MOUNT CARMEL
College of Nursing

Demystifying Instrument Development: A Practical Approach for Nursing Education Researchers

2016 STTI-NLN Nursing Education Research Conference

Darrell Spurlock, Jr. PhD, RN, NEA-BC, ANEF
Director, Scholarship and Institutional Effectiveness
Mount Carmel College of Nursing, Columbus, OH

Amy Wonder, PhD, RN
Assistant Professor
Indiana University School of Nursing, Bloomington, IN

Conflicts and Disclosures

Neither the planners or presenters have any real or perceived vested interests that relate to this presentation

Objectives

At the conclusion of the workshop, attendees will be able to:

- Explain various approaches to instrument development, focusing on application of common models/theories of measurement.
- Distinguish current conceptions of validity and reliability and the methodological strategies required to produce robust validity and reliability evidence.
- Select appropriate basic psychometric analyses required for instrument/scale development.
- Construct an empirically sound approach for instrument/scale development and initial testing.
- Evaluate and apply guidelines for appraisal of instrument/scale development research.

Methods

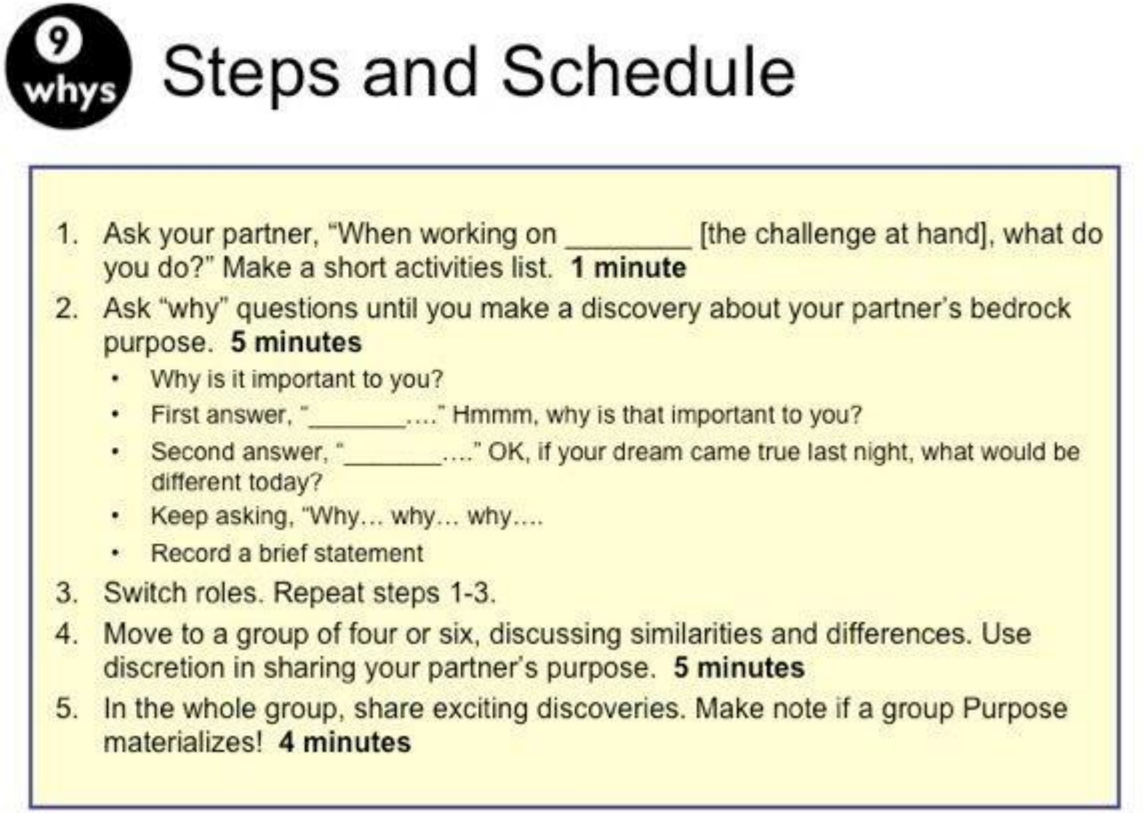
- Liberating structures
- Lecture/visuals
- Group work/discussion

Liberating Structure: 9 Whys (15-20 min)

What is made possible?

Rapid clarity about what is *essentially important* about our work -- individually, and as part of the nursing education community.

When we discover an *unambiguous shared purpose*, this can unleash both increased *freedom* and *responsibility*.



9 whys Steps and Schedule

1. Ask your partner, "When working on _________ [the challenge at hand], what do you do?" Make a short activities list. **1 minute**
2. Ask "why" questions until you make a discovery about your partner's bedrock purpose. **5 minutes**
   - Why is it important to you?
   - First answer, "________...." Hmmm, why is that important to you?
   - Second answer, "________...." OK, if your dream came true last night, what would be different today?
   - Keep asking, "Why… why… why….
   - Record a brief statement
3. Switch roles. Repeat steps 1-3.
4. Move to a group of four or six, discussing similarities and differences. Use discretion in sharing your partner's purpose. **5 minutes**
5. In the whole group, share exciting discoveries. Make note if a group Purpose materializes! **4 minutes**

Core Challenges

- Lack of standardized nursing education outcomes
- Lack of standardized measures to evaluate nursing education outcomes
- Low/variable faculty skill to evaluate instrument use in nursing education research
- Low/variable researcher skill to develop and test new measures and instruments for use in nursing education
- Few options to prepare nursing education researchers to develop the skills necessary for instrument development
- Low skill in interdisciplinary collaboration to address the above issues

Nursing Education Outcomes

- Nursing education has a fairly robust framework for identifying *what* students should learn; less of a framework guiding *how* they should learn it; and almost no framework guiding *when* and *to what extent*.
- Less guidance is available on outcomes outside the knowledge domain. Examples: communication skills, team skills, attitudes, etc.
- This is not specific to nursing education. All of higher education struggles with these same questions…but not all of higher education is practice-based where KSAs for practice should be measurable.

