MAYO CLINIC

Contemporary Thinking on Assessment Validation

Arguments, Decision, & Kane's Framework

David A. Cook, MD, MHPE

SDRME June 2015

Workshop outline

- **Decisions**
- The validity argument
- Key inferences
- Pearls for scholars

A contemporary approach to validity arguments: a practical guide to Kane's framework

David A Cook,[1,2] Ryan Brydges,[3,4] Shiphra Ginsburg[3,4] & Rose Hatala[5]

Medical Education 2015: 49: 560–575

Assessment = decision

Decisions

- Selection
- Learning (direction, motivation, feedback)
- Tailoring (mastery, accelerate/remediate)
- Certification (competence, milestones)

How good are the assessments we're using?

Assessments are diagnostic tests

Pulmonary embolism?

- History
- Exam
- CBC
- D-dimer
- Chest x-ray
- CT angiogram

Competent physician?

- Rotation shelf exam
- OSCE
- Simulation procedural assessment
- Certifying exam
- Workplace observation

Integration → Decision / Action

Diagnostic Tests → Decisions

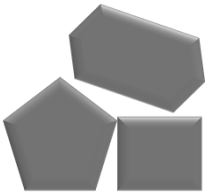
- Prostate specific antigen (PSA)
  - Decision: ***treat*** prostate cancer?
  - (Not just diagnosis)
- Learning styles
  - Decision: adapt instruction?
  - (Not just diagnosis)
- Mini-CEX
  - Decisions: What feedback? Remediation?

What's the decision?

Medical student	Resident	Practitioner
Preclinical Friday quiz	Simulation-based procedural test	Certification exam
Preclinical final exam	Mini-CEX direct observation	Patient satisfaction survey
Clinical skills OSCE	Clinical quality metrics	Clinical quality metrics
USMLE Step I	Clinical rotation grade	Workshop evaluation form
Clinical rotation direct observation	Program director final evaluation	Clinical teaching assessment
Clinical rotation shelf exam		

Programs of assessment

Diagnostic *strategies*: Programs of assessment



Workshop outline

- Decisions
- **The validity argument**
- Key inferences
- Pearls for scholars

A Parable … Collecting the evidence

- Detective #1 – trust your gut
- Detective #2 – Law & Order
- Detective #3 – formal model as trained
- Detective #4 – strategic use of model

A Parable … Interpreting the evidence

- Attorney #1 – run to the judge
- Attorney #2 – logical organization
- Attorney #3 – use framework, careful argument
- Attorney #4 – use framework to identify gaps in argument, fill gaps

A Parable … Making the decision

- Juror #1 – throw away the key
- Juror #2 – 55%, lock him up
- Juror #3 – 55%, insufficient evidence
- Juror #4 – 55%, but 90% for lesser charge

Validity

- "… the degree to which **evidence** and theory support the **interpretations** of test **scores** entailed by proposed **uses**"
  – AERA / APA 1999

Validity frameworks: A very brief history

1920: Types of validity
- **Criterion validity**
- **Content validity**

What if no gold standard? Risk of confirmation bias.

1950: Types of validity
- **Correlational validity**
- **Content validity**
- **Construct validity**

Too many types. Everything relates to the construct. Where to fit reliability?

1989: Sources of evidence
- **Content**
- **Response process**
- **Internal structure**
- **Rel. other variables**
- **Consequences**

How to prioritize evidence?

2000: Argument
- **Scoring**
- **Generalization**
- **Extrapolation**
- **Decision and use**

Argument-based approach

"The core idea is to state the proposed interpretation and use explicitly and in some detail, and then to **evaluate** the plausibility of these proposals."

M. Kane, 2013

Hypothesis

Kane – simplified!

