The meaning of validity: consensus, what consensus?

CAMBRIDGE ASSESSMENT

Paul E. Newton & Stuart D. Shaw

The Validity Symposia
February 29th, 2012

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

A rollercoaster ride to consensus

The primacy of validity has remained uncontested and irrefutable

"One of the major deities in the pantheon of the psychometrician"

Ebel, 1961, p.640

Overarching trajectory of evolution of validity characterised in terms of:

- differing, disparate and contested ideas
- range of inadequate formulations/ambiguous terminology/ confusing nomenclature
- neither a neat nor linear evolution
- early insights overlooked then rediscovered
- important insights partially formed reaching maturity over time

Shifting emphases

CAMBRIDGE ASSESSMENT

Trajectory of thought variously described as:

- a pervasive philosophical change (Angoff, 1988)
- an evolution (Shepard, 1993)
- a metamorphosis (Geisinger, 1992)

Move towards unified view of validity and developments in validity theory perceived in terms of shifts in emphasis:

“from numerous specific criterion validities to a small number of validity types and finally to a unitary validity conception …

“ from prediction to explanation as the fundamental validity focus”

(Messick, 1989, p.18)

“from a strictly positivistic position to a more post-positivistic orientation”

(Moss, Girard & Haniford, 2006)

“from a purely quantitative, positivistic approach to a conception of validity reliant on the interpretation of multiple evidence sources integrated into validity arguments.”

(De Luca, 2011, p.303)

Tracking the path to consensus

Standards for Educational and Psychological Testing (AERA, APA, & NCME) – 5 sets

1952 (prelim proposals); 1954/1955 (separate documents for psychological and achievement tests); 1966; 1974; 1985; 1999

Educational Measurement (NCME & American Council on Education) - 4 chapters

Cureton (Validity, 1951)
Cronbach (Test Validation, 1971)
Messick (Validity, 1989)
Kane (Validation, 2006)

Tracking the path to consensus

CAMBRIDGE ASSESSMENT

Other key publications:

- **Essentials of Psychological Testing** (Cronbach) – 5 eds.
  1949; 1960; 1970; 1984; 1990

- **Psychological Testing** (Anastasi) – 7 eds.
  1954; 1961; 1968; 1976; 1982; 1989; 1997 (Anastasi and Urbina)

Papers:

- MacCorquodale & Meehl (1948)
- Cronbach & Meehl (1955)
- Loevinger (1957)
- Campbell & Fiske (1959)
- Angoff (1988)
- Messick (1975; 1995; 1998)
- Kane (2004)

Educational Measurement

All four authors focus on intended interpretations and uses of test scores

- the interpretations and uses that are presumptively valid across individuals and contexts.

General assumption: **validity is a property of interpretations and uses** and not a property of the test

Validity is defined in terms of:

- intended uses (Cureton, 1951)
- interpretations and a range of possible uses (Cronbach, 1971)
- value implications and social consequences of testing outcomes (Messick, 1989):
- the tradition of interpretations and uses of test scores (Kane, 2006)

Standards for Educational and Psychological Testing

CAMBRIDGE ASSESSMENT

1954	Four Types of Validity: **content, predictive, concurrent, construct**
1955	Four Types of Validity: **content, concurrent, predictive, construct**
1966	Three aspects of validity corresponding to 3 aims of testing: **content, criterion-related, construct**
1974	Kinds of validity depend upon kinds of test score inferences: **content; criterion-related; construct**
1985	Means of accumulating validity evidence grouped into categories: **content-related, criterion-related, construct-related**. ("These categories are convenient - but use of category labels does not imply that there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use", p.9)
1999	Five sources of validity evidence based on: **test content; response processes; internal structure; relations to other variables; consequences of testing**

Points of consensus

CAMBRIDGE ASSESSMENT

- validity is not an inherent property of a test but refers to the specified uses of a test for a particular purpose (Sireci, 2007; 2009)
- validity pertains to the intended inferences or interpretations made from test scores (Cronbach, 1971; Messick, 1989; Kane,2006)
- it is the interpretations and uses of test scores that are validated, and not the tests themselves (Cronbach & Meehl, 1955; Cronbach, 1971; Kane, 2009)
- notion of discrete kinds of validity has been supplanted by the unified view of validity (Loevinger, 1957; Messick, 1989)
- all validity is construct validity (Messick, 1975; 1988; 1989)

Points of consensus

CAMBRIDGE ASSESSMENT

- validity is not expressed as a presence or absence of that characteristic but is a matter of degree