Nursing Education Outcomes

- Difficulty in measuring these important outcomes:
    - Attitudes
    - Thinking ability (e.g., critical thinking)
    - Ability to work in teams
    - Ability to perform procedures/skills – *are we making progress on this with simulation?*
    - Non-clinical skills: informatics, EBP, organizational dynamics, leadership
- Key questions:
    - *At what rate* are attitudes and skills developed?
    - What sort of practice is required to sustain them?
    - How quickly do they degrade over time?
    - What is the responsibility of the nursing education program? Of the practice setting?

Nursing Education Outcomes

- Knowledge is more easily measured than skills, attitudes, or abilities.
- *Knowledge*, as a concept, remains ill-defined for a practice profession.
  - Need updated language and better integration with the learning sciences (educational psychology, cognitive psychology, learning design, industrial and systems science)
- In pre-licensure nursing education, test vendors supply a variety of knowledge measures to support assessment and evaluation efforts, all aimed eventually at preparation for the NCLEX-RN.

Nursing Education Outcomes

- The NCLEX-RN test plan is revised every few years and demonstrates some perceptible changes in the last few cycles.
- The composition of and types of questions used on the NCLEX-RN is also changing, making the exam more difficult to pass.
- State boards of nursing continue to require schools to achieve a pass rate that is usually near or above the national passing average.
  - This approach ensures a perpetual focus on strategies to improve NCLEX-RN pass rates since it is impossible, mathematically, for *most* schools be to above the average.

Nursing Education Measures

- Common:
  - Pre-licensure entry assessments: Often based on high school exit tests or some other parallel assessment. Examples: ATI Test of Essential Academic Skills (TEAS).
  - Graduate entry assessments: Designed to predict success in graduate study. Becoming less common across higher education. Examples: GRE, MAT.
  - Mid-curricular assessments: Knowledge tests given at various points along the student's educational pathway, often tied to subjects like medical-surgical nursing, pediatrics, etc. Available in both graduate and undergraduate programs.
  - End-of-program comprehensive tests: May be used as high-stakes exams, integrated within a course, or scheduled outside formal courses. Used more in pre-licensure programs but also available in graduate programs.

Nursing Education Measures

- Less common:
  - Tests of critical thinking
  - Tests of math, writing, and reading comprehension skills
  - Tests of information literacy
  - Tests of specific knowledge, skills, or attitudes
  - Performance evaluations using rubrics or checklists
- Challenges:
  - Lack of normative data
  - Lack of validity and reliability evidence for nursing education
  - Challenges with implementing the assessments in a fair, ethical, and appropriate manner

Nursing Education Measures

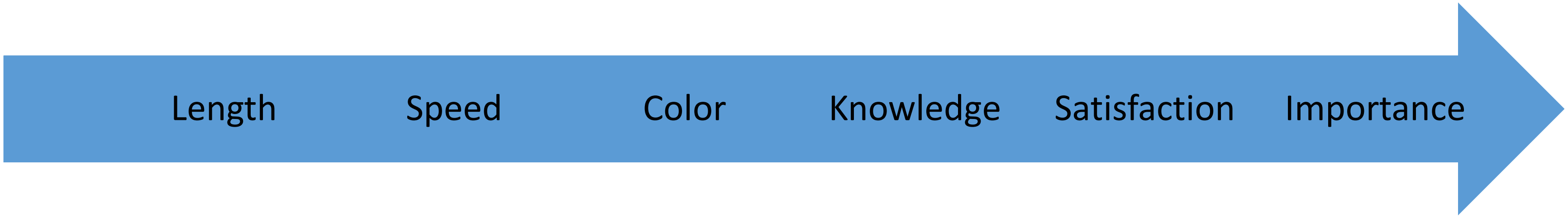
- Summary:
  - Knowledge-type tests are most commonly used – and sometimes misused.
  - Assessments of the affective domain are less common and normative data are lacking.
  - Performance assessments are increasing in number with the use of simulation, but a standardized approach is needed. Establishing normative data is expensive.
  - Observational rating scales are infrequently used and rarely available. They are difficult to develop, train to use, and apply in standardized ways across programs.
  - Few instruments specifically designed for use in nursing education focus on measuring traits or dispositions which may be of significant interest to nurse educators.