Validation = interpretation/use + evaluate claims

Validity is an hypothesis …

- About a specific interpretation or decision
- Focused on a specific construct
- Collect evidence to support or **refute**

Observations → *Scores* ➔ **Inference** @ ***Construct*** ➔ **DECISION**

Inference, not instrument

- Valid ~~instruments~~
- Valid ***scores*** (for inference/construct/use)

Observations → *Scores* ➔ **Inference** @ ***Construct*** ➔ **DECISION**

Building the validity *argument*

State proposed decision (use)

State interp/use argument (hypothesis, claims)

Test ***weakest*** assumption(s)

Revise instrument or use as warranted

Review evidence → **Reject**

Plausible (accept)

Cook, Brydges, Ginsburg, Hatala, Med Educ 2015

What people actually *do*

Vaguely mention construct

Test ***strongest*** assumption

Accept hypothesis

Strong vs weak validity arguments

- "The weak program is sheer exploratory empiricism; any correlation of the test score with another variable is welcomed. The strong program … calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges." – Cronbach 1988
- Current model does not require theory, but does "require that proposed interpretation and use be specified clearly" – Kane 2008

Why not focus on the test?

- Appropriateness depends on context: learner, environment, application
- Reliability varies by learners (uniformly smart people = low reliability)
- Alignment of domain (few generalizable skills)
- **Focus on decisions (interpretations)**
- → Validity is contingent on the question asked

Bottom line

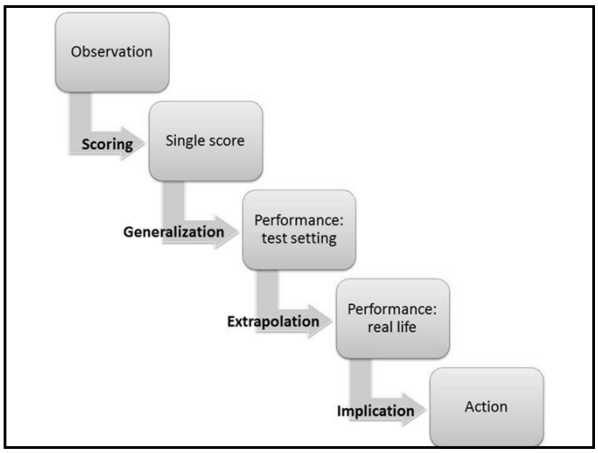
- All validity is construct validity
- Validity is an hypothesis, tested by evidence
- In the end, we want an inference (and decision)

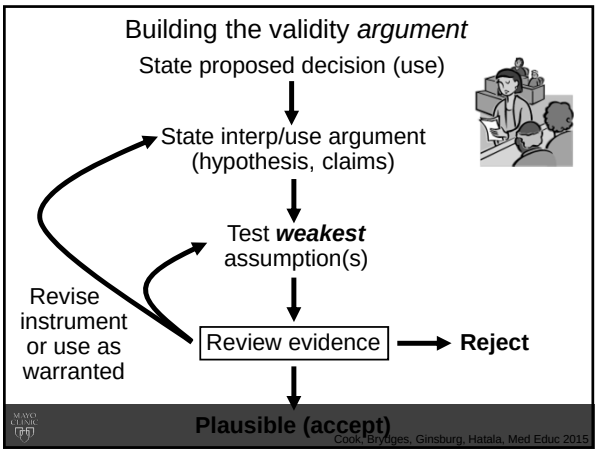
Observations →*Scores* → **Inference** @ ***Construct***

