Best Practices in Validity: A Primer for Simulation-Based Assessment

Mark W. Scerbo, PhD, FSSH
Matthew Lineberry, PhD
Stefanie Sebok-Syer, PhD
Aaron W. Calhoun, MD, FSSH

SIMULATION:
BRINGING LEARNING TO LIFE

#IMSH2021

Disclosures

- Mark Scerbo – no disclosures
- Matt Lineberry – no disclosures
- Stefanie Sebok-Syer – no disclosures
- Aaron Calhoun – no disclosures

Session Objectives

1. Encourage and practice critical thinking about simulation assessments
2. Distinguish key components of Messick and Kane's unified validity frameworks
3. Avoid common, but less helpful, approaches to validity
4. Apply validity concepts in educational assessment and research practice

Validity – A Relevant Example and Core Concepts

Reliability and Validity - Definitions

- Reliability
  - consistency of measurement
  - the extent to which results of an instrument yield consistent results

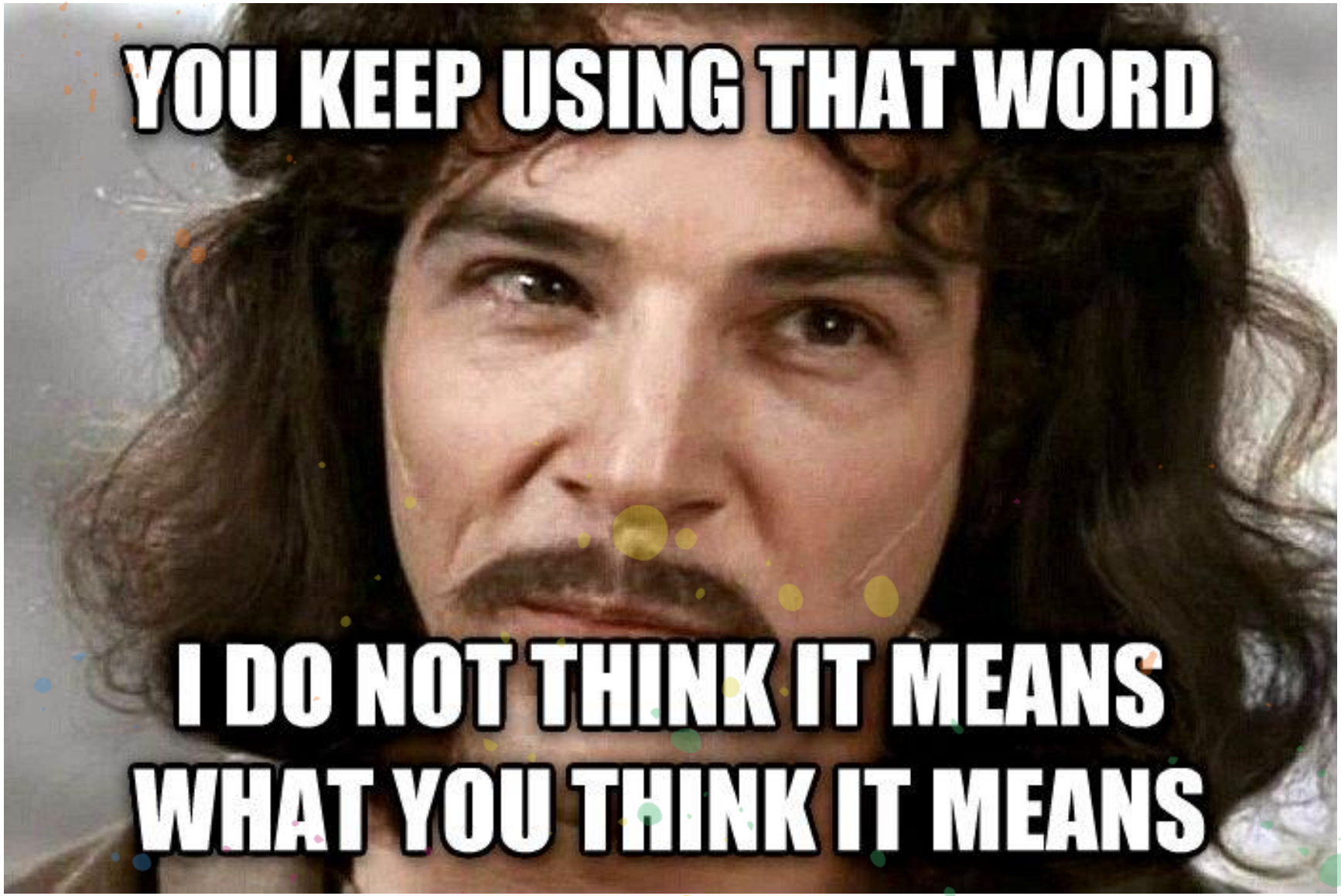
Types of Reliability - A Sample

- **Internal Consistency:** How well do specific items or subscales "hang together" in measuring one construct or a group of related constructs?
- **Inter-rater Reliability:** Do different raters score similarly?
- **Intra-rater Reliability** (Test-Retest): Does the same rater score similarly over time
- "**Classical" vs Generalizability** theory can used to tease out sources of score variability

Reliability and Validity - Definitions

- Validity
  - appropriateness of measurement
  - degree to which a test measures what it is supposed to measure

Validity -Wisdom from the Princess Bride…



Reliability and Validity - Definitions

- Validity and reliability **are not** traits intrinsic to specific tools
- Validity always relates to a specific **decision**
