Leading education and social research
Institute of Education
University of London

Does it matter what 'validity' means?

Professor Paul E. Newton

Date: 4 February 2013
Seminar: University of Oxford, Department of Education

www.ioe.ac.uk

The most elusive of all assessment concepts?

Leading education and social research
Institute of Education
University of London

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment."

Samuel Messick (1989)

The most fundamental of all assessment concepts?

Leading education and social research
Institute of Education
University of London

"validity [...] is the single most important criterion for evaluating achievement testing. The importance of validity is widely enough recognized that it finds its way into laws and regulations." (p.215)

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Leading education and social research
Institute of Education
University of London

MEANINGS OF VALIDITY

(ongoing mission: to explore strange new literatures, to seek out new validity forms)

Validity specific to fields beyond education (EPM)

Leading education and social research
Institute of Education
University of London

Law (e.g. Waluchow, 2009)

- Legal validity (as existence)
- Systemic validity
- Systemic moral validity
- Moral validity

Management (e.g. Markus & Robey, 1980)

- Organizational validity
- Technical validity

Validity for quantitative research conclusions

Leading education and social research
Institute of Education
University of London

Campbell (1957)

- Internal validity
- External validity

Bracht & Glass (1968)

- Population validity (ext.)
- Ecological validity (ext.)

Wolf (1978)

- Social validity

Cook & Campbell (1979)

- Statistical conclusion validity (int.)
- Internal validity (int.)
- Construct validity (ext.)
- External validity (ext.)

Validity for qualitative research conclusions

Leading education and social research
Institute of Education
University of London

Maxwell (1992)

- Descriptive validity
- Interpretive validity
- Theoretical validity
- Evaluative validity

Kvale (1994)

- Communicative validity
- Pragmatic validity

Cho & Trent (2006)

- Transactional validity
- Transformational validity

Lather (1986)

- Construct validity
- Face validity
- Catalytic validity

Lather (1993)

- Transgressive validity
- Ironic validity
- Paralogical validity
- Rhizomatic validity
- Voluptuous validity

Validity for educational and psychological measurement

Leading education and social research
Institute of Education
University of London

Abstract validity	Criteria validity	External test validity	Judgmental validity	Response validity
Administrative validity	Criterion validity	External validity	Known-groups validity	Retrospective validity
Aetiological validity	Criterion-oriented validity	Extratest validity	Linguistic validity	Sampling validity
Artifactual validity	Criterion-related validity	Face validity	Local validity	Scientific validity
Behavior domain validity	Criterion-relevant validity	Factorial validity	Logical validity	Scoring validity
Cash validity	Cross-age validity	Faith validity	Longitudinal validity	Self-defining validity
Circumstantial validity	Cross-cultural validity	Fiat validity	Lower-order validity	Semantic validity
Cluster domain validity	Cross-sectional validity	Forecast true validity	Manifest validity	Single-group validity
Cognitive validity	Cultural validity	Formative validity	Natural validity	Site validity
Common sense validity	Curricular validity	Functional validity	Nomological validity	Situational validity
Concept validity	Decision validity	General validity	Occupational validity	Specific validity
Conceptual validity	Definitional validity	Generalized validity	Operational validity	Statistical validity
Concrete validity	Derived validity	Generic validity	Particular validity	Status validity
Concurrent criterion validity	Descriptive validity	Higher-order validity	Performance validity	Structural validity
Concurrent criterion-related validity	Design validity	In situ validity	Postdictive validity	Substantive validity
Concurrent true validity	Diagnostic validity	Incremental validity	Practical validity	Summative validity
Concurrent validity	Differential validity	Indirect validity	Predictive criterion validity	Symptom validity
Congruent validity	Direct validity	Inferential validity	Predictive validity	Synthetic validity
Consensual validity	Discriminant validity	Instructional validity	Predictor validity	System validity
Consequential validity	Discriminative validity	Internal test validity	Prima Facie validity	Systemic validity
Construct validity	Divergent validity	Internal validity	Procedural validity	Theoretical validity
Constructor validity	Domain validity	Interpretative validity	Prospective validity	Theory-based validity
Construct-related validity	Domain-selection validity	Interpretive validity	Psychological & logical validity	Trait validity
Content sampling validity	Edumetric validity	Intervention validity	Psychometric validity	Translation validity
Content validity	Elaborative validity	Intrinsic content validity	Quantitative face validity	Translational validity
Content-related validity	Elemental validity	Intrinsic correlational validity	Rational validity	Treatment validity
Context validity	Empirical validity	Intrinsic rational validity	Raw validity	True validity
Contextual validity	Empirical-judgemental validity	Intrinsic validity	Relational validity	User validity
Convergent validity	Essential validity	Job analytic validity	Relevant validity	Washback validity
Correlational validity	Etiological validity	Job component validity	Representational validity	