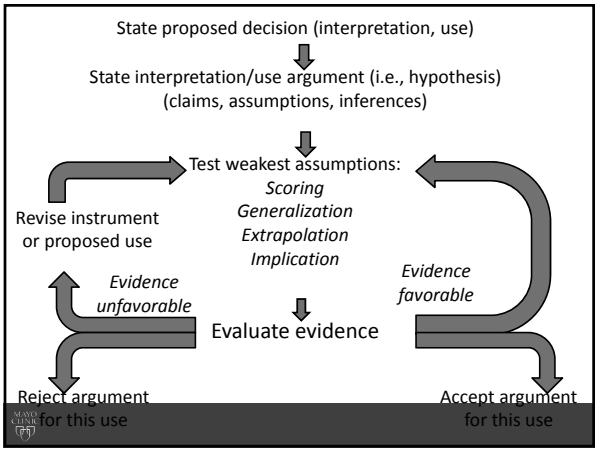
→ **DECISION**

Workshop outline

- Decisions
- The validity argument
- **Key inferences**
- Pearls for scholars







Categories to organize evidence

- Scoring
- Generalization
- Extrapolation
- Decision

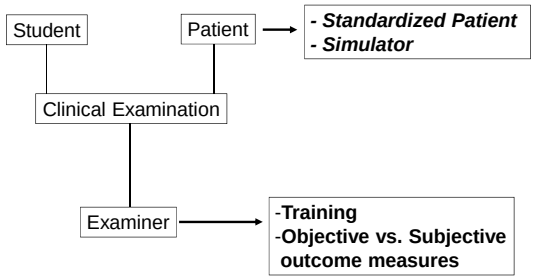
Kane M. J Educ Measurement. 2013; 50:1
Cook et al, Med Educ 2015

Scoring

- From observed performance to observed score
- Performance is the data, score is the claim
- "Did everyone get the same test?"
  - Standardization (including raters)
  - Test security
  - Item & response development (choices)
  - Scoring rubric, pass/fail standard

Variables in clinical examination

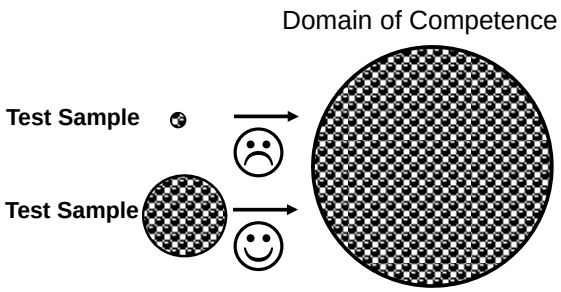
AIM: *Standardize* examiner and PATIENT



Generalization

- How well does observation + score reflect desired test domain
  - Content representative; adequate sample
- Scores reproducible across repeated test administration
  - Reliability (item, station, rater)
  - Generalizability Theory

Case Specificity

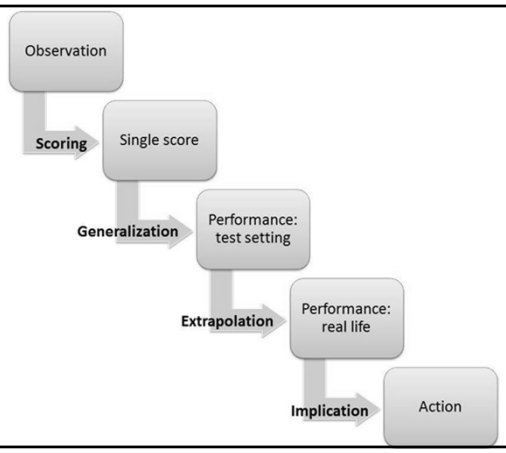


Extrapolation

- From "test" score to real-world proficiency
  - Empirical evidence shows that test scores relate to construct(s)
    - Experts >> Novices
    - Correlation with other measures
    - Improvement after training
  - Full breadth of real-world task

