APA ANNUAL CONVENTION
August 3-6, 2017 | Washington, D.C.

Tests, Test Scores, Constructs and Success in the World

Thorndike Career Award

Edward Haertel
Stanford University

APA 125
AMERICAN PSYCHOLOGICAL ASSOCIATION

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

Validity in 1951

- Cureton and “validity coefficients”

Validity is “the correlation between the actual test scores and the ‘true’ criterion scores”

Cureton, 1951

Validity in 1971

- Cronbach and “scientific inquiry into score meaning”

One validates, not a test, but an *interpretation of data arising from a specified procedure*

Cronbach, 1971

Validity in 1989

- Messick and "*adequacy* and *appropriateness* of *inferences* and *actions*" based on test scores …

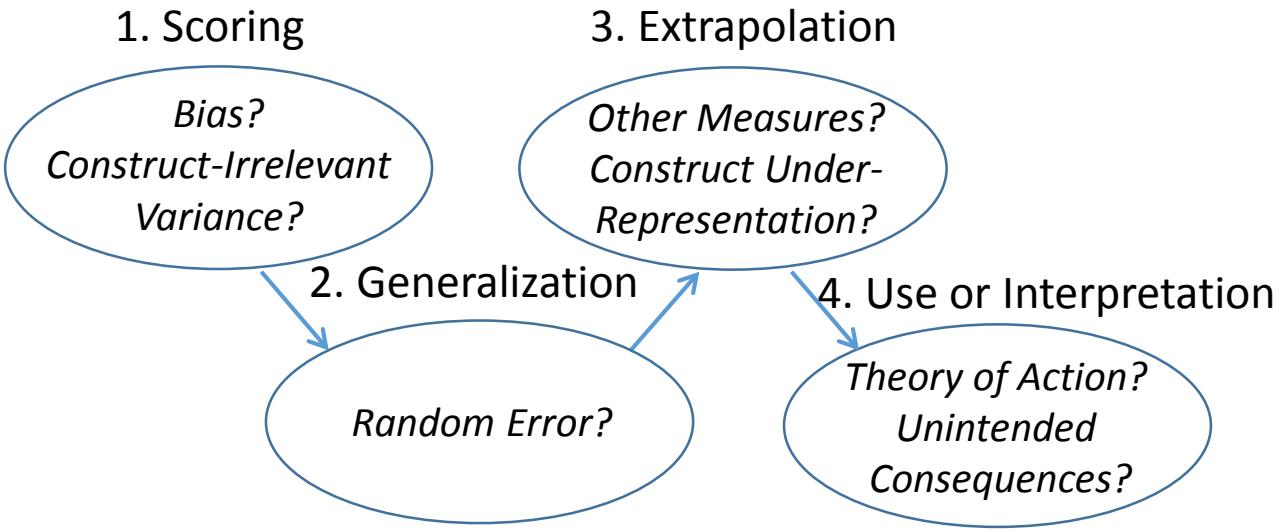
	(Inferences) Test Interpretation	*(Actions)* Test Use
(Adequacy) Evidential Basis	Construct Validity	CV + Relevance/Utility
(Appropriateness) Consequential Basis	CV + Value Implications	CV + R/U + VI + Social Consequences

Validity in 2006

- Kane
  - Development Phase of validation
    - Performed mainly by test developers
    - Focus is *Interpretive Argument*, laying out connections from test performances to intended inferences and actions
  - Appraisal Phase of validation
    - Various stakeholders may join in
    - Focus is *Validity Argument*, which investigates key propositions forming the Interpretive Argument

