developing a measure has the burden of proof for demonstrating that ***scores*** from a measure demonstrate adequate quality
  - Evidence needs to support the intended inferences and uses (Kane, 2006; Messick, 1989)

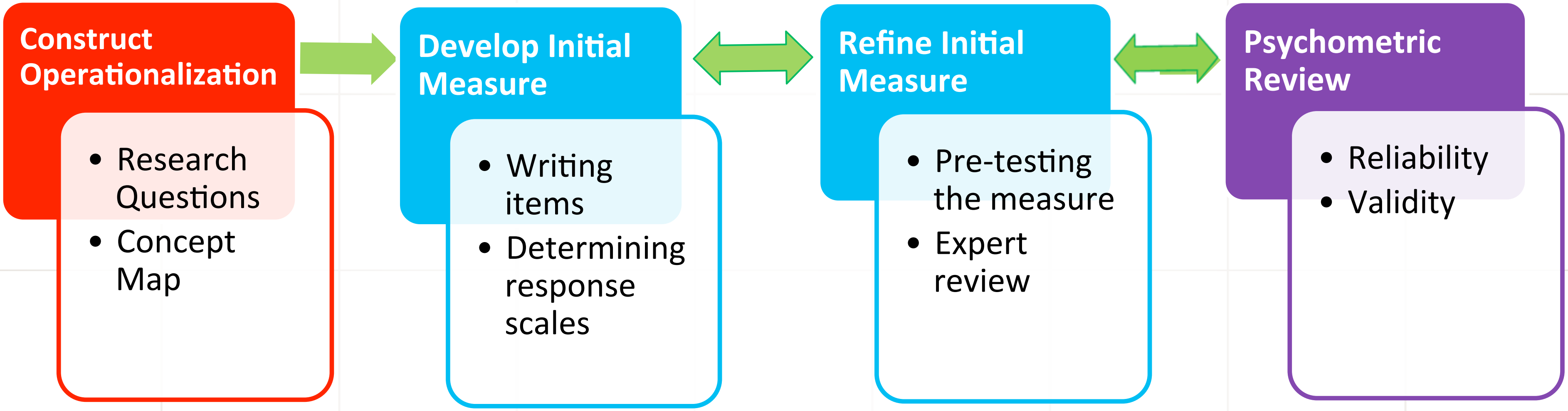
- A measure is never called “reliable” or “valid”
  - Interpretations and uses of **scores** and **intended inferences** are validated, not the measures themselves (Cronbach and Meehl, 1955; APA, AERA, & NCME, 2014; Kane, 2013)

Framework

- Reliability and validity are not absolutes
  - Reliability and validity are not referred to in terms of the presence or absence of reliability/validity, but rather as a matter of degrees (Messick, 1989)

- Evidence is sample and purpose specific (Messick, 1989; Sireci, 2009)
  - Psychometric information for a measure of teacher stress in grades 3-8 is specific to that population→ additional evidence would need to be collected if used on a population of high school teachers
  - Different interpretations/uses for scores require additional (and perhaps different kinds of) psychometric evidence

Framework



Construct Operationalization

- Research Questions
- Concept Map

Develop Initial Measure

- Writing items
- Determining response scales

Refine Initial Measure

- Pre-testing the measure
- Expert review

Psychometric Review

- Reliability
- Validity

- Actions surrounding the development, use or evaluation of a measure are all connected to validity
- Each step in the process provides a different source of evidence for the intended use(s) of scores

Framework

- Continual process
  - Accumulating validity evidence is neither static nor a one-time event, rather it is a continual process that uses multiple evidence sources (Shepard,1993; Messick, 1989; Kane, 2006)

RELIABILITY EVIDENCE

Reliability

- “Consistency of a measurement procedure” (John & Benet-Martinez, 2000, p.342)
- Degree to which scores remain consistent if the measure were given at a later time in similar conditions (Crocker & Algina, 1986)
- Indices of reliability describe the degree to which scores are reproducible

Reliability

Reliability

Reliability

- Individuals are not always consistent, so scores will have a small amount of measurement error and vary from one occasion to another

Reliability: CTT

- Scores will have a small amount of measurement error and vary from one occasion to another

- Classical Test Theory (CTT) assumes that there is a hypothetical average (true) score that is an error-free value resulting from several replications or alternate forms (APA, AERA, & NCME, 2014)