  (Cronbach, 1971; Messick, 1989; Zumbo, 2007; Kane, 2001)

- validation is neither static nor a one-time event but an on-going process that relies on multiple evidence sources

  (Shepard ,1993; Messick, 1989; Sireci, 2007)

- evidence needed for validation depends on claims made about test and proposed interpretations and uses
  - different interpretations/uses will require different kinds and different amounts of evidence for their validation (Kane, 2006, 2009)

- recognition that validation processes and validity evidence are value-laden (Messick, 1989)

Explicit criticism of consensus

- explicit refutation of notion that validity is a property of interpretive inferences and assertion that validity is a property of the test Borsboom et al (2004; 2009)

  "absurdities of construct validity theory" (2009, p.135)

  "almost everybody except construct validity theorists" believe that validity is a property of tests rather than interpretations (2009, p.138)

- proposals of new nomenclature to characterize validity and validation (Lissitz & Samuelson, 2007)

  adopt a narrower, operational definition of validity as an evaluation of internal characteristics of test - shifting concerns about external relationships/test-score uses to separate category

- ongoing debate over the role and importance of consequences in validity theory (Crocker, 1997; Maguire et al. 1994; Mehrens, 1997; Popham, 1997; Shepard, 1997; Moss, 1998; Cizek, 2011)

Explicit criticism of consensus

- well-established disjunction between modern validity theory and modern validity practice (Jonson & Plake, 1998; Hogan & Agnello, 2004; Cizek et al, 2008; Wolming & Wikstrom, 2010)

  "We have elevated the concept of construct validation to so high a level that it seems an 'out of reach' goal" (Fremer, 2000, p.1)

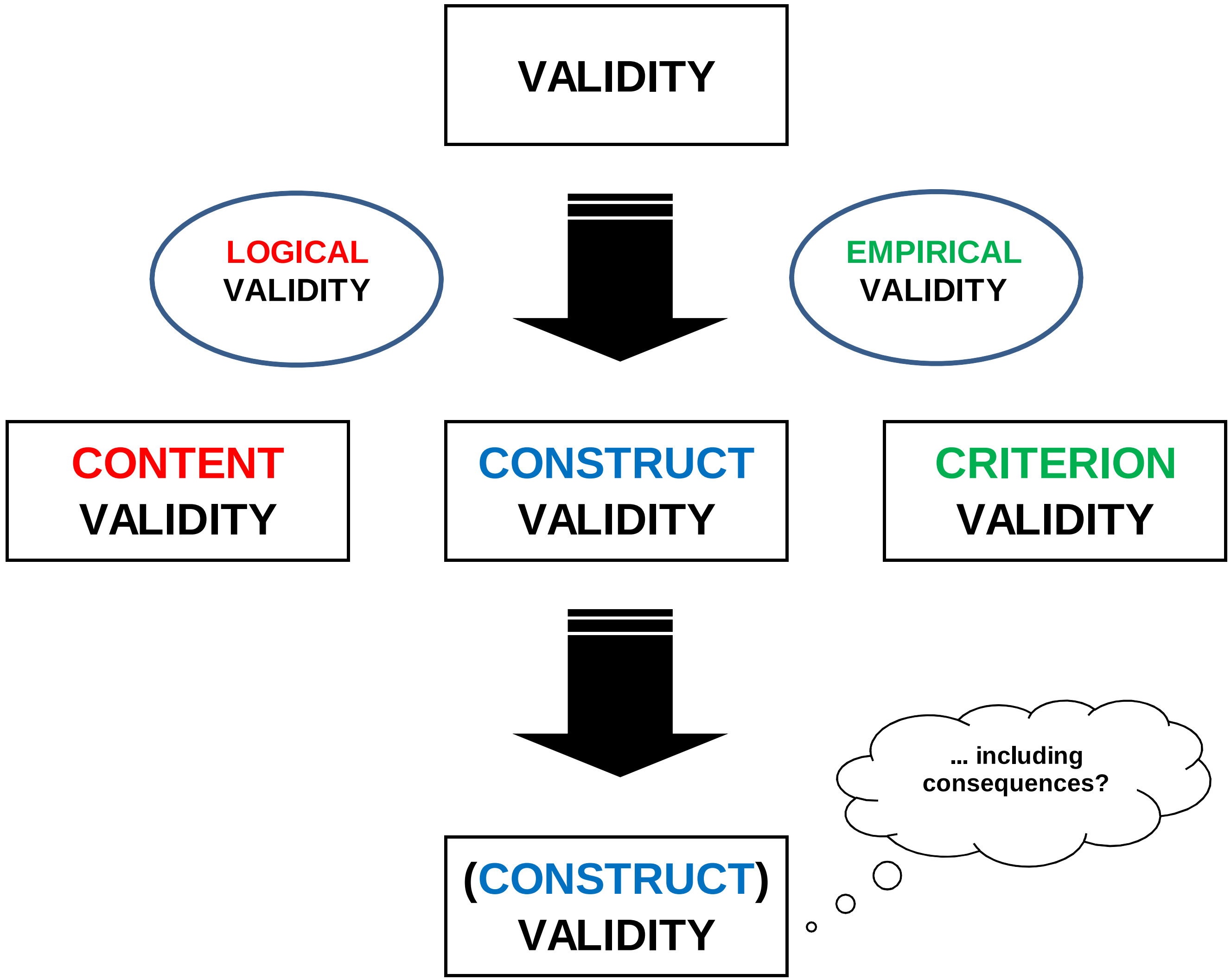
  "Sadly, in my opinion, the unified field of validity as articulated by Messick (1989) is not just historically unsettled at a theoretical level, but it offers little in the way of usable advice to people working on test construction efforts in the field." (Lissitz, 2009, p.5)