Validity in 2006

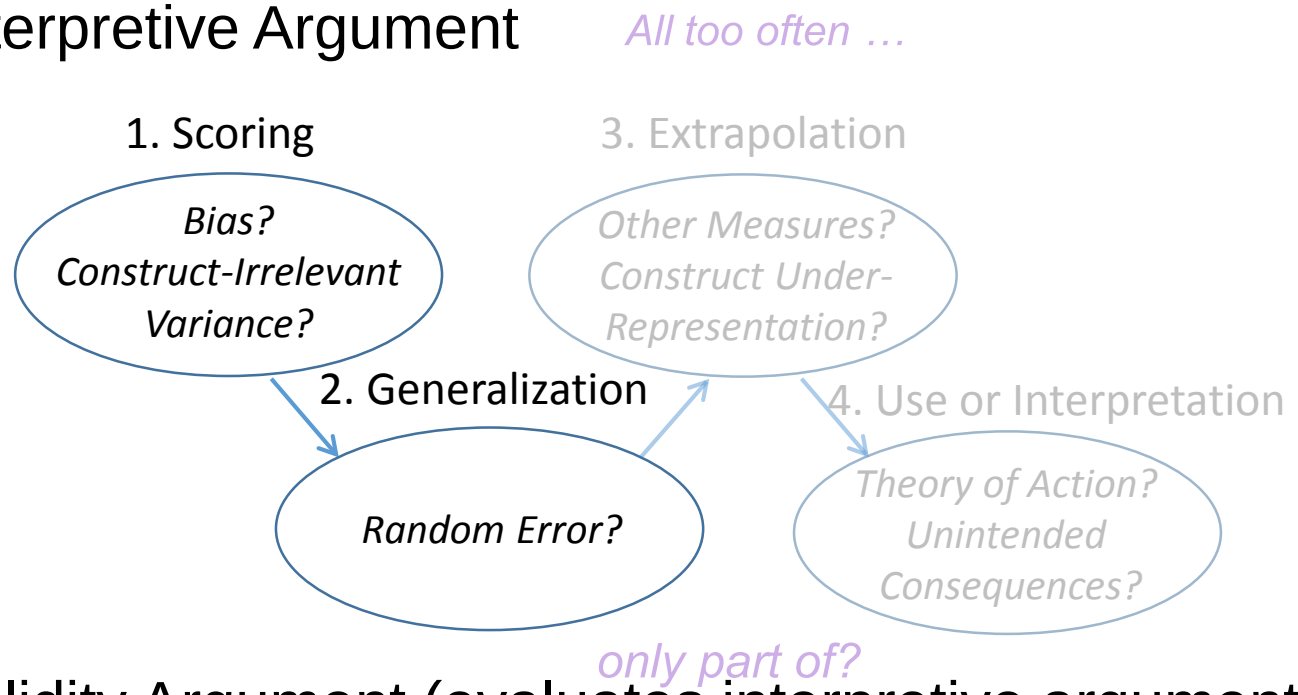
- Kane
  - Interpretive Argument



  - Validity Argument (evaluates interpretive argument)

Validity in 2006

- Kane
  - Interpretive Argument



  - Validity Argument (evaluates interpretive argument)

Validity in 2006

"The arguments for these … programs tend to claim that the program will lead to improvements in school effectiveness and student achievement by focusing the attention of school administrators, teachers, and students on demanding content. Yet, the validity arguments developed to support these ambitious claims typically attend only to the descriptive part of the interpretive argument … [focusing] on scoring and generalization to the content domain for the test. *The claim that the imposition of the accountability requirements will improve the overall performance of schools and students is taken for granted*"

(Kane, 2006, italics added).