Approaches to Instrument Development:

Models and Theories

Important Definitions

- Construct: an abstract mental representation that is often inferred / inferred to exist based on observable events, behaviors, or other phenomena (DeVellis, 2012)
- Sometimes used interchangeably with terms like *phenomena, concepts, or variables*.
- Vary in level of abstraction



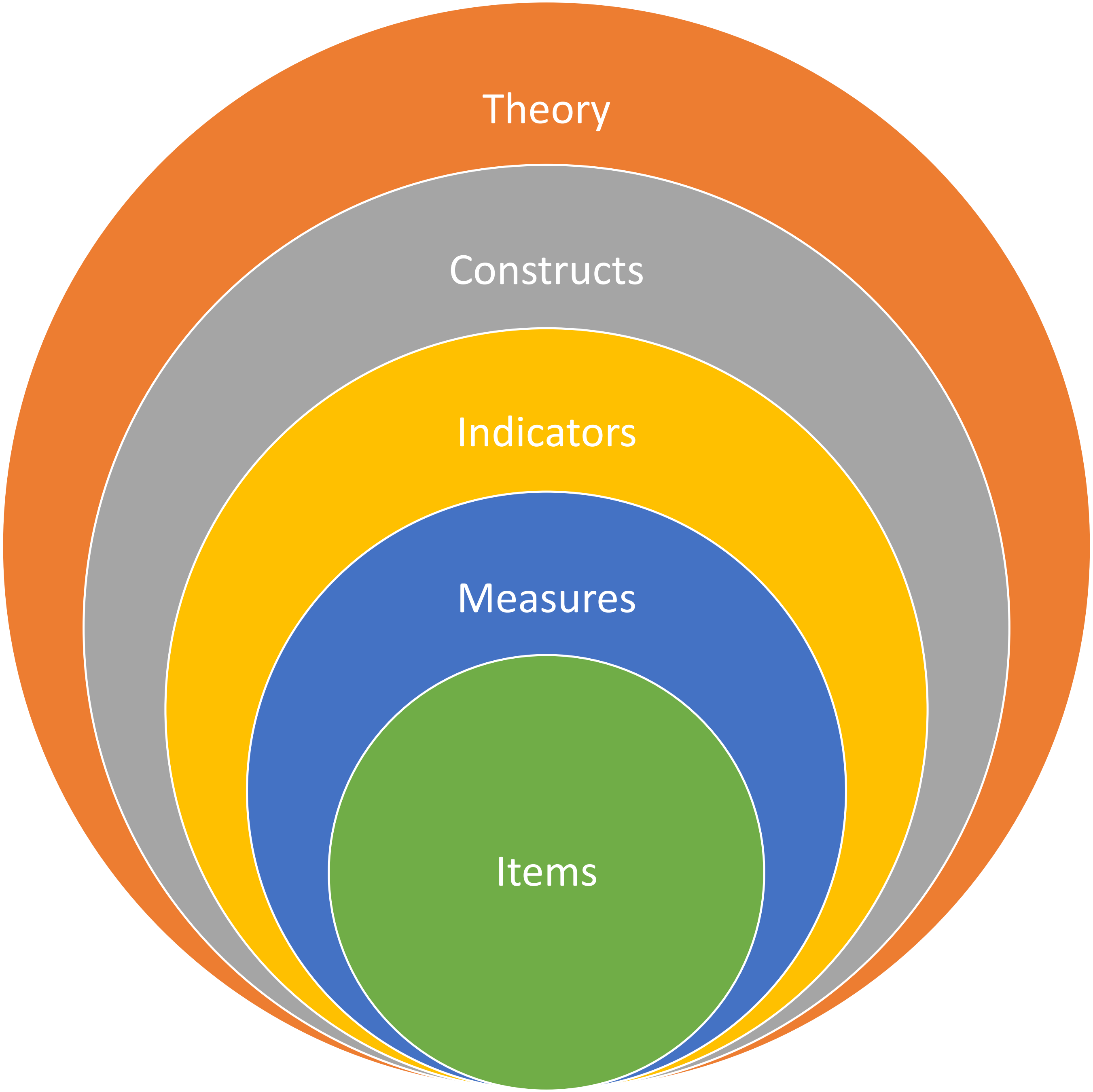
Length Speed Color Knowledge Satisfaction Importance

Important Definitions

- Latent trait: term used by researchers to describe characteristic or trait that is unobservable (using existing methods of measurement) but which can be inferred based measures of closely related constructs
- Examples: *agreeableness, readiness, introversion, resilience*, and even *knowledge*
- Latent trait measures must demonstrate evidence of validity and reliability in order to support claims that the latent trait is being measured.
- The claims a latent trait measure makes should be clear, explicit, and situated within a theory about the relationships between concepts/constructs.

Important Definitions

- Measurement: an attempt to represent a construct using an *indicator* of the construct.
- Measures: indicators are formalized into *measures*, often *scales* or *instruments*.
- Items: measures are usually comprised of *items* that when considered together with other items, form *scores*.
- Scores: scores are the basis upon which inferences about the latent trait are made.


Theory
Constructs
Indicators
Measures
Items

Why Measure?

- Polit and Yang (2016) describe three main types of measures, based on the measure's primary purpose:
  - To *predict*
    - Used to make decisions based on the likelihood of future events
  - To *discriminate*
    - Used to classify subjects into groups based on characteristics or scores
  - To *evaluate*
    - To evaluate the effectiveness of interventions or treatments
- Other organizational schemas have been suggested:
  - Thorndike and Thorndike-Christ (2011): *instructional, selection, placement, classification, and personal.*

Methods of Educational Measurement

- Though there is no universal agreement on a schema for methods of measurement, four main divisions common to nursing education are proposed (Spurlock, 2016):
    - Self-report: a method of measurement that relies on subjects to accurately report/respond to items on a measure.
    - Observation: a method that involves independent observation by another (usually the researcher) with recording of the data on a standardized rubric, checklist, or scoring sheet.
    - Psychometric: a method involving gathering of responses to a number of items, frequently statements, where the construct being measured is not explicitly clear to the respondent. Response are statistically combined to infer the amount of the underlying latent trait.
    - Tests: a method where respondents must provide the correct answer (frequently from among incorrect options) to questions or other cognitive challenges.