- Using a “validated” tool for a different purpose/target population or altering its content requires **new** data and/or logic to support this **new** usage
- Validity is **only one aspect of tool quality**

Historical Considerations

- Cronbach and Meehl (1955) - construct validity is an index of how well test results conform to the underlying construct.
- Campbell (1957) - discussed two types of validity related to experimentation: internal and external validity.
- Cook and Campbell (1979) – added construct and statistical conclusion validity
- Shadish, et al. (2002) - validity of information is determined by the consistency among empirical findings, previous findings, and theories.

Contemporary Validation Frameworks

- Like all good research, validity is **hypothesis driven**
- Conceptual frameworks are useful tools that can help organize your data as you build a case for a tool's validity
- Starting point - **the decision**
- Framework guide the collection of data to support a cohesive **validity argument**

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework. Medical Education 2015, 49: 560-575

An Example

You are the director of a simulation program at a nursing school. Recently, your dean has asked that all nursing students engage in an end-of-program OSCE focused on compassionate care. The dean further asks that a “validated assessment tool ” be used to evaluate their performance and contribute to their final grade. After searching the literature you find a wide array of tools that have been studied in various environments but are uncertain as to how to even begin to choose. The dean calls that afternoon and asks how the project is coming. How will you proceed?

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane’s Framework. Medical Education 2015, 49: 560-575