  - Thus, any individual score (X) is assumed to be a comprised of a True Score (T) and error (E)
    - X = T + E

  - CTT assumes after several replications or alternate forms, the average of the resulting errors approach zero

  - Across multiple replications/forms the expected value of X = True Score

Reliability: CTT

- Instead of conceptualizing reliability in terms of a single score, reliably is often conceptualized in terms of a sample of persons where:

  Var (X) = Var (T) + Var (E)

- Reliability = Var (T) / Var (X)
  - Proportion of variance due to “true scores” out of the total observed variance
  - If there is no error then reliability= 1; if there is only error then reliability = 0

Examples of traditional CTT reliability coefficients

- Test Re-test → stability over time
  - Intraclass correlation (ICC)
  - Pearson correlation
- Equivalence → stability over forms
  - ICC
  - Pearson correlation
- Internal Consistency → stability over judges/observers
  - ICC
  - Cohen’s Kappa
- Internal Consistency → stability over items
  - Coefficient alpha (more to follow)

Generalizability Theory

- In CTT, reliability is studied one aspect at a time (ignoring other sources of error) and we do evaluate the relative contribution of multiple sources of variance

- Remember in CTT, there is a single source of variance attributed to error "e" (i.e., X = T + e)

- Generalizability theory (GT) expands on the logic of Analysis of Variance (ANOVA) to disaggregate the multiple sources of variance that contribute to "e"
  - Measures multiple sources of variance in a single analysis
  - Researchers can deliberately test for specific sources that contribute to scores and estimate the degree of variance associated with each source (John & Benet-Martinez, 2000)

Generalizability Theory

- In GT, a person’s true or universe score is the mean of scores from different conditions or facets (Shavelson & Webb, 1991; Mushquash & O’Connor, 2006)

- G coefficient → ratio of universe score variance to observed score variance (Mushquash & O’Connor, 2006)
  - Variance component estimates which reflect the degree of observed variance due to a particular source or interactions between sources
  - Ex: 15% of variance is due to time; 25% of variance is due to the interaction between judge and item content

Generalizability Theory

- Researchers need to design specific development and evaluation plans for collecting information across multiple sources (facets)
  - Ex: Forms, items, occasions, and raters
  - Sources can be crossed (information on all facets) and/or nested (does not include information on all facets)

Software programs for GT

- No specific programs available in SAS or SPSS
  - Researchers have developed syntax programs available for use in SPSS, SAS, and MATLAB (Mushquash & O'Connor, 2006)
- GENOVA-suite programs (Brennan, 2001)

COEFFICIENT ALPHA

Coefficient Alpha

- Alpha ranges from 0 -1
  - higher values indicate greater internal consistency*

    *pending assumptions (more to follow)

- Coefficient alpha tends to be the default coefficient for evaluating internal consistency reliability in the social and behavioral sciences
  - Yet, there are several limitations to alpha

Coefficient Alpha

- Alpha is determined by:
  - Interrelatedness of items
  - Length of the measure

Coefficient Alpha: Example

	Measure A: 10 Items								
Item	1	2	3	4	5	6	7	8	9
1	--								
2	0.3	--							
3	0.3	0.3	--						
4	0.3	0.3	0.3	--					
5	0.3	0.3	0.3	0.3	--				
6	0.3	0.3	0.3	0.3	0.3	--			
7	0.3	0.3	0.3	0.3	0.3	0.3	--		
8	0.3	0.3	0.3	0.3	0.3	0.3	0.3	--	
9	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	--
10	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3

- What is your best guess for alpha?

(John & Benet-Martinez, 2000, p. 344)

Coefficient Alpha: Example

	Measure B: 6 Items				
Item	1	2	3	4	5
1	--				
2	0.6	--			
3	0.6	0.6	--		
4	0.3	0.3	0.3	--	
5	0.3	0.3	0.3	0.6	--
6	0.3	0.3	0.3	0.6	0.6

- What is your best guess for alpha?