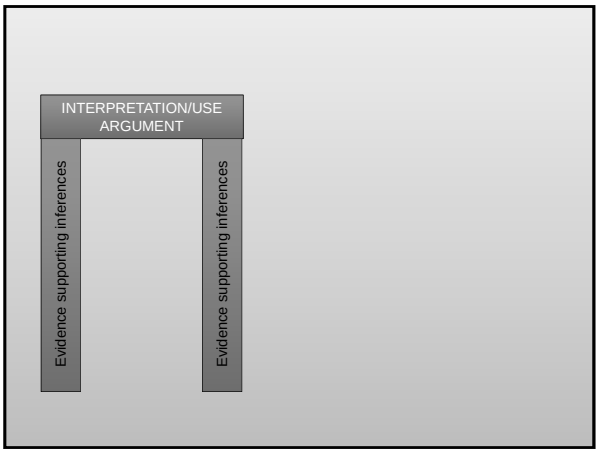
Decision

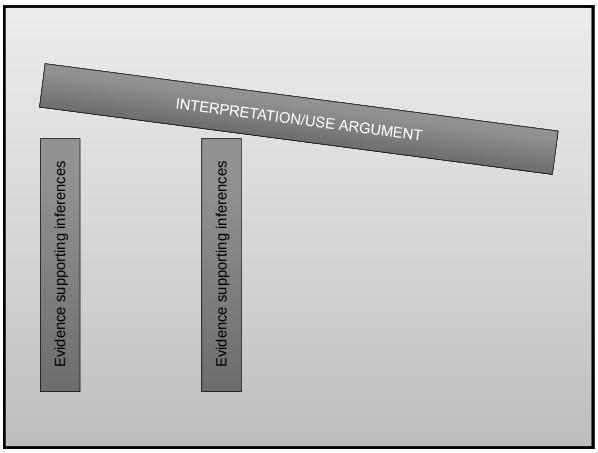
- From person's score to decision about person
- Evaluate consequences of different decisions for people with different scores
  - Intended outcomes achieved?
  - Differential impact on groups?
- Impact on learner, program, society

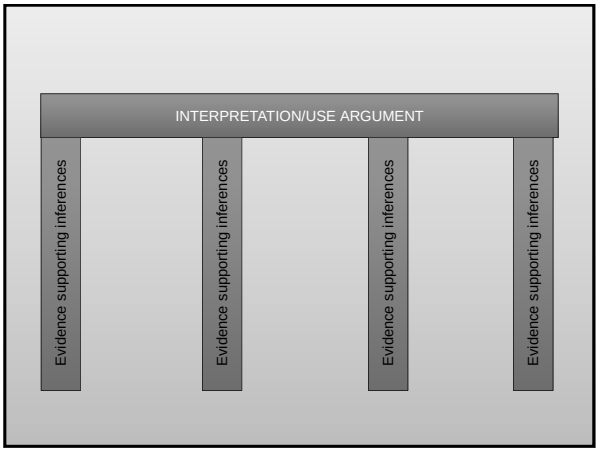


Advantages of Kane

- Formalize the hypothesis
- Focus on inferences
- Programmatic assessment
- Qualitative assessments (narrative data!!!)







MATCHING GAME

- On an OSCE, between-station correlation is low (alpha=0.33 for one station). However, acceptable reliability is achieved with six stations (alpha=0.75).

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

- The camera angle during video recording did not permit viewing of a key step in the procedure

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

- Scores for interns, senior residents, and staff were essentially the same (no significant difference)

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

- A program of assessment and remediation in laparoscopic surgery is found to reduce operative time by 33%

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

- Raters agreed on nearly all codes, except for item 17. Analysis of these responses revealed that each rater had a slightly different expectation for performance on this item.

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

- Scores showed modest correlation (r=0.55) with similar ratings with real patients

A. **Scoring** (items/response options, test format, equating, security, rater selection/training)
B. **Generalization** (sampling [blueprint], reliability)
C. **Extrapolation** (scope, authenticity [expert panel, think-aloud], responsiveness, correlation, discrimination, factor analysis)
D. **Implications, decision** (impact, standard setting, differential functioning)

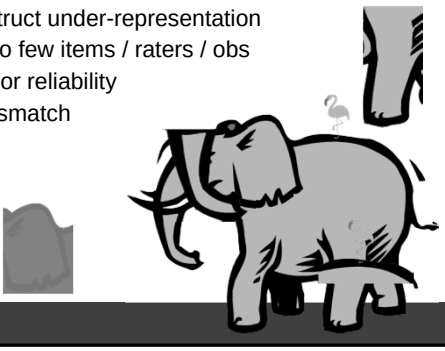
What about Face Validity?

