EPSY 8224 March 25, 2019

Sources of Validity Evidence for Performance Assessments (Lane, 2010)

Definition of Validity:

* Extent to which theory and evidence support intended score interpretation for proposed uses.

Two components of validity and validation:

1. Interpretation and Use Argument – interpretive framework
2. Validity Argument – body of evidence from validation studies to defend interpretation of scores for proposed uses (to defend IUA)
3. Content representation (challenge with KSAs – fewer tasks, minimal observations, narrowly defined content domain)
	1. Decisions are based on multiple assessments and multiple measures
	2. Vary item and task formats
	3. Assessment blueprint
	4. Clearly defined content domain – articulating assessable content/contexts
	5. Expert support for the KSADs included in the assessment – these are core, relevant KSADs – expert review
	6. Rubric clearly reflects and represents targeted KSADs – might include expert review
4. Cognitive complexity
	1. Deeper and higher-level thinking skills are measured
	2. Rubric reflects and represents targeted cognitive skills (e.g., expert review)
	3. Response process evidence – think-aloud cognitive interviews, examine item-feature effects in explanatory item-response models
5. Meaningfulness and transparency
	1. Interviews of participants and score users
	2. Clarity of domains – blueprint – rubrics
6. Generalizability
	1. G-theory analyses to investigate the extent to which we can generalize over relevant facets of the measurement procedure
7. Fairness – differential validity
	1. Validity evidence works equally well across relevant groups of test takers
8. Consequences
	1. Often takes time to follow-along outcomes of score use
9. Instructional sensitivity (accessibility is another concern)
	1. Instruction causes higher performance (pre-post) – differences in the performance scores are sensitive to differences in instruction
	2. Those who had opportunity to learn (cross-sectional) (because of instruction in the specific KSADs) perform better than those who have not
10. Psychometrics

*Testing Standards*: (a) content, (b) response processes, (c) internal structure, (d) associations with other constructs (validity coefficients), (e) consequential bases.

Validating Measures of Performance (Kane, Crooks, Cohen, 1999)



Observation 🡪 observed score 🡪 universe score 🡪 target score

Inferences in Performance Assessment

1. Scoring – inference from performance to the observed score
2. Criteria for scoring are appropriate

Theory and current standards in the field are represented in the criteria

Criteria focus on high-level cognitive tasks – accommodate complex performances

1. Scoring criteria are applied as intended

Rater training and qualification process

Rater agreement – monitoring rater accuracy and consistency

Periodic rescoring of performances

1. Alternative interpretations

Controlling for construct-irrelevant features – by design and documenting as they are discovered

Bias – training and monitoring

Selection and background qualifications of raters

Clarity of task expectations (sensitivity review, expert review)

Functional equipment – user experience and familiarity

Administration conditions

1. Generalization – inference from observed score to universe score
	1. Generalizability

Over tasks, occasions, raters, setting, other facets of the measurement procedure

* 1. Alternative interpretations

The tasks are random and/or representative

Demonstrate that administered tasks are representative of the universe of tasks (universe of generalization)

1. Extrapolation – inference from universe score to target score
	1. Target domain coverage in the universe of generalization

What had to be left out?

Criterion-related validity evidence (associations with other aspects of the target domain)

To the extent that the process skills (because that why we’re using PA) involved in the performance are well understood, the link between the universe and target scores could be justified theoretically.

* 1. Alternative interpretations

Conditions of assessment may be restricted – due to practical constraints

Format of the assessment may be simulated or not authentic – high fidelity tasks are expensive

Implications for Design

* Standardization

May be improving consistency at the expense of narrowing the universe of generalization and introducing construct-irrelevant factors

* Length of the assessment
* Complexity of the assessment