- explicit revisions of the construct validity concept (Zumbo, 2009)

- Messick's framework does not offer practical guidelines as to how to arrive at conclusions during a validation study (Brennan, 1998; Crocker, 2003; Kane, 1992, 2004, 2006)

Evolution in a nutshell

CAMBRIDGE ASSESSMENT



PHASE 1

Validity of the test

“By *validity* is meant *the degree to which a test or examination measures what it purports to measure*.”

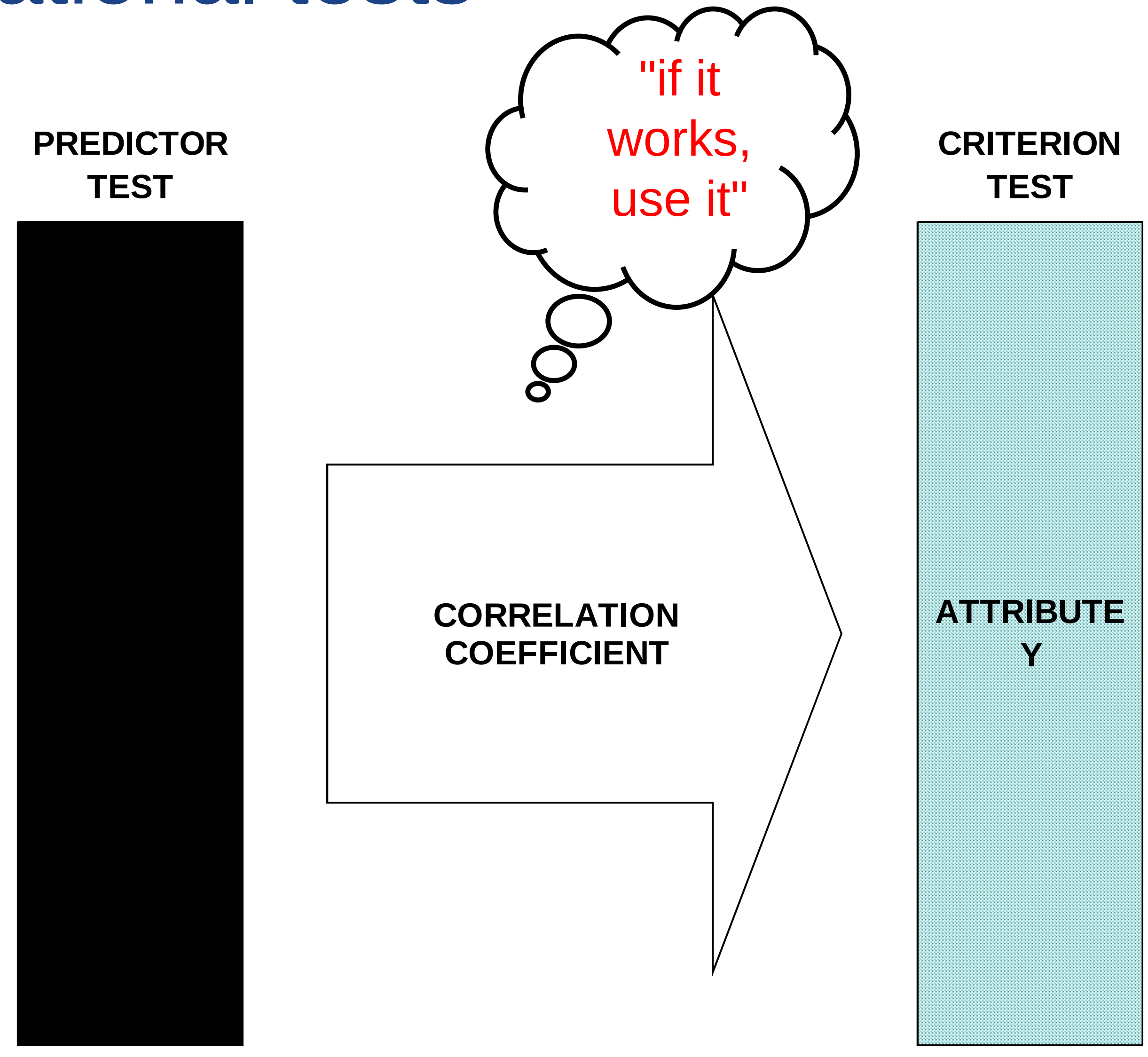
(Ruch, 1924, p.13)