(John & Benet-Martinez, 2000, p. 344)

Coefficient Alpha: Example

- Both measures have the same alpha (.81) but there are noticeable differences between the two measures

- Interrelatedness of items
  - Measure A has 10 items with a mean $r = .33$
  - Measure B has 6 items with a mean $r = .42$

- Length of the measure
  - Length can compensate for lower levels of inter-item correlation
    - As long as items do not decrease the mean interitem correlation, reliability always increases as number as items increases
    - Utility of adding items diminishes quickly (i.e., less increase in alpha for 10th item as opposed to the 4th)

(Example from John & Benet-Martinez, 2000)

Coefficient Alpha: Example

- A high alpha does not indicate you have a homogeneous and/or unidimensional measure

- Measure A: Completely homogeneous (all *rs* = .3; SD = 0)
- Measure B: Non-homogenous (*rs* = .3 & .6; SD = .15)
  - Potential multidimensionality in Measure B
    - Further investigation needs to be done via confirmatory factor analysis (CFA)
      - Allows for the ability to empirically test whether or not your measure is unidimensional
    - Alpha should not be used when a measure is multidimensional because it will underestimate reliability

(Example from John & Benet-Martinez, 2000)

Coefficient Alpha

- Alpha of .80 is not a benchmark for all conditions
  - High alpha can mask item redundancy or narrowness of content that can lead to:
    - Less efficient tests
      - 25 items were used when 5 would have sufficed
      - Redundant items increase alpha but do not add unique information
    - Less content coverage for certain areas
      - Redundant items that emphasize one aspect of the construct more than another may increase alpha at the expense of decreasing validity
      - Depending on the goal of the researcher, narrow content representation leads to less useful measures

(John & Benet-Martinez, 2000)

Coefficient Alpha

- Alpha also has strong assumptions (Cho, & Kim, 2015)
  - Tau equivalence
    - Items have equal discriminating power (equal factor loadings)
  - Error terms are uncorrelated (independent)
    - When this is violated alpha overestimates reliability
- Modern model-based reliability approaches offer alternatives to alpha with less strict assumptions
- Coefficient Theta (Teo & Fan, 2013)
  - Does not assume unidimensionality
  - Uses the number of items and largest eigenvalue from a principal components analysis

Coefficient Alpha

- Coefficient Omega (Teo & Fan, 2013)
  - Latent variable model based method that uses parameter estimates of the items
    - Evaluates the ratio of the variance due to the factor (construct of interest) to the total variance
    - Does not assume items have equal discrimination (tau-equivalence) or uncorrelated errors

- Although alpha is the default coefficient, there are alternative options that may paint a more accurate picture of reliability

