CAVEWAS / CARP 2003

Annual conference

Vancouver, BC, Canada

Why am I here today?

- To discuss validity of sources of information used in vocational evaluation…
- To present the current consensus in the inter-discipline of measurement regarding validity.
- To give my personal perspective on use of standardized tests in Vocational assessments.

Is vocational evaluation something which can be done perfectly?

- No: A Voc. Eval. is the process of arriving at a professional opinion which will be informed by various **sources of information**.
- Does this mean that once you become a qualified professional your opinion will be valid no matter how you arrive at it?
- No: you are typically retained to provide an opinion that is grounded in empirical evidence.

A scientific research perspective

- Voc. Eval. is an investigative method with a particular epistemology (way of knowing something).
  - Voc. Eval. is perhaps best viewed as a form of qualitative research – which has as its focus the investigation of vocational capacities, abilities and interests of a single subject.

Does Voc. Eval. Rely on Empirical data? Are findings Evidence-based?

- Absolutely Yes (to both questions)
  - Assessments ought to be scientifically crafted.
  - No evidence ⇨ no claim to authority.
- Your opinion itself is not evidence!!
- Evidence is what you use to form an opinion.

What is this evidence?

- How should we view the kind of empirical method and social science "logic of inquiry" that is used in Voc. Eval.?
- Empirical evidence consists of multiple data sources – converging to say more or less the same thing about the examinee.

The Claim to authority of Qualitative methods.

- Interestingly, it is anthropologists who have perhaps written most about the class of empirical method which is used in Voc. Eval.
- Ethnographers study populations and cultures. Though we study individuals, the claim to authority which we make for our finding rests on the same logic of inquiry as that of ethnography.
- Interestingly social workers have used this method for longer than anyone, but current formal discussion of the empirical methods of case study perhaps begins with cultural anthropologists.

Massive overdetermination of pattern (MOP).

- Valid data from all sources can be combined in an empirical method which relies on MOP (Michael Agar, 2nd ed 1996)
- Another good book related to Case-study, is provided by Robert Yin (3rd ed., 2003).
- Do we need to use the same technique every time in our voc. Evals.? Is consistency good?
- Our job is to provide the most meaningful and useful vocational information about an individual – so perhaps not.

Validity of the overall evaluation comes from meaningfulness and usefulness of conclusions.

- By seeking convergence (pattern), we may converge upon a meaningful & useful finding.
- But it is the meaningfulness and usefulness of the finding that determines its validity (not merely that it converges with other data).

What about contradiction in the data?

- If there are no contradictory data no matter how widely you sample, then great!
  - Your finding can only support what was already known.
- But contradictory data are even better!!
  - Outliers in the qualitative data may indicate the presence of “rich points”.
  - Rich points lead the investigator to learn something new.
  - This learning represents real value in terms of acquisition of new information.

Validity and Standardized Test data

- Validity of standardized test data that are used to inform Voc. Eval.

Validity & Reliability of test data used in Voc. Eval:

- Tests are just one source of information.
  - But all sources of information can be evaluated as to their validity
- We will now focus on the special case of validity of test scores (which has been studied extensively)
- However, bear in mind that All sources of information must contribute positively to valid inferences being made in your evaluations.
- Measurement Validity (relating to standardized tests), is considered to be a property of test-scores.

Test scores are sources of information.

- Tests are used to take measurements.
- Tests do not have validity!! Scores and inferences based on test-scores have validity.
- "Measurement validity" refers to the
  - Appropriateness
  - Meaningfulness &
  - Usefulness

  … (AMU) … of the specific inferences which will be made from test scores.

Good summary of current consensus on validity

- Canadian Psychological Association, (1996). ***Guidelines for Educational and Psychological Testing.*** Ottawa, Ontario: Author. Retrieved from the internet April 7th, 2003 http://www.cpa.ca/guide9.html
- In-depth information is contained in the Standards (Jointly authored by AERA, NCME & APA).

Inferences & Scores have validity: Not tests

- Recall that we speak of validity as a unitary construct.
- There are no different types of validity, but merely different *facets* or aspects of the same ***unitary*** construct.

If validity is multi-faceted how do we assess validity?

- Modern consideration of validity always concerns itself with threats to validity.
  - Such things as calculation of validity coefficients (as is done in a multi-trait multi-method index for example), can be viewed as numerical methods for estimating threats to validity from one or more sources.

Reliability is one facet of validity

- Reliability may be defined as ***the extent to which tests scores are free from measurement error.***
- Lack of reliability is a threat to validity
  - Low reliability => score-data are comprised significantly of measurement error.
- We know by definition that error variance is variance that is uncorrelated with our construct of interest (provided the assumptions of our measurement model hold true).

Score reliability

- Each test score (X_i) is comprised of some element of trait or true score (T), and some component of measurement error (E).

$$X_i = T + E$$

$$E = X_i - T$$

Reliability is merely one aspect of validity!

- It is clearly an important aspect (but not the only aspect of validity that is important).
- We may note that reliability is a necessary but not sufficient condition for validity.
  - Without reliability there can be no validity.
  - But adequate reliability does not imply that sound inferences could be made from test scores.