PHASE 2

Content validity, especially for educational tests

- Does the content of the test match the content of the curriculum?

Criterion validity, especially for occupational tests



PHASE 3

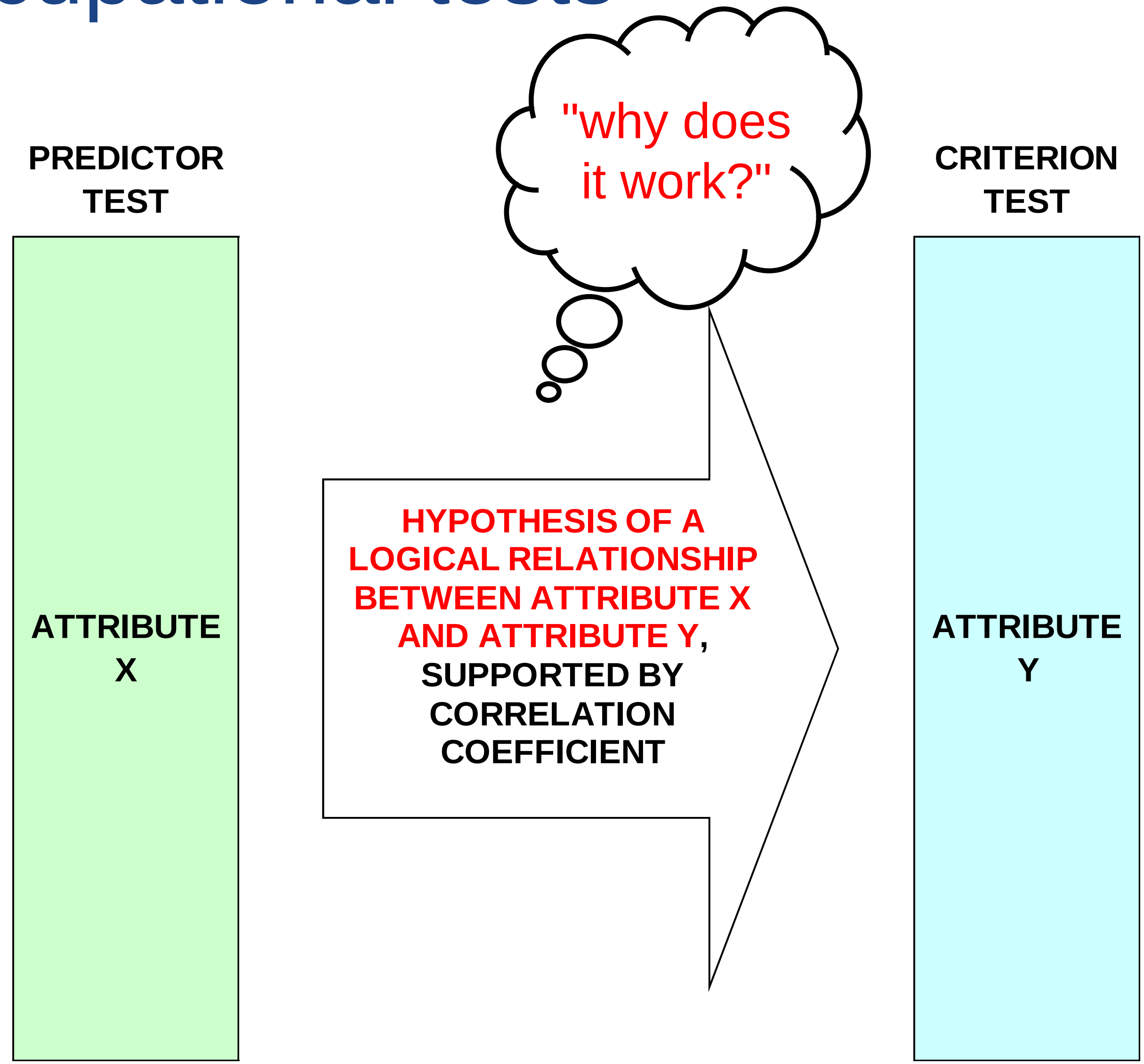
New-style construct validity, for educational tests