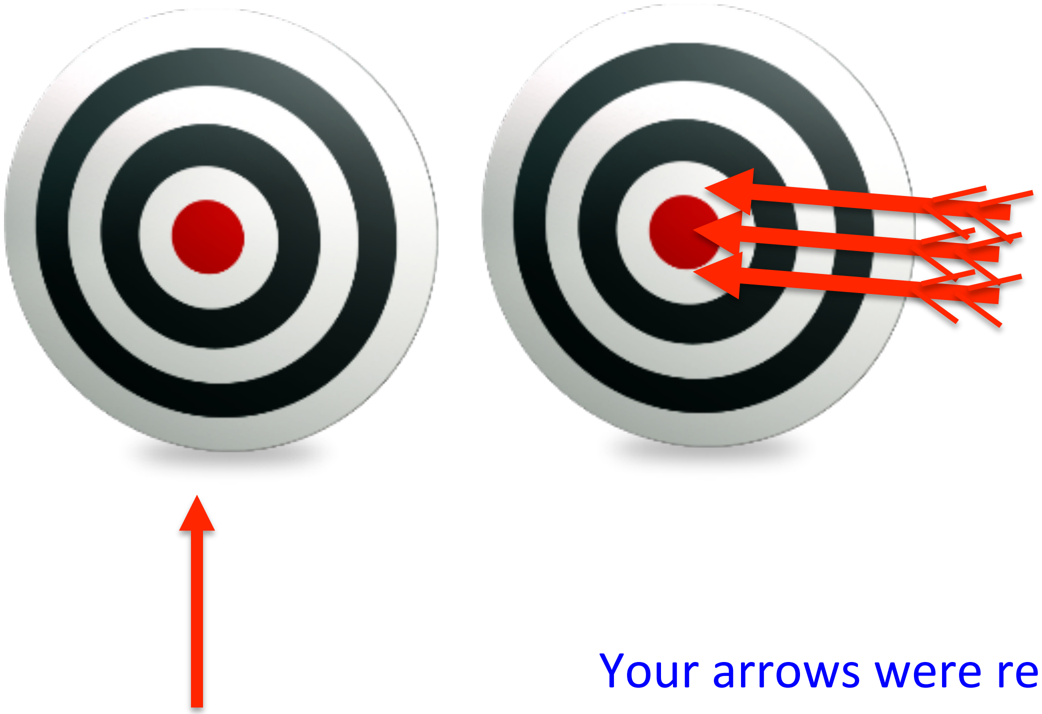
RELIABLITY ≠ VALIDITY

Reliability

- Reliability is a **necessary**, *but not sufficient*, condition for validity
    - Scores that demonstrate reliability are not necessarily valid
        - You could be measuring something the same way every time (consistent) but you could be measuring something other than what you intended

Reliability

Reliability



Actual target

Your arrows were reliable but not valid because you did not hit the intended target

VALIDITY EVIDENCE

Validity Evidence

- *Standards for Educational and Psychological Testing* (*Standards*) defines validity as:
  - "a unitary concept"
  - "the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA, APA, & NCME, 2014, p. 14)

- The *Standards* considers validity:
  - "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 2014, p. 11)