Leading education and social research
Institute of Education
University of London

THE CONSENSUS DEFINITION OF VALIDITY

(and its evolution)

The first consensus definition of validity

Leading education and social research
Institute of Education
University of London

"Two of the most important types of problems in measurement are those connected with the determination of what a test measures, and of how consistently it measures. The first should be called the problem of validity, the second, the problem of reliability." (p.80)

Buckingham, B.R., McCall, W.A., Otis, A.S., Rugg, H.O., Trabue, M.R. & Courtis, S.A. (1921). Report of the Standardization Committee. *Journal of Educational Research*, 4 (1), 78-80.

"By *validity* is meant *the degree to which a test or examination measures what it purports to measure*." (p.13)

Ruch, G.M. (1924). *The improvement of the written examination.* Chicago: Scott, Foreman and Company.

The second consensus definition of validity

Leading education and social research
Institute of Education
University of London

Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, AERA, NCMUE, 1954)

1. dissemination
2. interpretation
3. validity
   - introductory section (pp.13-18)
   - 19 validity standards (pp.18-28)
4. reliability
5. administration and scoring
6. scales and norms.

Standards # 1 (1954)

Leading education and social research
Institute of Education
University of London

"When validity is reported, the manual should indicate clearly what type of validity is referred to." (pp.18-19)

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (2), Supplement.

Standards # 1 (1954)

Leading education and social research
Institute of Education
University of London

1. **Content validity**
2. **Concurrent validity**
3. **Predictive validity**
4. **Construct validity**

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (2), Supplement.

Standards # 2 (1966)

Leading education and social research
Institute of Education
University of London

1. **Content validity** (e.g. achievement tests)
2. **Criterion-related validity** (e.g. aptitude tests)
3. **Construct validity** (e.g. personality tests)

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington, D.C.: American Psychological Association.

Standards # 4 (1985)

Leading education and social research
Institute of Education
University of London

1. **Content-related evidence**
2. **Criterion-related evidence**
3. **Construct-related evidence**

... i.e. it was now officially incorrect to think of validity as a specialised, fragmented concept (following Messick, Guion, and others)**.**

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Psychological Association.

Standards # 5 (1999)

Leading education and social research
Institute of Education
University of London

"In the current standards, all test scores are viewed as measures of some construct [...] The validity argument establishes the construct validity of a test." (p.174)

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Standards # 5 (1999)

Leading education and social research
Institute of Education
University of London

1. **Evidence based on test content**
2. **Evidence based on response processes**
3. **Evidence based on internal structure**
4. **Evidence based on relations to other variables**
5. **Evidence based on consequences of testing**

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Leading education and social research
Institute of Education
University of London

A FRAGILE CONSENSUS

Occasional rejection of the consensus position

Leading education and social research
Institute of Education
University of London

Cattell (1964)

- Concrete validity ⟺ Concept validity
- Natural validity ⟺ Artifactual validity
- Direct validity ⟺ Indirect validity

Gradual appropriation of the consensus position

Leading education and social research
Institute of Education
University of London

APA, AERA, NCME The *Standards*		**1954** Content Predictive Concurrent Construct		**1966** Content Criterion-related Construct	
Cronbach *Essentials of Psychological Testing*	**1949** Logical Empirical Factorial		**1960** Content Predictive Concurrent Construct		**1970** Content (valdn.) Criterion-oriented (valdn.) Construct (valdn.)
Anastasi *Psychological Testing*		**1954** Face Content Factorial Empirical	**1961** Content Predictive Concurrent Construct		**1968** Content Criterion-related Construct
Thorndike & Hagen *Measurement and evaluation in psychology and education*		**1955** Content Predictive Concurrent Congruent Concept (Construct)	**1961** Rational (Logical, Content) Empirical (Statistical) Construct		**1969** Content Criterion-related Construct

Growth of 'black market' in types of validity (pre-1966)

Leading education and social research
Institute of Education
University of London

Loevinger (1957)

- internal validity
- substantive validity
- structural validity
- external validity

Tryon (1957)

- domain validity

Campbell and Fiske (1959)

- convergent validity
- discriminant validity

Campbell (1960)

- trait validity
- nomological validity

Shaw and Linden (1964)

- common sense validity

Cureton (1965)

- raw validity
- true validity
- intrinsic validity

Growth of 'black market' in types of validity (post-1966)

Leading education and social research
Institute of Education
University of London