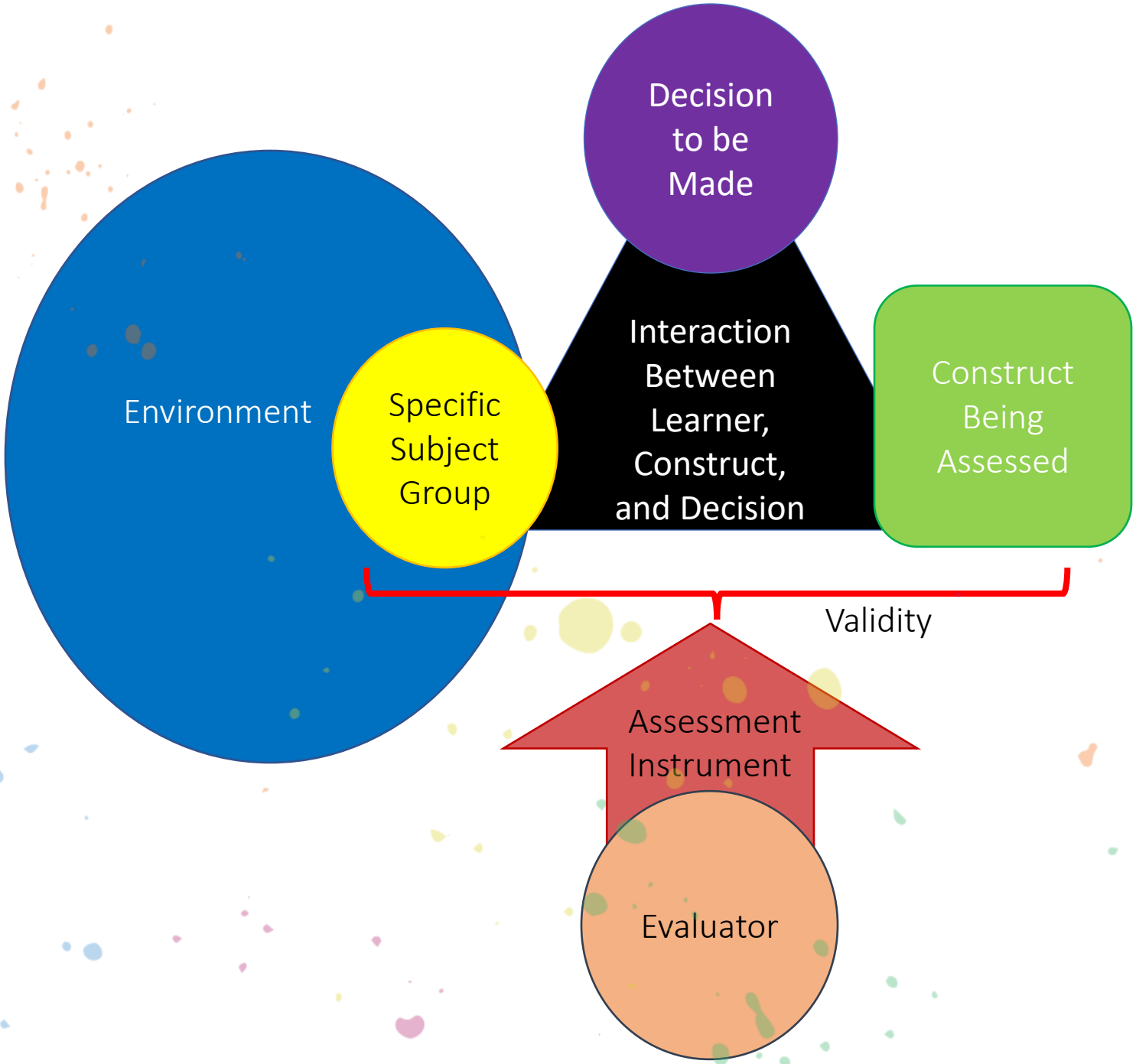
Key questions

- What is being assessed?
- Why should this be assessed?
- How will tools be selected?
- How will the results be used?
- What barriers exist to the assessment process?
- What does the phrase “validated assessment tool” really mean?

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane’s Framework. Medical Education 2015, 49: 560-575

The Intended Use Argument

- Does an instrument actually measure what we think it measures in the people we are using it on?
- *Validity* refers to a *complex relationship* between the **tool**, the **construct** being measured, and the specific learner **population** being assessed
- Argued with reference to a specific **decision** that is to be made using the scores
- *…thus, it is not transferrable between populations without further evidence to support that decision.*


Decision to be Made
Environment
Specific Subject Group
Interaction Between Learner, Construct, and Decision
Construct Being Assessed
Validity
Assessment Instrument
Evaluator

Example – Are most medical school applicants empathic?