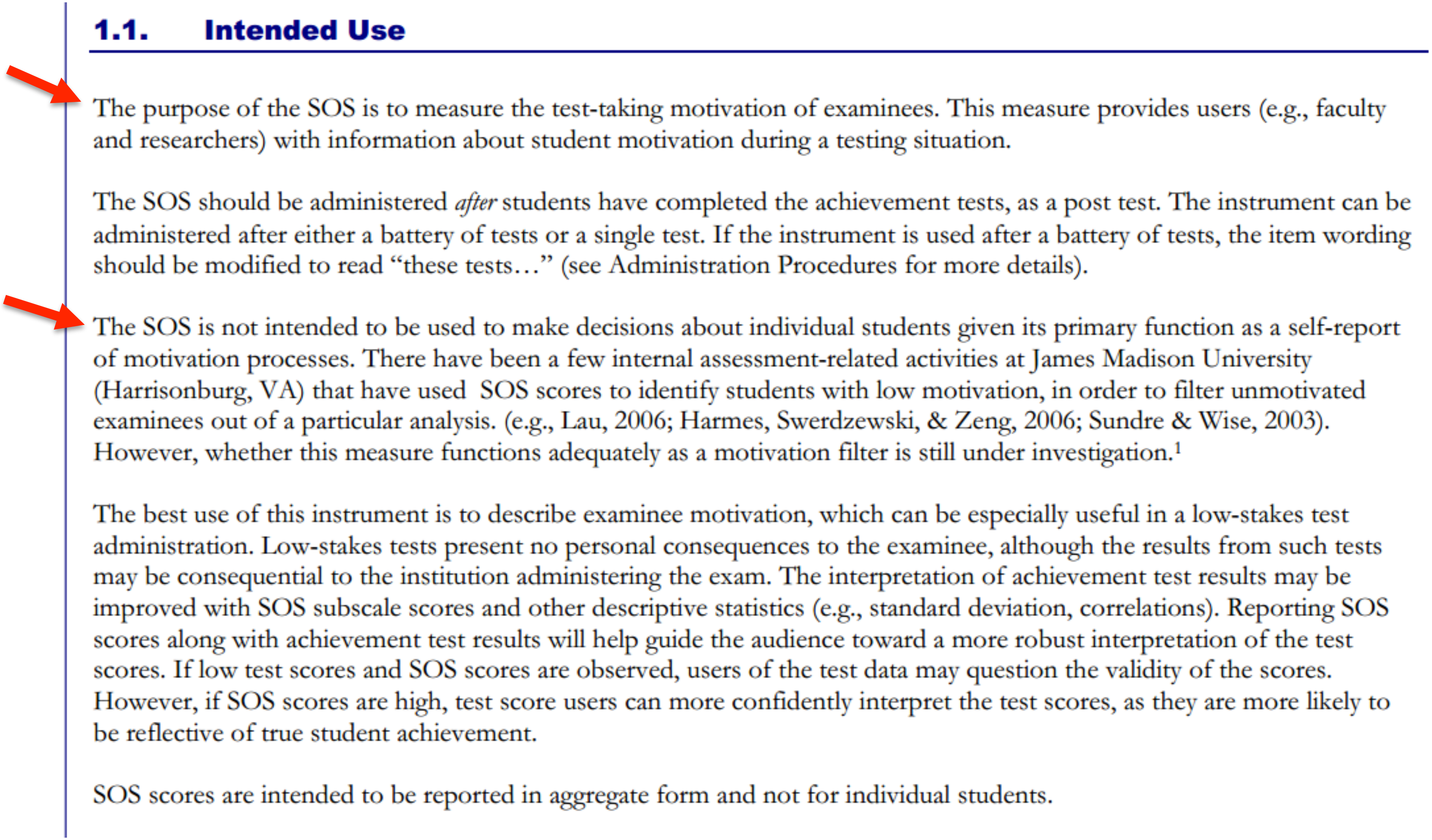
Validity Evidence

- Unitary concept → Construct validity evidence
- What seems like different types of validity are different sources of evidence related to the overarching concept of construct validity
  - "all validity is of one kind, namely, construct validity" (Messick, 1998, p.37)
- Construct validity evidence
  - Umbrella approach that subsumes all validation processes
  - Includes, but is not limited to:
    - Reliability evidence
    - Statistical conclusion validity evidence
    - Content evidence
    - Convergent and discriminant evidence
    - Evaluation of group differences

Validity Evidence

- Constructs
  - Unobserved, latent characteristics given meaning through the combination of measurable attributes, skills, or traits
    - Ex: Depression, IQ, Conflict, Self-Efficacy, Motivation
  - Operalization of constructs is guided by theory and previous research
    - Need to specifically define your construct of interest
      - Determine what it is and what scores are intended to measure
      - Determine what it is NOT and what scores are NOT intended to measure

Example Manual



1.1. Intended Use

The purpose of the SOS is to measure the test-taking motivation of examinees. This measure provides users (e.g., faculty and researchers) with information about student motivation during a testing situation.

The SOS should be administered *after* students have completed the achievement tests, as a post test. The instrument can be administered after either a battery of tests or a single test. If the instrument is used after a battery of tests, the item wording should be modified to read "these tests…" (see Administration Procedures for more details).

The SOS is not intended to be used to make decisions about individual students given its primary function as a self-report of motivation processes. There have been a few internal assessment-related activities at James Madison University (Harrisonburg, VA) that have used SOS scores to identify students with low motivation, in order to filter unmotivated examinees out of a particular analysis. (e.g., Lau, 2006; Harmes, Swerdzewski, & Zeng, 2006; Sundre & Wise, 2003). However, whether this measure functions adequately as a motivation filter is still under investigation.[1]

The best use of this instrument is to describe examinee motivation, which can be especially useful in a low-stakes test administration. Low-stakes tests present no personal consequences to the examinee, although the results from such tests may be consequential to the institution administering the exam. The interpretation of achievement test results may be improved with SOS subscale scores and other descriptive statistics (e.g., standard deviation, correlations). Reporting SOS scores along with achievement test results will help guide the audience toward a more robust interpretation of the test scores. If low test scores and SOS scores are observed, users of the test data may question the validity of the scores. However, if SOS scores are high, test score users can more confidently interpret the test scores, as they are more likely to be reflective of true student achievement.

SOS scores are intended to be reported in aggregate form and not for individual students.

Reference: http://www.jmu.edu/assessment/resources/resource_files/sos_manual.pdf

Validity Evidence

- Validation requires a clear argument for the proposed interpretations and uses of scores (Kane, 2006)
  - Interpretive argument → inferences from the observed data to any claims we hypothesize
    - Outlines reasoning and provides specific claims that need to be evaluated
    - Framework for evaluation
  - Validity argument → evaluation of the interpretive argument

- "Validity is an inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use" (Messick, 1989, p.5)