Lord and Novick (1968)
- **empirical validity**
- **theoretical validity**

Bemis (1968)
- **occupational validity**

Dick and Hagerty (1971)
- **cash validity**

Boehm (1972)
- **single-group validity**

Carver (1974)
- **psychometric validity**
- **edumetric validity**

Popham (1978)
- **descriptive validity**
- **functional validity**
- **domain-selection validity**

Hambleton (1980)
- **decision validity**

Ebel (1983)
- **intrinsic rational validity**
- **performance validity**

Official change in the consensus position

Leading education and social research
Institute of Education
University of London

1985 *Standards* (4th edition)

- validity modifier labels officially dropped
- now just 'sources of evidence' of validity

1999 *Standards* (5th edition)

"These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. [...] To emphasize this [...] the treatment that follows does not follow traditional nomenclature (i.e., the use of the terms *content validity* or *predictive validity*)." (p.11)

The ‘black market’ still continues to trade in types

Leading education and social research
Institute of Education
University of London

Prevalence study

- **analysed (only the) titles of articles**
- **from 22 EPM journals**
- **published between 01/01/05 and 31/12/10**

- **how frequently was the ‘X validity’ formulation observed?**

Validity Modifier Label	Freq.
Construct validity	61
Incremental validity	27
Predictive validity	22
Convergent validity	17
Discriminant validity	14
Criterion-related validity	12
Concurrent validity	9
Criterion validity	9
Factorial validity	8
Construct-related validity	3
Structural validity	3
Content validity	2
Consequential validity	2
Differential validity	1
Internal validity	1
Cross-cultural validity	1
Cross-validity	1
External validity	1
Population validity	1
Consensual validity	1
Diagnostic validity	1
Extratest validity	1
Incremental criterion-related validity	1
Operational validity	1
Local validity	1
Concurrent criterion-related validity	1
Criteria validity	1
Cross-age validity	1
Elemental validity	1
Predictive criterion-related validity	1
Synthetic validity	1
Treatment validity	1

In fact, the 'black market' still continues to grow...

Leading education and social research
Institute of Education
University of London

Tenopyr (1986)
- general validity
- specific validity

Foster & Cone (1995)
- representational validity
- elaborative validity

Jolliffee et al. (2003)
- prospective validity
- retrospective validity

Allen (2004)
- formative validity
- summative validity

Freebody & Wyatt-Smith (2004)
- site-validity
- system-validity

Briggs (2004)
- design validity
- interpretive validity

Willcutta & Carlson (2005)
- diagnostic validity

Trochim (2006)
- translation validity

… and grow

Leading education and social research
Institute of Education
University of London

Hill et al. (2007)
- **structural validity**
- **elemental validity**

Shaw &Weir (2007)
- **cognitive validity**
- **context validity**
- **scoring validity**

Larsen et al. (2008)
- **manifest validity**
- **semantic validity**

Lievens et al. (2008)
- **operational validity**

Hopwood et al. (2008)
- **extratest validity**

Brookhart (2009)
- **decision validity**

Karelitz et al. (2010)
- **cross-age validity**

Evers et al. (2010)
- **retrospective validity**

Guion (2011)
- **generic validity**
- **psychometric validity**
- **relational validity**

Leading education and social research
Institute of Education
University of London

THE BREAKDOWN OF CONSENSUS

(and ambiguity in the consensus definition)

Samuel Messick

Leading education and social research
Institute of Education
University of London

Two major evaluation questions (Messick, 1965)

Scientific question (technical accuracy)

- **Is the test any good as a measure of the characteristic it purports to assess?**

Ethical question (social value)

- **Should the test be used for its present purpose?**

Early Messick: validation is (ultimately) policy analysis

Leading education and social research
Institute of Education
University of London

Matrix represents TEST VALIDITY (essentially a political judgement)

cf. Construct Validity in Cell 1 (essentially a scientific judgement)

	Test Score Interpretation	Test Score Use
Scientific (technical) Evaluation	1 **Evaluation of measurement**	3 **Evaluation of decision-making**
Ethical (social) Evaluation	2 **Evaluation of social values underlying TBDMP**	4 **Evaluation of social consequences of TBDMP**

TBDMP = Test-Based Decision-Making Procedure

Did Messick regret having created a monster?