Examples

Self-Report

- Moral Distress Questionnaire (MDQ; Eizenberg, Desivilya, & Hirschfeld, 2009)
- Nurse Practitioners Evidence-Based Practice Survey (Melnyk et al., 2008)
- Evidence-based Practice Questionnaire (Upton & Upton, 2006)

Observation

- Hostility Coding Scale (Loucks & Shaffer, 2014)
- Clinical Judgement Rubric (Lasater, 2006)
- Surgical Procedure Feedback Rubric (SPR ;Toprack et al., 2015)

Psychometric

- Post-Exam Self-Efficacy Measure (Cleary et al., 2015)
- HIV/AIDS Stigma Instrument—Nursing Students (HASI-NS; Rosenburg, Taliaferro, & Ercole, 2012)
- Activism Orientation Scale (AOS; Corning & Myers, 2002)

Tests

- NCLEX-RN®; NCLEX-PN®
- Evidence-based Practice Knowledge Assessment for Nursing (EKAN; Spurlock & Wonder, 2015)
- Basic Knowledge Assessment Test (BKAT) – various versions (Toth, 2005)

Measurement Theory

- Measurement models help guide instrument development, and testing – which is a research endeavor at its base.
- Two main theories relevant to nursing education research: classical test theory (CTT) and item response theory (IRT).
- CTT is the most commonly used and best understood in nursing education.
- The equation used to summarize CTT is depicted as $X = T + e$, where X is the true score, T is the observed score, and e represents measurement error.
- Nursing education researchers and faculty are most familiar with terms like *internal consistency reliability* and *reliability coefficients,* such as *Cronbach's alpha.*

Measurement Theory

- One challenge with CTT is that the error term is a composite term, where the *sources* of the error are not differentiated.
- Error resulting from test construction problems, sample characteristics, changes in scores across time, etc. cannot easily be evaluated.
- Error is generally examined through the lens of reliability, which is limited often to internal consistency reliability.
- Many measures are developed on an *atheoretical* basis, without serious consideration as to which measurement model would best guide development.

Measurement Theory

- Item response theory (IRT) is an approach to measurement where both the trait being measured and the characteristics of the respondents are modeled on the same scale, a logarithmic scale.
- There are multiple techniques and models categorized under IRT, including 1-, 2-, and 3+ parameter models. The Rasch model is one form of 1-parameter IRT.
- An illustrative example of how IRT works differently from CTT:
  - Assume we weigh a bag of apples and the scale indicates 10 pounds. And then we weigh a bag of oranges and it too indicates 10 pounds. We have the same amount of fruit present in both bags. In this way, the nature of the things being measured becomes less important than the scale used to do the measuring.

Validity and Reliability

Reliability

- Reliable measures perform in consistent, predictable ways.
- Ideally, when the amount of a variable (or trait) changes, the measure can detect that change and reflect it in the score.
- Reliability is, in essence, the *extent to which a measure detects and accurately presents a change in the amount of the underlying variable or trait being measured*.
- Reliability under CTT (and in some cases, with IRT) is usually indexed using a variety of reliability coefficients.

Reliability

- Frequently used indicators of reliability:
  - Cronbach’s alpha: Internal consistency of score responses; limited because not all measures *should* be internally consistent.
  - Intraclass correlation coefficient (ICC): ANOVA-based technique to compare the variance accounted for vs. not accounted for by a measure.
  - Split-half reliability: Index of reliability calculated by examining the internal consistency of ½ of the items on a measure when compared with the other ½ of the items.
  - Test-retest reliability: Calculated by comparing the correlations between measurements when the trait or variable being measured should be stable over time. (The time interval matters here.)
  - Inter-rater reliability: The ability of a measure to produce similar measurements when the measures are conducted by different raters.

Validity

- *“Whereas reliability concerns how much a variable influences a set of items, validity concerns whether the variable is the underlying cause of item covariation.”* (DeVillis, 2012, p. 59)
- Three commonly accepted “types” of validity – though other definitions exist:
  - <u>Content validity</u>: The extent to which a set of items measures the construct of interest.
  - <u>Construct validity</u>: The extent to which the score from a measure interacts in a theoretically-consistent way with other variables. Messick (1995) suggests construct validity is the over-arching form of validity.
  - <u>Criterion validity</u>: The extent to which a scale predicts (or is predicted by) an empirically important outcome or variable.

Validity and Reliability - Notes

- Validity and reliability are not static properties of an instrument. They must be continually evaluated over time and across different groups.
- Statistical evidence alone is not sufficient evidence of validity.
    - Factor analysis only reveals the structure of the data. The structure must be interpreted by the researcher.
- Validity and reliability are demonstrated through *evidence*, developed over time, and contributed to by researchers other than the instrument developers.
- Why not be creative?

The Nursing Student Stress Scale is a valid and reliable….

Developing Instruments

Instrument development and testing is a type of research activity that is fundamental to advancing the science of nursing education.

General Steps in the Process

DeVellis (2012) outlines the following general steps in the instrument development process:

1) determining clearly the construct to be measured,
2) generating the pool of scale items,
3) determining the format of the scale,
4) enlisting experts to review the candidate items,
5) considering the inclusion of validation items,
6) pilot administration of items, and
7) evaluation and analysis of the pilot data, and
8) optimizing the length of the scale.

1. Determining the Construct

- Instrument development is, in many ways, an act of theory development and testing.
- Theories allow others in the field to know how constructs are defined, the relationships between constructs, and flowing from this, how best they can be measured.
- For many constructs of interest in nursing education, the theoretical basis for the theory may be quite limited.
  - There are often nuggets available in the literature and these should not be overlooked.
- It is important to be both sensitive and specific in developing new measures.