- Environment - medical school
- Subject Group – medical school applicants
- Construct - empathy
  - the ability to feel an appropriate emotion in response to another's emotion and the ability to understand the others' emotion
- Assessment instrument – the Empathy Quotient (Baron-Cohen & Wheelwright, 2004).
- Evaluator – Admissions Director
- Decision – Admission/Rejection

The Validity Argument- Making a Case

The Validity Argument

- Similar to arguing a case in court
- **Overall structure of the case**, from opening to closing statements, makes the case
- Different **streams of evidence** (witnesses, forensics, etc.) are woven together throughout
- Two commonly used frameworks address these different aspects of the process
  - **Messick** – Addresses streams of evidence/categories of data
  - **Kane** – Addresses the overall structure of the case/validity argument

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework. Medical Education 2015, 49: 560-575

Messick's Framework- Streams of Evidence

Content Evidence	Does the content of the tool match the construct being measured?
Response Process	What is happening between the question being asked and the score being generated? How is this affecting the quality of the data generated by the question?
Internal Structure	How reproducible is that data (overlaps with Reliability)? Are all the items assessing the same construct?
Relationship to Other Variables	How well does that data match other accepted measures of the same construct (if they exist)?
Consequences	What will the data be used for and can that use be supported?

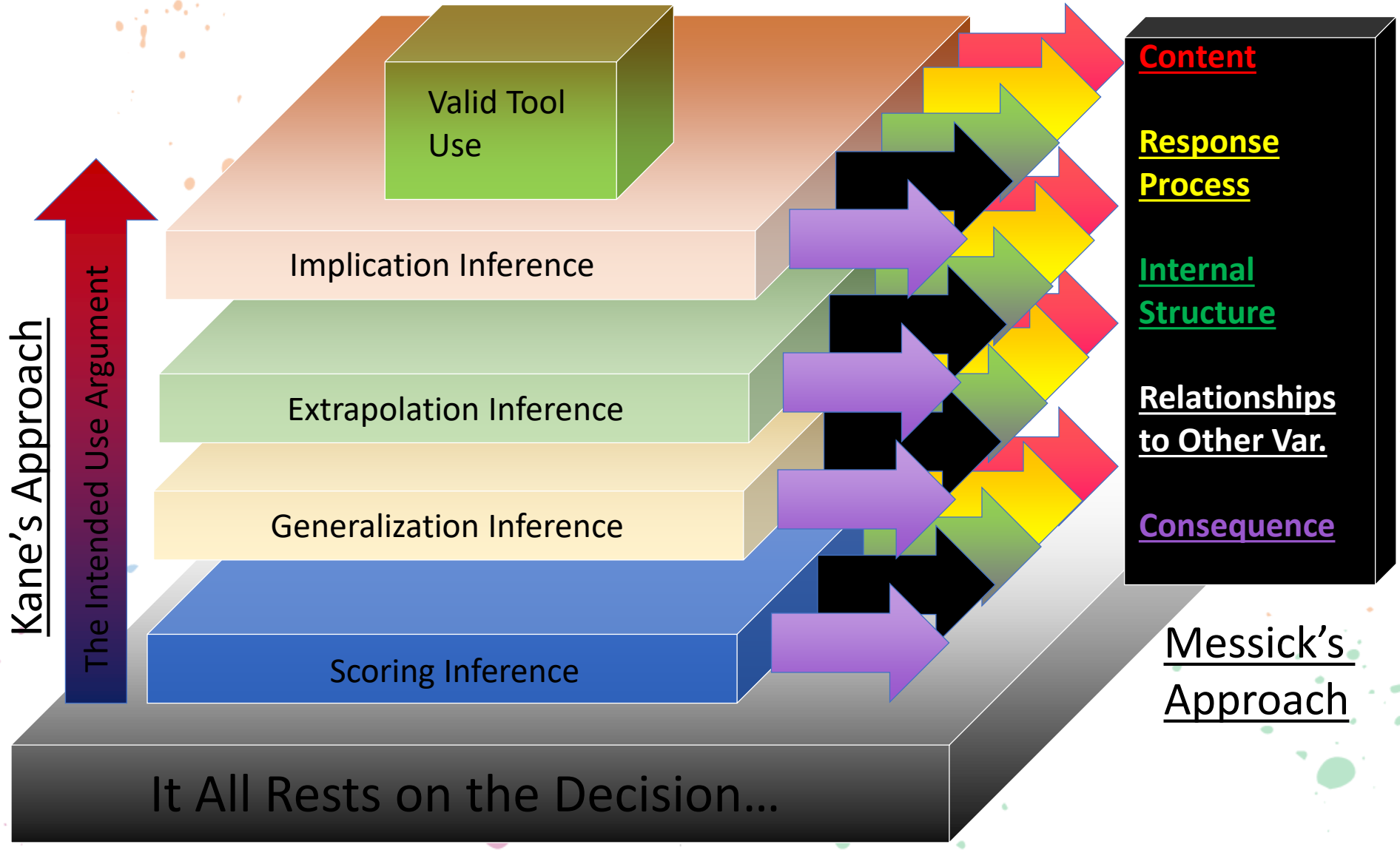
Downing S. M. Validity: On the Meaningful Interpretation of Assessment Data. Medical Education 2003;37:830-837.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing (2014) Washington, DC: American Educational Research Association.