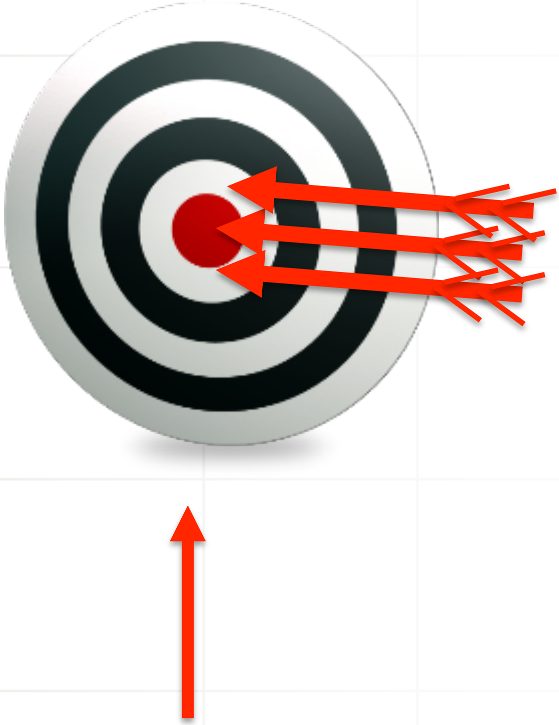
Validity Evidence

- Evidence is based on a particular use and interpretation
  - Specific to how we define our construct
  - Determines how we can interpret scores from our measure

- Evidence should be multifaceted
  - Variety of sources and methods
  - Need to provide “a convincing, comprehensive validity argument” (Sireci, 2009, p.33)
  - "multiple lines of evidence .. consonant with the inference, while establishing that alternative inferences are less well supported" (Messick, 1989, p.5)

Validity Evidence

Validity Evidence



Actual target & multiple sources of evidence

Your arrows were reliable and valid because you consistently hit your intended target

SOURCES OF VALIDITY EVIDENCE

Example Measure

- For the subsequent slides I will reference a hypothetical measure called:
  - "Teacher Accountability Stress Index (TASI)"

- 10 items administered to teachers in grades 3-8
- 1-5 Likert-type scale
- Purpose:
  - Measure teacher stress as it pertains to accountability
- Potential use of the measure:
  - Provide administrative intervention for highly stressed teachers

- Example items from the hypothetical measure:
  - "Accountability testing has led to pressure to increase student test scores"
  - "I worry about my job security if students underperform on their accountability tests"

Validity Evidence

- Multiple sources for accumulating validity evidence
  - Three areas were chosen for discussion in today's presentation: Content, Criterion-related, and Construct

<u>Content evidence</u>

- Potential questions of interest:
  - How well does the measure reflect the intended construct, knowledge, skills?
    - Relevance
      - A depression measure should ask questions about feelings related to sadness
    - Representativeness
      - Comprehensive
      - Ideally there are multiple items for a particular construct
  - How were items developed?
  - Were items evaluated prior to administration?
  - Were multiple groups (e.g., women, minorities) represented in the development process?

Validity Evidence

Examples of relevant content evidence for "TASI"

- How well does the measure reflect the intended construct, knowledge, skills?
  - Multiple items were developed to measure the construct
  - Items addressed accountability testing and potential areas of stress
- How were items developed?
  - TASI items were developed based on pilot qualitative research with teachers and extensive literature review of existing measures
- Were items evaluated prior to administration?
  - Prior to administration TASI items were reviewed for language and content by a small group of teachers from grades 3-8
    - What if I could only have university professors review items?
- Were multiple groups (e.g., women, minorities) represented in the development process?
  - Demographics for the review teachers included:
    - 60% Female; 40% Male
    - 65% White, 25% African-American, 5% Asian, & 5% Other

Validity Evidence

Criterion-related evidence

(Evidence based on relations to other variables)

- Potential questions of interest:
  - How well do scores from a measure relate to a particular criterion
    - How well do scores on a new measure of teacher stress (TASI) relate to a more established measure of teacher stress, the "Teacher Stress Index (TSI)"?
  - What exactly is the measure valid for?
    - Scores from the new measure of teacher stress may predict scores on an established measure of teacher stress but not a potentially unrelated construct such as Classroom Organization (CO) from the Classroom Assessment and Scoring System (CLASS)

Validity Evidence

Example of criterion-related evidence for "TASI"

- How well do scores from a measure relate to a particular criterion?