2. Generating the Item Pool

- Items should be generated based on definitions of the construct which have been previously set out.
- Generate more items than is necessary (for the final product) since you will eliminate some along the way.
- Use best practice guidelines when possible, especially for test item writing.
- Do not copy items from other scales, either in whole or in part, to generate a new form.
  - Other scales can *inspire* a new scale, but unless it is very clear that a new scale is an attempt to improve or revise and existing one, taking items from other scales is intellectually dishonest.
- It is probably best for those generating the items to be steeped in the literature surrounding the construct to aid in conceptual clarity and efficiency.

2. Generating the Item Pool

- Duplication/redundancy can be both a feature and a bug.
- Desired scale length should play a role in determining how many items to develop.
  - Brief scales tend to be more difficult to develop than longer scales due to issues of power
  - Long scales increase respondent burden and thus may be less useful in some study designs
- Consider how the items will be administered – via computer, in person, etc.
- “Design thinking” is important at this stage too.
  - Measures must align well with how they will be used in practice.

2. Generating the Item Pool

- With the format of the scale in mind, begin to write statements which seem to address the construct of interest.
- Review items for clarity in language. The reading difficulty level is always important, and may be especially so with certain types of measures.
- Follow rules of grammar even in items for your measure. Avoid ambiguous pronouns, misplaced modifiers, etc.
- Avoid confusing uses of positively or negative worded items whenever possible. In some cases, it is strategic to use them to detect response bias.

3. Determining Scale Format

- Thurstone scaling: Items are developed to examine the range or intensity of the target trait/latent variable, often crossing 0/neutral into negatively worded items.
  - Often uses sequence of words in the items like, *only*, *usually*, *none*, or *against*.
- Guttman scaling: Items are developed to measure increasing intensity, without the obvious need to go below 0/neutral in intensity.
- Width of response categories: When possible, the widths between response categories should be approximately equal; if not, this will be lost in most quantitative analyses

3. Determining Scale Format

- Seek to be clear in item answer options/categories.
- More options/response categories tend to be more confusing/more difficult to discriminate for respondents.
    - Example: Many, Some, Few, None vs. – Many, Several, Some, Few, One or Two, None.
- Precluding equivocation in scales is becoming more common in practice due to research which shows mildly forced options yield better measures of actual attitudes/feelings.
    - For example, eliminating “neither agree nor disagree”
    - In some cases, equivocation may be the subject of interest, so there is no requirement to exclude neutral response options.

3. Determining Scale Format

- Likert scale: Scale item is presented in a statement and the response scale involves levels of agreement with the statement.
- Likert-type scale: Other types of scale items similarly designed and presented, but there the response scale is not an agreement scale.
  - Examples: Frequency, amounts (counts), importance.

Example:

Li

Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree

Likert-type: The government provides better services than the private sector.

Completely True	True	Neither True nor Untrue	Untrue	Completely Untrue

3. Determining Scale Format

- Semantic differential: A bipolar descriptor is presented and respondents place an X or line in the spot between the two words to indicate the location between the descriptors. Responses are typically then categorized on a response scale. Example:

Completely Truthful ___ ___ ___ ___ ___ ___ Completely Untruthful

- Visual analog: Two descriptors, usually polar opposites, are separated by a line usually, 100 mm, and respondents are asked to mark the place on the line that reflects where they fall between the two descriptors. Example:

Worst Possible ____________________________Best Possible

3. Determining Scale Format

Other considerations:

- Binary items: Constructed as Yes/No, True/False. Useful when paired with scaled statements reflecting a range of the trait or variable being measured.
- Time frame: Time frame is important in item wording, or in the instructions leading to the items. If the underlying trait or variable being measured is likely to change, it may be important to specify a time frame, such as, “Within the last month,….”
- Scale length: Mentioned earlier – overall scale length can impact the usefulness of a measure. When would a 20-item measure be preferred over a 40-item one?

4. Expert Review

- Experts in the construct of interest should be empaneled to provide feedback on the items, and answer a very important question:
  - To what extent does this item measure (or reflect) the construct of interest?
  - This question supplies data to calculate the Content Validity Index (CVI) for an item and the scale (and subscales, if proposed)
- Experts can also be asked the extent to which the item is *clear* and *concise*.
- Experts can provide narrative feedback or suggestions for improving certain items.
- Instrument developers must then take this feedback and make improvements or edits to the items. Items may be eliminated or retained. Several rounds of review may be necessary.

5. Validation Items

- Some scale developers may chose to include validation items in their scale.
- This can take the form of including items from existing social desirability scales, or items from scales parallel to the new one being developed (as applicable, and with permission, of course).

6. Administer the Items

- The pool of candidate items should be administered to a group representative of the population in which the instrument will ultimately be used.
- Depending on the measurement model, power calculations (or other sample size requirement parameters) may come into play.
  - In general, the larger the sample, the better. If the study is underpowered, it may be hard to make decisions about items that should stay, be revised, or be discarded
- Though optional, respondents may be asked to provide comments or feedback about item clarity, suggested revisions, etc.
- Consider methods which provide for validity evidence during the first administration of the item pool/draft scale. For example, if a new scale was designed to measure *test anxiety*, and if the literature has shown a correlation between test anxiety and generalized anxiety, perhaps co-administering an existing anxiety measure would be prudent in this case.