- Not a source of evidence
- Sometimes "face validity" is really *content* evidence
- Usually "face"-type data does little to support validity of inferences or decisions

Downing, Med Educ, 2004

Threats to validity

- Construct under-representation
  - Too few items / raters / obs
  - Poor reliability
  - Mismatch

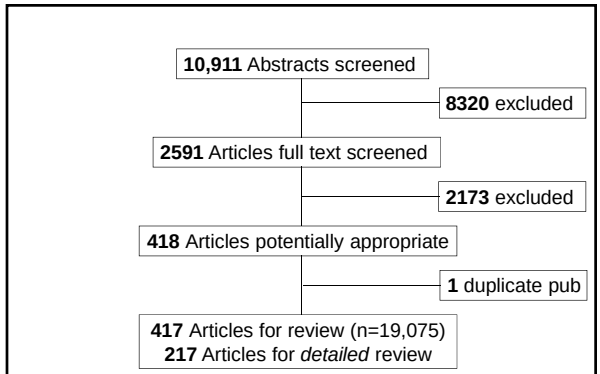


Threats to validity

- Construct-irrelevant variance
  - Flawed or biased items / rater
  - Too easy / hard, teach to test, cheating

Workshop outline

- Decisions
- The validity argument
- Key inferences
- **Pearls for scholars**



Cook et al, Acad Med 2013

> Validity theory is rich, but the practice of validation is often impoverished.
>
> - R. Brennan

How to convince a jury

Collect evidence
Plan ahead
Use an accepted framework
Prioritize evidence

Interpret evidence
Logical argument
Use a framework
Look for gaps

Impartial judgment
Specific application
Evidence needs vary

How to convince a jury
1. Plan, organize (framework)

- What evidence will you seek?
- How to interpret the results?

Framework!

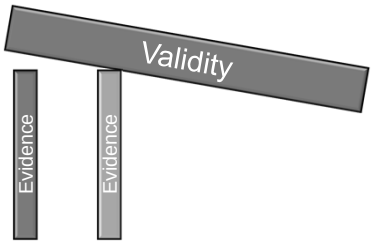
Classical	**5 Evidence Sources**	**Kane**
Content Criterion Construct *(outdated, **35%**)*	Content, Response process, Internal structure, Relationships with other variables, Consequences *(newer, **3%**)*	Scoring Generalization Extrapolation Decision *(newest, **0%**)*

How to convince a jury
2. What's the decision?

Pulmonary embolism?	Competent physician?
• History	• Rotation shelf exam
• Exam	• OSCE
• CBC	• Portfolio
• D-dimer	• Simulated procedure
• Chest x-ray	• Certifying exam
• CT angiogram	• Workplace observation

Integration → Decision / Action

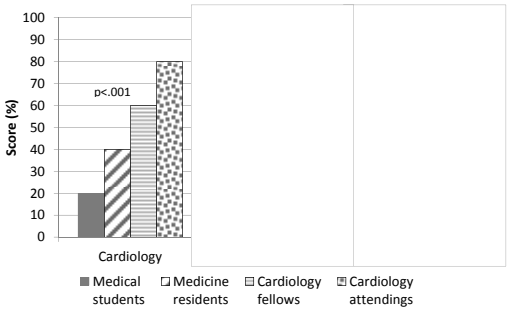
How to convince a jury
3. Complementary evidence sources