  An established measure of teacher stress is the "Teacher Stress Index (TSI)"

  - Teachers were administered both our new measure (TASI) and the established criterion (TSI)
  - Pearson correlation was used to evaluate the relationship between the two measures
    - Significant strong correlation (*r* = .75) between the TASI & TSI

Validity Evidence

Cautions for criterion-related evidence

- Restriction of range
  - Relationship between TASI (Spring 2015) and later Job Satisfaction (Spring 2016)
    - Teachers who are very stressed may leave the profession
  - Reduces the strength of the relationship you would find with the entire group

- Attenuation
  - Low reliability of one variable or both may reduce correlation

- Overall, researchers need to be thoughtful about choosing a criterion
  - Sometimes it is difficult to identify and measure an objective criterion
  - No one criterion can account for all aspects you may be trying to measure

Validity Evidence

Construct evidence

- Potential questions of interest:
  - How well does my hypothesized structure fit the data?
    - Do items thought to define the construct load onto the same single factor?
  - Do measures of the same construct (teacher stress) correlate more highly than measures of another construct (classroom organization)?
  - Does my hypothesized structure demonstrate differences across subgroups?
  - Is my hypothesized structure stable over time?

Validity Evidence

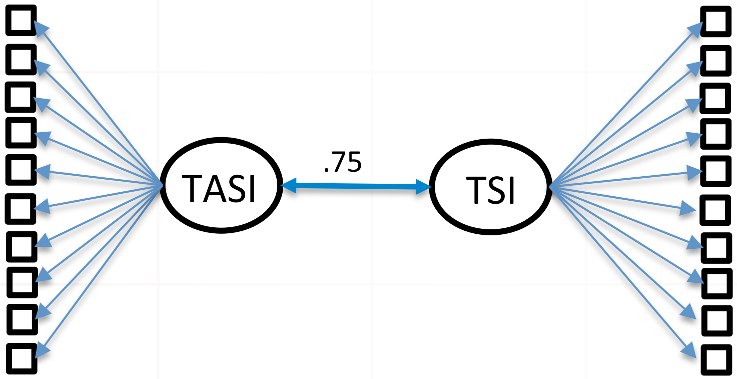
Examples of construct evidence for "TASI"

- How well does my hypothesized factor structure fit the data?
  - Exploratory Factor analysis (EFA)
    - Since the TASI is a brand new measure, our first step is to conduct an EFA to examine the factor structure
    - Assuming we have 200 teachers who filled out the TASI, we would conduct an EFA with 100 randomly chosen teachers
      - EFA provides both factor structure and item information
  - Confirmatory Factory Analysis (CFA)
    - Once we have established the TASI has a particular factor structure, we would perform a CFA with the 100 teachers not included in our original sample to confirm this structure

Validity Evidence

Examples of construct evidence for "TASI"

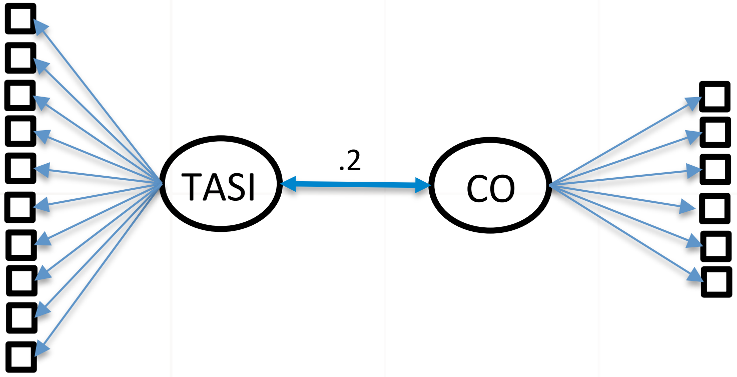
- Do measures of the same construct (teacher stress) correlate more highly than measures of another construct (classroom organization)?
  - Convergent and Discriminant evidence
    - Evidence for what a measure does and does not assess
  - Latent variable approach [CFA; Structural Equation Modeling (SEM)]
    - Convergent Evidence
      - Ex: Positive correlation between latent factors of TASI & TSI, similar measures of teacher stress



Validity Evidence

Examples of construct evidence for "TASI"