7. Evaluating Items

- Primarily a statistical undertaking:
  - Descriptives (mean, SD, var) for the items and scale
  - Item analysis for “exam-type” items
    - Difficulty, discrimination, etc.
  - Inter-item correlations and item-scale correlations
  - Internal consistency reliability coefficients (Cronbach’s α)
  - Factor analysis: can be *exploratory* if there was no theory *a priori* or *confirmatory* if there was a proposed structure or theory proposed
  - If IRT is used, other parameters: item and person infit and outfit statistics, item and person reliability, DIF

7. Evaluating Items

- Choosing which items to drop and which to keep is an iterative, theoretical, and statistical exercise:
  - Items with low item-scale correlations can usually be dropped
  - Item that are too difficult or easy (in the case of an exam-type instrument) may be candidates – unless you want items to measure a wider range of knowledge on the instrument form
  - Item content may also be a factor; eliminating items based purely on statistical considerations may carve away the assessment of a particular part of the construct
- Cronbach's α is impacted by both scale length and sample size. Longer scales will have higher αs than shorter ones; short scales have achieve high αs when administered to large groups

Evaluating Instruments and Instrument Reporting

Evaluating Existing Instruments

- Barry et al. (2014) evaluated 967 articles published in health education or behavior journals between 2007 and 2010.
  - Using the journal title to group the papers, between 40-93% of papers did not report validity information and 35-80% of papers did not report reliability information.
- Many instruments used in nursing education (and other areas of nursing) have limited psychometric data available…and the data are often old.
  - Often, only the primary/first report on the measure/scale is cited by authors using the tool in later studies.
  - Apart from new reports of reliability coefficients (when appropriate), many instruments never undergo additional validity testing or revision.

Evaluating Existing Instruments

- The types of measures and instruments useful to the nursing education researcher varies widely, making appraisal of measure or instrument quality even more difficult.
- Most measurement and scale development texts provide sound procedures for developing, testing, and using knowledge tests and attitude questionnaires.
- Less clear guidance is available for developing observational measures or measures based on non-standard measurement theories.

Critiquing Existing Practices

- Under- or non-reporting of validity and reliability evidence for new measures.
- Failing to set limits on how/when instruments should be used.
- Conflating *dispositions* or *attitudes* with *ability* or *actions*. Classic case: critical thinking.
- Using proxy measures when objective measures could or should be used.
- Treating self-reports as interchangeable with more objective measures. Example: Asking subjects to rate their knowledge vs. giving them test questions.
- Translating instruments into new languages without considering how the underlying theory may be impacted by this practice.
- Creating new instruments without an appropriate underlying theory or measurement theory.
- Failing to use best practices for item writing or construction; form design, etc.

How Accurate are Self-Assessments

- Zell and Krizan (2014) meta-analyzed data from 357,547 subjects comparing self-assessments with objective measures across a range of tasks. The overall correlation was *r* = .29 (.11).