How to convince a jury
4. Don't rely on expert-novice comparisons

Cook, Adv Health Sci Educ 2014

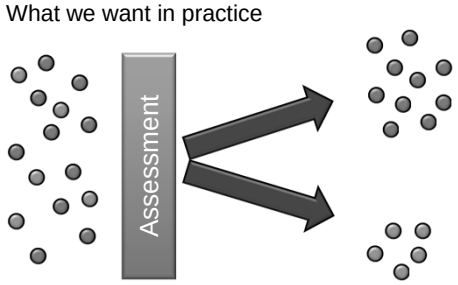
Expert-novice comparisons



Cook, Adv Health Sci Educ 2014

A side note:
Expert-novice comparisons

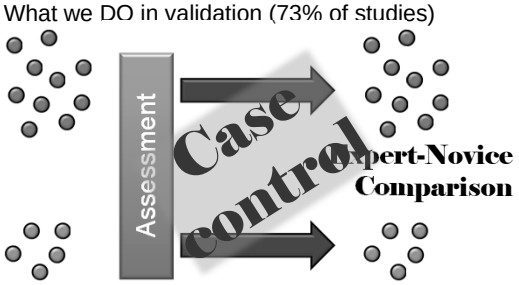
• What we want in practice



Cook, Adv Health Sci Educ 2014

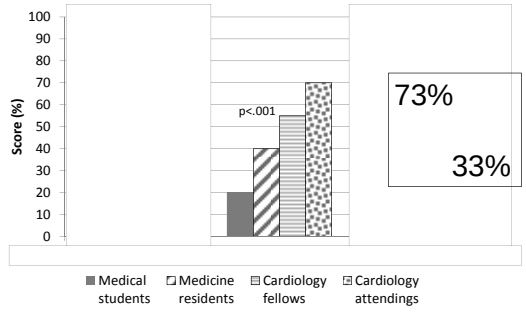
A side note:
Expert-novice comparisons

• What we DO in validation (73% of studies)



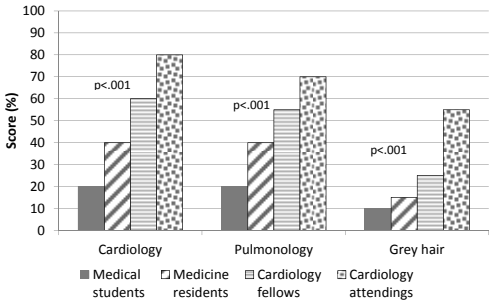
Cook, Adv Health Sci Educ 2014

Expert-novice comparisons



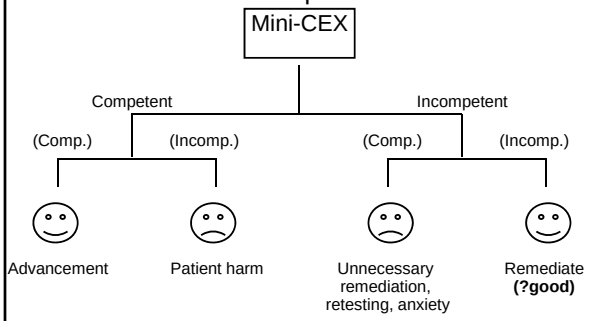
Cook, Adv Health Sci Educ 2014

How to convince a jury
4. Don't rely on expert-novice comparisons



Cook, Adv Health Sci Educ 2014

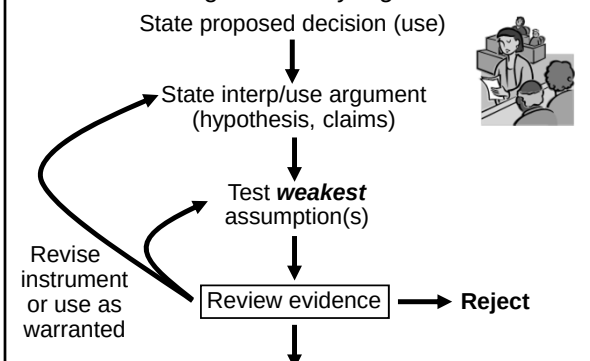
How to convince a jury
5. Think about Consequences



How to convince a jury
6. Stand on the shoulders of giants!

- 417 studies
  - Maximum 27 for any given **instrument**
  - (**Most were one-off**)
- → Don't start from scratch!

Building the validity *argument*



Cook, Brydges, Ginsburg, Hatala, Med Educ 2015

A few references

- Kane, "Validation" in Assessment Measurement 2006
- Kane, J Educ Meas 2013 – nice summary of IUA
- Cook, Am J Med 2006 – overview of Messick
- Cook, Acad Med 2013 – validity evidence and reporting quality (simulation)
- Cook, AHSE 2013 – operational details on Messick's evidence sources
- Cook, Med Educ 2015 – how to use Kane