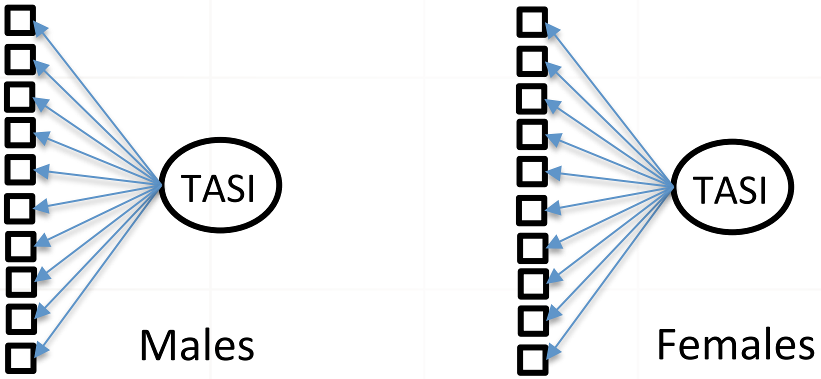
- Do measures of the same construct (teacher stress) correlate more highly than measures of another construct (classroom organization)?
  - – Discriminant Evidence
    - Ex: Low correlations between TASI and classroom organization (CO), variables that measure different (or less related) constructs



Validity Evidence

Examples of construct evidence for "TASI"

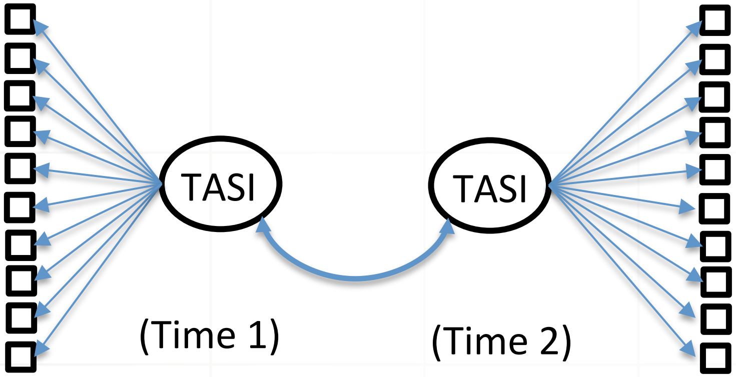
- Does my hypothesized structure demonstrate differences across subgroups?
  - We assume the TASI is assessing the same construct across all types of groups
  - If this does not hold then the TASI does not represent the construct equally well and we cannot interpret scores from the TASI across groups
    - Given that males and females may react differently to stress, we evaluated the construct invariance of the TASI across gender
      - Results demonstrated factor loadings were invariant across groups (metric invariance; Kline, 2011)
      - Indicates our factors have the same meaning across groups



Validity Evidence

Examples of construct evidence for "TASI"

- Is my hypothesized structure stable over time?
  - We assume the TASI is assessing the same construct across time
  - If this does not hold then we cannot interpret change in the TASI over time because our construct is not being measured in the same way
    - We evaluated longitudinal invariance of the TASI using the same group of teachers at time 1 (Spring 2015) and time 2 (Spring 2016) (Kline, 2011)
      - Results demonstrated factor loadings were invariant across time (metric invariance), so our factors have the same meaning at both time points



Validity Evidence

Aspects to keep in mind

- Potential threats to validity
  - Construct underrepresentation
    - Measure fails to fully capture construct
  - Construct irrelevant variance
    - Some aspect was included in the measure that was not part of the intended construct
- Consequences
  - Be mindful to consider and evaluate potential consequences of score interpretation/use

FINAL THOUGHTS

Final Thoughts

- Psychometric information is ***sample and purpose specific***

- The validation process (accumulation of evidence) is a continual process
  - Your job is never done
  - It is up to you to build a body of evidence

- A single evaluation with a single population is not sufficient to claim scores are reliable/valid for a particular purpose
  - A single evaluation provides support but more evidence is always warranted

Questions?
lhawley2@unl.edu

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological tests.* Washington, DC: American Educational Research Association.

Brennan, R. L. (2001). Generalizability theory. New York: Springer-Verlag.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*, 207-230.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in Psychological tests. *Psychological Bulletin, 52*, 281-302.

John, O. & Benet- Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. *In* H.T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339 – 369). Cambridge University Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of *Educational Measurement, 50*, 1-73.

References

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kline, R. B. (2011). Principals and practice of Structural Equation Modeling (3rd ed.). Guilford.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35-44.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542-547.

Raykov, T., & Marcoulides, G. A. (2011). Introduction to psychometric theory. Routledge.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer (Vol. 1). Sage Publications.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, 405-450.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In Lissitz, R. W. (Ed.) *The Concept of Validity* (pp.19-37). Information Age Publishing.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. The Asia-Pacific Education Researcher, 22(2), 209-213.