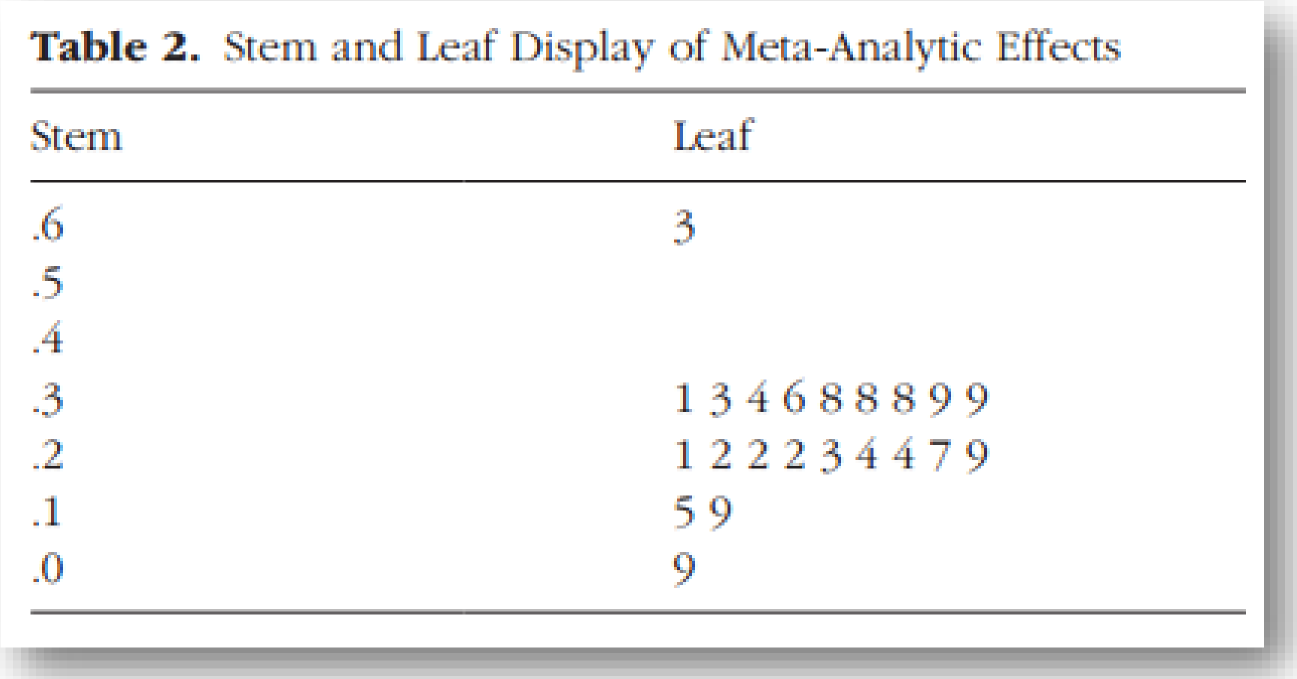


Table 2. Stem and Leaf Display of Meta-Analytic Effects

Stem	Leaf
.6	3
.5	
.4	
.3	1 3 4 6 8 8 8 9 9
.2	1 2 2 2 3 4 4 7 9
.1	5 9
.0	9

Domain and Task Moderators

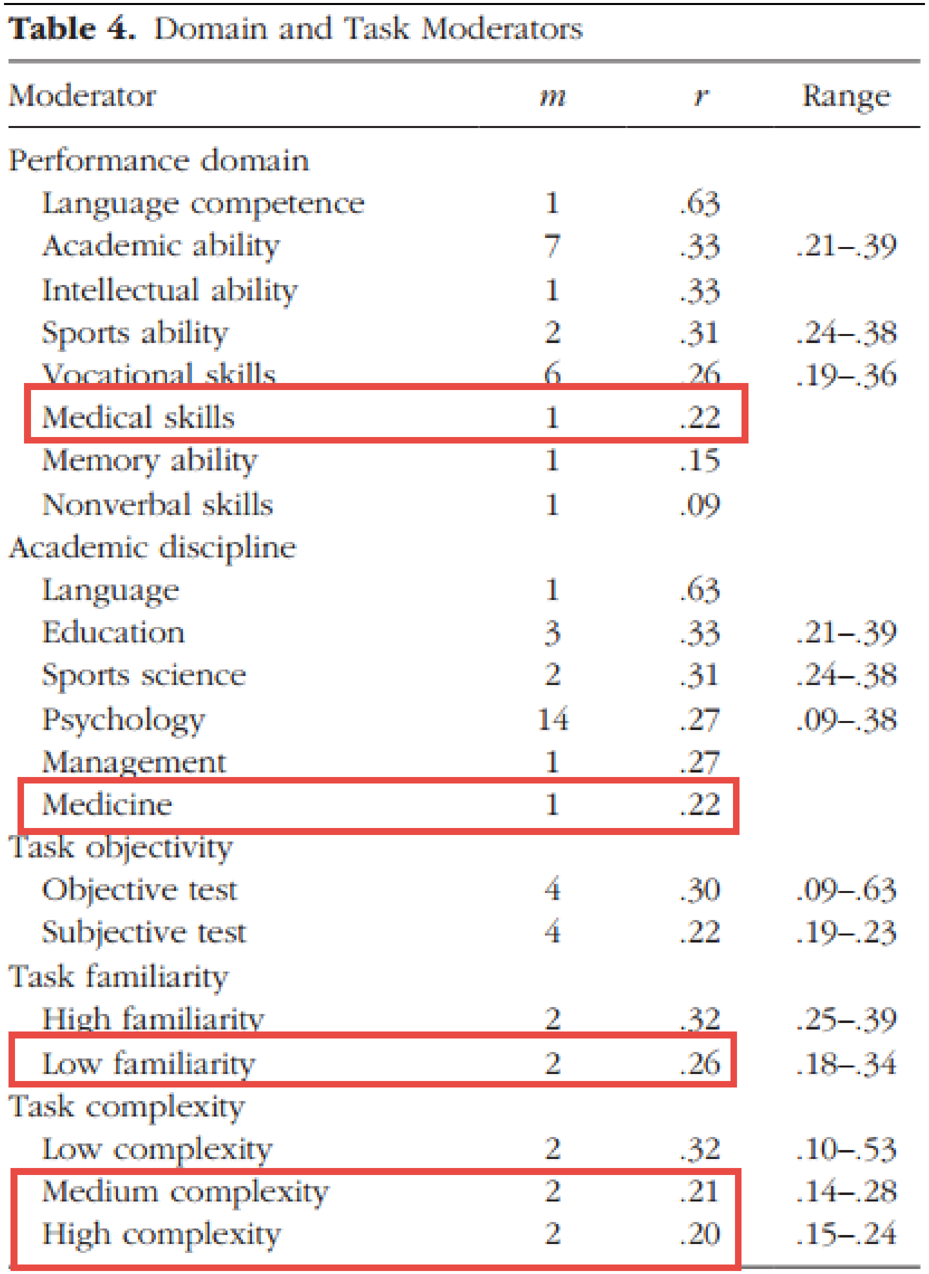


Table 4. Domain and Task Moderators

Moderator	*m*	*r*	Range
Performance domain			
Language competence	1	.63	
Academic ability	7	.33	.21–.39
Intellectual ability	1	.33	
Sports ability	2	.31	.24–.38
Vocational skills	6	.26	.19–.36
Medical skills	1	.22	
Memory ability	1	.15	
Nonverbal skills	1	.09	
Academic discipline			
Language	1	.63	
Education	3	.33	.21–.39
Sports science	2	.31	.24–.38
Psychology	14	.27	.09–.38
Management	1	.27	
Medicine	1	.22	
Task objectivity			
Objective test	4	.30	.09–.63
Subjective test	4	.22	.19–.23
Task familiarity			
High familiarity	2	.32	.25–.39
Low familiarity	2	.26	.18–.34
Task complexity			
Low complexity	2	.32	.10–.53
Medium complexity	2	.21	.14–.28
High complexity	2	.20	.15–.24

Looking Ahead

- Most PhD programs in nursing require a course on tests/measures but the content varies widely. Courses are often cross-listed with other departments.
  - Health-focused measurement courses may not be entirely useful for nursing education-focused PhD students.
- Measurement content in DNP programs is often situated within quality improvement and outcome management courses, with little connection to education.
- Master's programs with a focus on nursing education often require an assessment/measurement course, but the focus is usually on test item writing skills required for the classroom teacher.