Kane's Framework-Arguing the Case

Decision/Intended Use Argument	The framework begins with a statement of the decision that the tool is intended to facilitate and an outline of the argument you will use to make the case.
Scoring Inference	Do the individual sub-scores within the tool accurately reflect the specific observations made by raters?
Generalization Inference	Can those sub-scores be effectively combined to create an overall score that accurately reflects global subject performance in the setting of the assessment?
Extrapolation Inference	How accurately does that overall performance score reflect performance or phenomena in the "real world"?
Implication Inference	Given the strength of the prior three inferences, how confident can we be in our use of the tool to help us with the decision we started with?

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework Medical Education 2015, 49: 560-575

Combining the Two Approaches- An "Orthogonal" Relationship



“You Don’t Need to Catch Them All”

Validating a Tool: A Practical Approach

- Define the construct and proposed interpretation
- Make explicit the intended decision(s) or conclusion(s) that your data will need to address.
- Define the interpretation-use argument (i.e. the validity hypothesis), and prioritize needed validity evidence
- Identify candidate outcome measures
- Appraise existing evidence and collect new evidence as needed
- Keep track of practical assessment issues including cost
- Formulate/synthesize the validity argument in relation to the interpretation-use argument
- Make a judgment: does evidence support the intended use?

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation*, 1–12.

Back to our Example

- Armed with this new knowledge, how would you approach the question of how best to assess your nursing students on their ability to give compassionate care?
- How would you go about selecting a tool?
- What would you be looking for in terms of the “validity argument” for using these tools in your context?
- How would you decide how to implement it?
- How would you ultimately decide if the approach you chose resulted in “valid” scores for the decision you are trying to make?

Threats to Validity

- Construct Contamination
- Under-Representation

Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework. Medical Education 2015, 49: 560-575

Validity in Differing Contexts

- Verification, Validation, and Accreditation (VV&A) in Engineering Contexts (Balci, 2003; Barnes & Konia, 2018)
  - **Verification** – is the product or system built correctly, according to specification?
  - **Validation** – does the product or system meet the customer's needs?
  - **Accreditation** - does the product/simulation or system meet a third party's requirements?

Take Home Lessons

- The statement “this is a valid tool” is wrong by definition
- Tools are only valid to make particular decisions in particular populations/contexts
- Arguing that a tool is valid for a specific use should be done using a framework
- Its usually easier to find one than to make one
- When looking for a tool in the literature, consider how they made the argument

Key Articles For Reference

- Barnes JJ, Konia MR. Exploring validation and verification: How they different and what they mean to healthcare simulation. Simul Healthc 2018. 13(5):356-362.
- Cook DA, Brydges R, Ginsburg S, Hatala R. A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework. Medical Education. 49(6):560-575. Jun 2015.
- Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin, 1955;52:281-302.
- Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003. Sep;37(9):830-7.
- Kane MT. Validating the interpretations and uses of test scores. Journal of Educational Measurement. 2013 Mar 1;50(1):1–73.
- Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50(9):741–9.
- Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin, 2002.

Best Practices in Validity: A Primer for Simulation-Based Assessment

Mark Scerbo, PhD, FSSH
Matthew Lineberry, PhD
Stefanie Sebok-Syer, PhD
Aaron W. Calhoun, MD, FSSH

THANK YOU!