CAMBRIDGE ASSESSMENT

- Does the content of the test match the content of the curriculum?
- Do students answer questions in the manner intended?
- Do examiners mark answers in the manner intended?
- Do students' answers to different questions cohere in the way we would expect them to?
- Do users interpret results in the manner intended?

New-style construct validity, for occupational tests

CAMBRIDGE ASSESSMENT



It’s *all* about score meaning... it’s *all* about construct validity

CAMBRIDGE ASSESSMENT

“[...] construct validity may ultimately be taken as the whole of validity in the final analysis.”

(Messick, 1989, p.21)

PHASE 4?

(Construct) validity of the interpretation?

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests."

(AERA/APA/NCME, 1999, p.9)

(Construct) validity of the use?

CAMBRIDGE ASSESSMENT

"To claim that a proposed interpretation or use is valid is to claim that the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible."

(Kane, 2006, p.23)

Other (construct) validity referents

- validity of the hypothesis of a relationship between predictor and criterion (Guion, 1978)
- validity of the scores produced by a test (Thorndike & Thorndike-Christ, 2010, p.154)
- validity of the validation evidence (Ellis & Blustein, 1991)
- validity is the explanation for the test score variation (Zumbo, 2009)

(Construct) validity of the argument?

CAMBRIDGE ASSESSMENT

"All of the inferences and assumptions must be sound if the interpretive argument is to be considered valid [...]."

(Clauser, Kane & Swanson, 2002, p.419)

(Construct) validity of anything?

Miller et al (2009), on a single page (p.104), refer to:

- “the validity of an assessment”
- “the validity of the assessment for that use or interpretation”
- “the validity of interpretations of tests and assessments”
- “the validity of test and assessment results”
- “the validity of the uses and interpretations”

Or just the validity of the test?

CAMBRIDGE ASSESSMENT

"[...] what really matters in validity is how the test works, and this is certainly not a property of test score interpretations, or even of test scores, but of the measurement instrument itself [...] the relevant capacity to pick up variation in the targeted attribute"

(Borsboom et al, 2009, p.149)

Validity of the test

“By *validity* is meant *the degree to which a test or examination measures what it purports to measure*.”

(Ruch, 1924, p.13)

ANY QUESTIONS?

When assessment professionals use the term ‘valid’, do they mean the same as:

– a social scientist would?

– a philosopher would?

– a lawyer would?

– an average Jo would?

i.e. does (or should) the concept of validity have a meaning specific to the assessment professions?

When an assessment professional claims that something is valid: (c) does that claim require quantitative elaboration?

– a tiny little bit valid, nearly valid, extremely valid?

i.e. is validity a matter of degree, or is it an either/or kind of thing?

3

When an assessment professional claims that something is valid or invalid: (a) what kind of 'thing' can that claim legitimately refer to?

- an item, a test, a response, a score, an interpretation of a score, an entire assessment procedure, the use of an assessment procedure?

i.e. is the test valid (a la Borsboom) or is the interpretation valid (a la construct validity theory)?

When an assessment professional claims that something is valid or invalid: (b) what should count as that ‘thing’ being valid or invalid

– an item, a test, a response, a score, an interpretation of a score, an entire assessment procedure, the use of an assessment procedure?

i.e. what is the basic criterion for judging whether or not a ‘thing’ is valid?

When assessment professionals use the term ‘valid’, should they all mean the same?

– or is it OK for different assessment professionals to mean different things?

i.e. should all assessment professionals sign-up to the same concept of validity?