Looking Ahead

- Instrument development is often more complex and more time consuming than other forms of research.
  - Extensive literature reviews are often required.
  - Enlisting the help of experts to review items is time-intensive.
  - Sample size requirements are often higher, so recruitment can take longer and incentives are often needed.
  - The analytic methods needed are often outside the comfort zone of even experienced researchers.
- To adequately report on the essential aspects of instrument development, a longer manuscript may be required…and publishers are pushing for shorter – not longer –manuscripts.

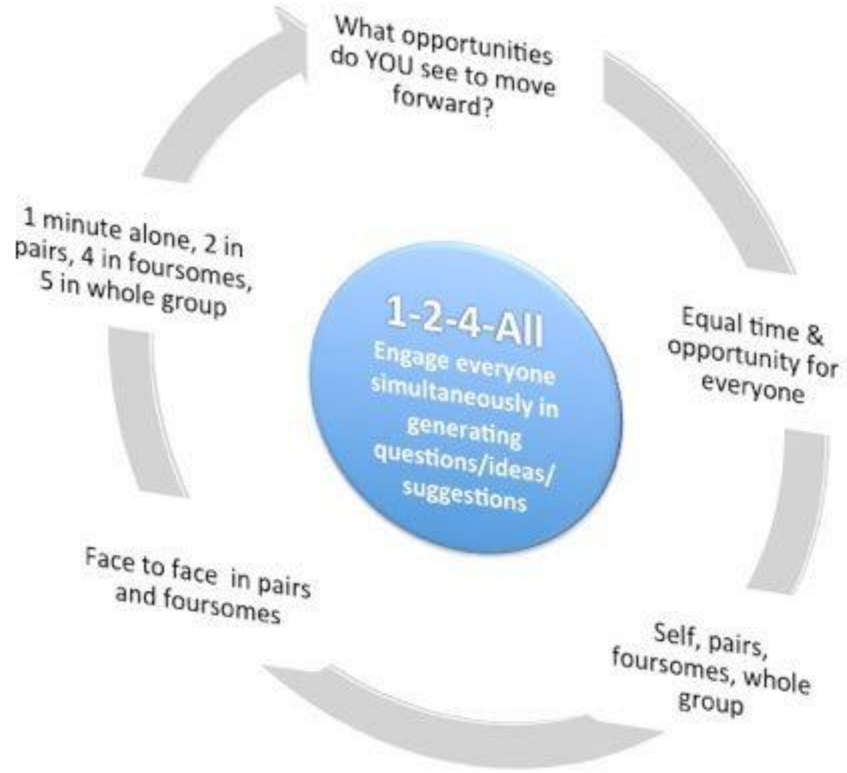
Looking Ahead

- To improve our current situation, we'll need:
  - **Collaboration**: Bringing together skills sets to develop new measures.
  - **Funding**: Instrument development work is not glamorous but still requires funding. Incentives, consultants, and site support funds are often required.
  - **Coordination**: Through developing research networks and with NLN's help, we can coordinate and prioritize development of a core of instruments/measures useful to both nursing education researchers and nursing faculty in educational settings.

Liberating Structure: 1-2-4-All (12-15 min)

What is made possible?

Inducing open, generative conversation about questions, ideas and solutions.


What opportunities do YOU see to move forward?
Equal time & opportunity for everyone
Self, pairs, foursomes, whole group
Face to face in pairs and foursomes
1 minute alone, 2 in pairs, 4 in foursomes, 5 in whole group
1-2-4-All
Engage everyone simultaneously in generating questions/ideas/ suggestions

Liberating structures images used with permission.

References

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*, 12–18. http://doi.org/10.1177/1090198113483139

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Los Angeles, CA: SAGE Publications.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741– 749.

Spurlock, D., & Wonder, A. H. (2015). Validity and reliability evidence for a new measure: The evidence-based practice knowledge assessment in nursing. *Journal of Nursing Education*, *54*(11), 605–613. http://doi.org/10.3928/01484834-20151016-01

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. http://doi.org/10.1177/1745691613518075

Resources

Artino, A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, *36*, 463–474. http://doi.org/10.3109/0142159X.2014.889814

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*, 109–117. http://doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, N.J.: L. Erlbaum.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*, 96–106. http://doi.org/http://dx.doi.org/10.1016/j.jclinepi.2010.03.002

Patient Reported Outcome Measurement Information System. (2013, May). PROMIS® Instrument Development and Validation Scientific Standards Version 2.0 (revised May 2013). Retrieved December 5, 2015, from http://www.nihpromis.org/Documents/PROMISStandards_Vers2.0_Final.pdf

Polit, D. F., & Yang, F. (2016). *Measurement and the measurement of change: A primer for the health professions*. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins.

Streiner, D. L., & Kottner, J. (2014). Recommendations for reporting the results of studies of instrument and scale development and testing. *Journal of Advanced Nursing*, *70*, 1970–1979. http://doi.org/10.1111/jan.12402

Contact

Darrell Spurlock, Jr. PhD, RN, NEA-BC, ANEF
Director, Scholarship and Institutional Effectiveness
Mount Carmel College of Nursing, Columbus, OH
dspurlock@mccn.edu

Amy Wonder, PhD, RN
Assistant Professor
Indiana University School of Nursing, Bloomington, IN
awonder@iu.edu

EKAN Information:
http://nursingmeasure.org