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

IUA Example: IQ-Based Tracking

- Student tracking based on IQ test scores
  - Demonstration projects by Terman, Cubberley, et al. in Oakland, Palo Alto, San Jose *(around 1910 – 1925)*
- Assumptions include…
  - Education as direct teacher-to-student transmission
  - IQ as inherited, largely innate, stable
  - Homogeneous classroom grouping optimal

This was a bad idea, largely abandoned long ago

IUA Example: Instructional Management

- Measure “learning, not learners” with tests keyed to narrow learning objectives
  - Winnetka Plan
  - Programmed Instruction
  - Criterion-Referenced Testing (CRTs)
- Assumptions include…
  - Decomposability – *break complex tasks into small pieces*
  - Decontextualization – *learn skill in one context, apply in another*

IUA Example: Evaluation Research

- Measure learners to study treatment outcomes
  - ESEA Title I evaluations
  - Equality of Educational Opportunity report (Coleman, 1966)
  - Evaluations of NSF-funded curricula
    - (PSSC Physics, BSCS Biology, CHEM Study Chemistry)
  - What Works Clearinghouse
- Assumptions include…
  - Comparable test-curriculum alignment for all treatments
  - Research design supports causal claims
    - (e.g., random assignment)
  - Testing itself does not influence treatments

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

Direct versus indirect effects of testing

- Direct: Interpretation / action based on test scores
- Indirect: Influence of testing per se, not via scores
  - Incentive effects (spur efforts to raise scores)
  - Messaging effects (use testing to influence perceptions)

"This will be on the test!"

Look here for unintended consequences

Direct versus indirect effects of testing

"The arguments for these … programs tend to claim that the program will lead to improvements in school effectiveness and student achievement by focusing the attention of school administrators, teachers, and students on demanding content."

Michael T. Kane, 2006

*Instead of testing to **evaluate** educational treatments, testing **becomes** the treatment*

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

New forms of (derived) test scores

- State-level school accountability systems
  - e.g., Kentucky's KIRIS, California's API, NCLB
- Additional examples:
  - Teacher accountability using "Value-Added" models
  - Student proficiency levels (e.g., "Proficient")
  - Identification / Reclassification of "English Learners"
  - College-and-Career Readiness indicators

New forms of (derived) test scores

- Familiar derived scores based on a single raw score
  - Scale scores (make different forms comparable or maintain score scales from year to year)
  - Percentiles, grade equivalents (add meaning)

New forms of (derived) test scores

- Newer derived scores based on a single raw score
  - Below Basic / Basic / Proficient / Advanced (adds meaning)
  - College Readiness (adds meaning)
  - English Learner designation (adds meaning)
    - May use additional information

      along with English language test scores

New forms of (derived) test scores

- More complex derived scores
  - Growth scores (including Student Growth Percentiles)
    - derived from a series of student scores over time
  - Value-Added scores for teachers
    - derived from student-level test scores over time; may use additional information
  - School-level accountability scores
    - E.g., Adequate Yearly Progress (AYP) determinations under NCLB

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

AYP: A complicated, school-level derived score

- Schools either make “Adequate Yearly Progress” (AYP) or are “In Need of Improvement”
- Thinking of this as a derived score raises questions …
  - What construct does it measure?
  - How reliable is it?
  - What might be some sources of construct-irrelevant variance?
  - What might be some sources of construct under-representation?
  - …

AYP: A complicated, school-level derived score

- Binary school AYP score is built up layer by layer
  - Student-level achievement scores in reading and mathematics
  - Student achievement levels (Below Basic, …, Advanced)
  - School-level "Percent Proficient" scores (reading and math, whole school plus designated student subgroups)
  - Annual school-level AYP determination

AYP: A complicated, school-level derived score

- Validation for step 1: Scores → Achievement Levels
  - Questions about methods for determining cut scores
  - Surplus meaning of labels like “Proficient”
  - Confusion when “Proficient” is used …

    across tests

    across testing programs (e.g., state tests and NAEP)