Leading education and social research
Institute of Education
University of London

- **Performance assessment has good consequences**
- **Good consequences mean high (consequential) validity**
- **Therefore, performance assessment has high validity**

Late Messick: validity is (basically) a scientific concept

Leading education and social research
Institute of Education
University of London

Matrix represents CONSTRUCT VALIDITY (essentially a scientific judgement)

	Test Score Interpretation	Test Score Use
Scientific (technical)	1 **Evaluation of measurement**	**Implications of decisions for 1**
	Implications of values for 1	**Implications of consequences for 1**

Standards # 5 (1999)

Leading education and social research
Institute of Education
University of London

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p.9)

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Varieties of meaning now associated with 'validity'

Leading education and social research
Institute of Education
University of London

1

	Measurement	Decisions	Impacts
Scientific (technical) Evaluation	Borsboom (?) Cizek (?) Scriven (?)		
Ethical (social) Evaluation			

2

	Measurement	Decisions	Impacts
Scientific (technical) Evaluation	Later Samuel Messick (?) 1999 *Standards,* narrow (?)		
Ethical (social) Evaluation			

3

	Measurement	Decisions	Impacts
Scientific (technical) Evaluation	1999 *Standards,* broad (?)		
Ethical (social) Evaluation			

4

	Measurement	Decisions	Impacts
Scientific (technical) Evaluation	Earlier Samuel Messick (?) Later Lee Cronbach (?) Later Mike Kane (?)		
Ethical (social) Evaluation			

Should we define validity as both scientific *and* ethical?

Leading education and social research
Institute of Education
University of London

If we reject ethical dimensions from validity, then ethical evaluation may fall by the wayside.

If we embrace ethical dimensions within validity, then validity theory may become too large and too complex to structure effective validation practice.

- Who should be responsible for which aspects of validation, including the overall judgement of validity?
- Is an overall judgement of validity even meaningful when, for example, good tests are used badly?
- How feasible would it be to conduct a thorough validation programme, as the basis for any claim to validity?

There is no consensus over the meaning of 'validity'

Leading education and social research
Institute of Education
University of London

Leading theorists disagree radically over its scope:

- measurement vs. measurement + decision-making vs. overall policy
- scientific vs. scientific *and* ethical

The most recent edition of the *Standards* is quite ambiguous:

- measurement vs. overall policy (if only from a technical perspective)

The *Standards* only ever sustained a fragile consensus, anyhow:

- proliferation of kinds of validity pre-1985 (cf. only 3 kinds officially)
- proliferation of kinds of validity post-1985 (cf. only 1 kind officially)

Does it matter what 'validity' means?

Leading education and social research
Institute of Education
University of London

- **If we want to use the term to communicate effectively, then yes.**
- **If there is no consensus over the meaning of validity (whether by formal definition or by the way it is used) then effective communication is not possible.**
- **It matters especially if "The importance of validity is widely enough recognized that it finds its way into laws and regulations."** (Koretz, 2008, p.215)

Leading education and social research
Institute of Education
University of London

HAS THE TERM ‘VALIDITY’ OUTLIVED ITS USEFULNESS?

Could we ditch the term ‘validity’?

Leading education and social research
Institute of Education
University of London

Ridiculous idea

- **it’s been our watchword for a century**
- **that which we call a rose by any other name would smell as sweet**

Possible reasons to ditch the term 'validity'

Leading education and social research
Institute of Education
University of London

- **More disagreement over how to apply the term 'validity' than over how to characterise quality in EPM.**
- **The term 'validity' has become too big for specialists to understand and, therefore, too big to be useful.**
- **Genuine difference of opinion over how to characterise quality in EPM is being obscured by debate over how to apply the term 'validity'.**

What if we stopped talking about:

1. **validity**... and thought more about **quality**?
2. **validation**... and thought more about **evaluation**?

Back to the drawing board, having ditched 'validity'

Leading education and social research
Institute of Education
University of London

	Focus for Evaluation (What needs to be investigated in order to evaluate the policy)		
	Measurement	**Decisions**	**Impacts**
Scientific (technical) Evaluation	what does 'quality' mean here?	what does 'quality' mean here?	what does 'quality' mean here?
Ethical (social) Evaluation	what does 'quality' mean here?	what does 'quality' mean here?	what does 'quality' mean here?
Legal Evaluation	what does 'quality' mean here?		

Neo-(Early)-Messickian matrix for policy analysis

Leading education and social research
Institute of Education
University of London

	Focus for Evaluation (What needs to be investigated in order to evaluate the policy)		
	Measurement	**Decisions**	**Impacts**
Scientific (technical) Evaluation	Potential of measurement procedure to support accurate measurement of attribute (defined by its construct)	Potential of measurement-based decision-making procedure to support accurate decisions	Potential of measurement-based decision-making policy to achieve other desired impacts
Ethical (social) Evaluation	Potential of construct to scaffold shared meaning within a wider community ('street credibility')	Likelihood that benefits accrued from accurate decisions will be judged to outweigh costs from inaccurate ones	Likelihood that benefits accrued from all non-decision-related impacts will be judged to outweigh their costs
Legal Evaluation	Potential to implement the measurement-based decision-making policy without infringing the law.		

Leading education and social research
Institute of Education
University of London

Validity is dead.
Long live evaluation.