Messick's characterization

- Messick and others have proposed that many facets of measurement validity relate to construct validity.
- Formerly:

1. construct validity
2. content validity
3. criterion-related validity

… were often considered separately

- Now content and criterion (as well as what was formerly considered construct validity) are viewed as aspects of construct validity.

Messick's perspective

- Samuel Messick has for decades been one of the most influential writers on measurement validity.
- In addition to urging us to consider that validity is a unitary construct, Messick urges us to consider that inferences based on measurements have CONSEQUENCES for people.
- We cannot therefore ignore consequential aspects of measurement.

What are some of the possible consequences of Voc. Eval. (and test scores obtained during Voc. Eval.)?

- Individual consequences
- Social consequences

Validity as we now define it, implies that we cannot ignore:

1. Morality & ethics
2. The purpose of Vocational Evaluation
    1. To provide meaningful evaluative information which may make a positive difference in an individual's life.
    2. …

Messick's 2 x 2 matrix

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance and Utility
Consequential Basis	Value Implications	Social Consequences

Messick

- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), ***Test validity*** (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), ***Educational measurement (3rd ed.)*** (pp. 13-103). New York: Macmillan.

Recap.

- A sophisticated view of measurement validity goes beyond the notion that test validity is the extent to which a test measures what it is supposed to measure.
- Instead … look at AMU of the inferences we may make from test scores.

Appendix: Classical view of validity

- Recall that we said reliability can be measured by the correlation between parallel forms.
- Trait (T) is reflected in a portion of a test score such that $X = T + E$, or $T = X - E$
- Now suppose that we know the reliability of two measures and we want to know the extent to which the two measures assess the same or similar constructs?

Multi-trait / multi-method

- Early measurement specialists noted that score distributions on different measures could be compared in a correlation matrix.
- These measures could differ in terms of method of assessment used, and in terms of the construct they purported to measure.
  - Thus the term multi-trait/multi-method matrix.

Multi-trait multi-method matrix

	Grip test	Verbal test	Numeric test	Dexterity test
Grip test	1.00			
Verbal test	0.0	1.00		
Numeric test	0.0	.8	1.00	
Dexterity test	.6	.2	.1	1.00

Multi-trait multi-method matrix

Reliability on diagonal	Grip test Smedley	Grip test Pneumatic	Numeric written	Numeric test Oral
Grip test Smedley	.85			
Grip test Pneumatic	.82	.88		
Numeric written	0.01	-0.01	.96	
Numeric oral	.12	.11	.90	.93

Mono-trait multi-method

Reliability on diagonal	Grip test Smedley	Grip test Pneumatic
Grip test Smedley	.85	
Grip test Pneumatic	.82	.88

How is the information in M-T/M-M matrix related to validity?

Reliability on diagonal	Grip test Smedley	Grip test Pneumatic	Numeric written	Numeric test Oral
Grip test Smedley	.85			
Grip test Pneumatic	.82	.88		
Numeric written	0.01	-0.01	.96	
Numeric oral	.12	.11	.90	.93

Why is information in M-T/M-M matrix necessary, but still insufficient to assess validity?

- Does not deal with all possible threats to validity.
- Does not speak to issues related to AMU of specific inferences which we will make from test-scores.

Vocational testing.

- May be comprised of present
  - functional capacity
  - aptitude/learning ability
  - Interest
  - Work related values assessments.
- Inferences are often related to specific occupations or classes of occupation (NOC).
  - We may note training requirements and the individual's capacity and interest in engaging in the work pre-requisite to a particular occup.

Interest tests examples

- Pictorial interest tests.
- Ashland Interest Assessment (1997)
  - 144 items
  - grade 3 readability
  - 12 scales (related to specific job-areas)
- CopSystem (CAPS, COPS and COPES)
- Jackson Vocational Interest Survey (JVIS)

Aptitude

- GATB
- DAT
- CAPS

Values

- COPES

END

Proposition

- The sum of knowledge pertaining to Voc. Eval. is sufficiently broad that no one person could master all the relevant fields that make up this interdisciplinary practice.
- Further: The fast pace of new developments in many fields will ensure that it will remain impossible to master everything pertinent to Voc. Eval.

The vast scope of knowledge in Voc. Eval. may represent a dilemma of minimum competence

- Possible resolutions of this dilemma.
  - Lifelong learning & professional dev.
  - Join CAVEWAS & CARP and associate with professionals from other disciplines to share information.
  - Sample widely. Read and Study the work of different people.
    - If you find nothing new at conferences or in books – you just aren't sampling widely enough.

Bill Angus: Introduction

- M.D. Angus & Associates Ltd. President/owner (since 1988)
- Former measurement consultant Psychological Corporation (1987)
- Registered Clinical Counsellor, BC.
- Member CPA, AERA, NCME, CCACC, AACC, Associate Member APA
- UBC Grad student (current) Measurement Evaluation & Research Methodology (MERM), Faculty of Ed. (ECPS).