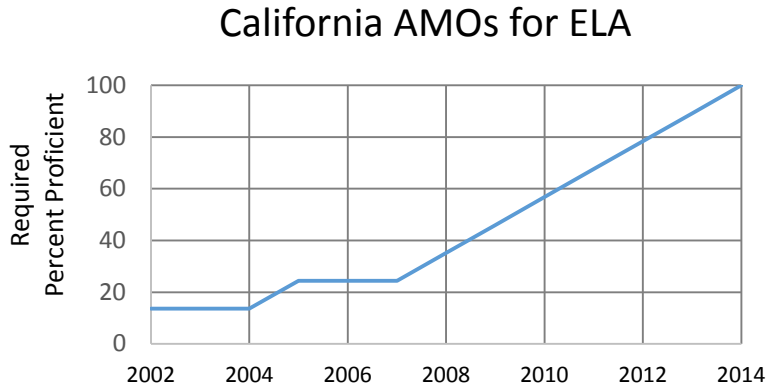
    across subject areas and grade levels

AYP: A complicated, school-level derived score

- Validation for Step 2: Student Achievement Levels → School-level "Percent Proficient" scores
  - Reliability affected by student group sizes as well as measurement error
  - Construct-Irrelevant Variance and Construct Under-Representation distort inferences as to school quality
  - "% Proficient" invites statistically faulty uses / interpretations

Annual Measureable Objectives (AMOs)

- Required to reach 100% Proficient by 2014
  - Expecting reauthorization in 2007, States went for "balloon mortgage" scenarios:



  - By 2014, instead of approaching 100% of students "Proficient," we were approaching 100% of schools "In Need of Improvement"

- AYP: A complicated, school-level derived score
  - Student scores → "proficient" determinations
  - → "percent proficient" for groups
  - → met/didn't meet AMO
    - For multiple groups (special rules for counting English Learners)
      - but only if numerically significant
    - Excluding 1%, 2% with severe disabilities
    - Except for "safe harbor"
      - as adjusted by "margin of error"
  - Conjunctive school-level decision rule
    - And don't forget 95% participation rate criteria …

AYP: A complicated, school-level derived score

Compared to AYP, IRT is simple!

AYP: A complicated, school-level derived score

- Validation for Step 3: AYP interpretations
  - Scoring
    - Score meaning is just the beginning
  - Generalization
    - Reliability
  - Extrapolation
    - What does it tell us about a school, beyond test scores?
  - Interpretation and Use
    - Did NCLB's test-based accountability regime advance its policy objectives?

Validating Policy Uses of Test Scores

- Validation: yesterday, today, and tomorrow
- Familiar IUAs (examples)
- Direct versus indirect effects of testing
- New forms of (derived) test scores
- AYP: A complicated, school-level derived score
- Expanding our view of validity research

Expanding our view of validity research

- Validity theory has evolved as the field has …
  - Better understood threats to validity
  - Wrestled with adverse impact and perceptions of test bias
  - Responded to new kinds of test uses and interpretations

Expanding our view of validity research

- Find and evaluate IUAs for new test uses
  - Describe constructs
  - Lay out justifications / mechanisms of action
    - **Both direct and indirect**
  - Identify key propositions requiring support
  - Bring to bear both empirical evidence and theoretical rationales

Expanding our view of validity research

- Actual versus Intended Score Uses and Interpretations
  - One prime example: NAEP Achievement Levels

"RECOMMENDATION … Research is needed to articulate the intended interpretations and uses of the achievement levels and to collect validity evidence to support these interpretations and uses. In addition, research is needed to identify the actual interpretations and uses commonly made by the National Assessment of Educational Progress's various audiences and to evaluate the validity of each of them. This information should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations." *NRC Evaluation, 2017*

Expanding our view of validity research

"In the final analysis, we suspect that this nation may be placing far too much weight on accountability to achieve its reform agenda." *Shavelson, et al., 1992*

"We should not shy away from critiquing policies and programs that are based on intuitive test theory. This involves telling lots of people (some of them very important) that what they want to do won't work and that doing something right is harder or takes longer than they might like." *Braun & Mislevy, 2005*

THANK YOU!

APA ANNUAL CONVENTION

August 3-6, 2017 | Washington, D.C.

[THE POWER OF]

APA 125
AMERICAN PSYCHOLOGICAL ASSOCIATION

http://people.stanford.edu/haertel