Concepts of Validity: Bridging Concepts and Practices

UBC-Paragon Webinar Series 2021-3

June 4, 2021

Bruno D. Zumbo

Professor & Distinguished University Scholar

Tier 1, Canada Research Chair in Psychometrics and Measurement

Paragon UBC Professor of Psychometrics and Measurement

University of British Columbia

Citation:

Zumbo, B.D., & Shear, B.R. (2021). Concepts of Validity: Bridging Concepts and Practices [Video]. *UBC-Paragon Webinar Series 2021-3*. University of British Columbia, Vancouver, B.C.

- For a PDF copy of the presentation slides or a link to the video, please go to URL: https://brunozumbo.com/?page_id=31

Many parts of this presentation reflect a decade-long exchange of ideas and collaboration with Dr. Benjamin R. Shear based on mutual interests in psychometrics and applied statistics, including validity theory, and differential item functioning.

- We continue to address questions such as, “what do test scores measure, and how do we know?”
  - **Dr. Shear** is an assistant professor in the Research and Evaluation Methodology program, in the School of Education, University of Colorado, Boulder.

This lecture has grown out of and was shaped by feedback at 20 invited addresses by Bruno Zumbo over the last 16 years- see a list in acknowledgements at the end of the presentation.

Topics in Today's Webinar

1. Introductory Remarks
2. Concept of Validity
3. Bridging Concepts & Practices
   - Transitioning from the concept of validity to research and validation with the aid of an argument-based approach
4. Concluding Remarks
5. End material
   - How this webinar fits into my broader program of research
   - Acknowledgements
   - Bibliography

Section 1

INTRODUCTORY REMARKS

General Remarks

- The concept, method, and process of validation are central to social, psychological, and health science research, for without validation, any inferences made from a measure are meaningless.

- Throughout this presentation, the terms measure, instrument, test, assessment, survey, and scale will be used interchangeably and in their broadest senses to mean any coding or summarization of observed phenomenon.

- Furthermore, lest we fall into traditional camps and comfortable silos, validity applies equally to tests or measures used in

• language assessment, • educational measurement, • certification and licensure testing, • social indicators, • psychological instruments	• health measurement, • measures of health status, • patient-reported outcome measures (PROMS), • patient-reported experience measures (PREMS),

to name but a few of the common applications.

General Remarks

- Integrating and summarizing such a vast domain as validity invites, often rather facile, criticism.
- Nevertheless, if someone does not attempt to identify similarities among apparently different psychometric, methodological, and philosophic views and to synthesize the results of various theoretical and statistical frameworks, we would probably find ourselves overwhelmed by a mass of independent models and investigations with little hope of communicating with anyone who does not happen to be specializing on "our" problem, techniques, or framework.
  - Hence, in the interest of avoiding the monotony of the latter state of affairs, even thoroughly committed measurement specialists must welcome occasional attempts to compare, contrast, and wrest the kernels of truth from disparate validity positions.
  - However, while we are welcoming such attempts, we must also guard against oversimplifications and confusions, and it is in the interest of the latter responsibility that I write to the more general aim.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

General Remarks

- Although a lot of ground will be covered in this webinar, several themes should be evident from the material I will present. Let me speak to just a few of these themes.
  - There are no widely accepted series of steps that one can follow to establish validity of the inferences one makes from measures in the varied and disparate fields wherein measurement is used.
  - The process of validation, as I see it, involves a weighing and integrating the various bits of information from the whole of psychometric activities from specifying a theory of the phenomenon of interest to test design, scoring and test evaluation, and back to the theory itself.
  - I fall clearly in the camp of validity theorists who see the process of validation as an integrative disciplined activity.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

General Remarks

- That is, historically, we have moved from a correlation (or a factor analysis to establish “factorial validity”) as sufficient evidence for validity to an integrative approach to the process of validation involving the complex weighing of various bodies, sources, and bits of evidence – hence, by nature bringing the validation process squarely into the domain of disciplined inquiry and science.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

General Remarks

- Throughout my program of research, I have highlighted the importance of data modeling and assumptions as empirical commitments.
- Zumbo and Rupp (2004) remind us that it is the responsibility of mathematically trained psychometricians to inform those who are less versed in the statistical and psychometric theory about the consequences of their statistical and mathematical decisions to ensure that examinees are assessed fairly.
  - As Zumbo and Rupp state, everyone knows that a useful and essential tool such as an automobile, a chainsaw, or a statistical model can be a very dangerous tool if put into the hands of people who do not have sufficient training, handling experience, or lack the willingness to be responsible users.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible Modeling of Measurement Data For Appropriate Inferences: Important Advances in Reliability and Validity Theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oak s, CA: Sage Press.

General Remarks

- As Zimmerman and Zumbo (2001) note, formally, test data are the realization of a stochastic event defined on a product space $\Omega = \Omega_I \times \Omega_J$ where the orthogonal components, Ω_I and Ω_J, are the probability spaces for items and examinees respectively.
- The joint product space can be expanded to include other spaces induced by raters or occasions of measurement, a concept that was formalized in *generalizability theory* from an observed-score perspective and the facets approach to measurement from an IRT perspective.
- Hence, modeling of test data minimally requires sampling assumptions about items and examinees as well the specification of a stochastic process that is supposed to have generated the data.

Zimmerman, D. W., & Zumbo, B. D. (2001). The Geometry of Probability, Statistics, and Test Theory. *International Journal of Testing, 1*, 283-303.

General Remarks

- In summary, then, the item (or task) responses created by the interaction of examinees with items (or tasks) on a measure are considered to be *indicators* or markers of unobservable or *latent* variables.
  - I use the term *latent variable* to refer to a random variable that is deliberately constructed or derived from the responses to a set of items and that constitutes the building block of a statistical model (e.g., θ scores in IRT or factor scores in factor analysis).
- The statistical problem of measuring a latent variable can be characterized as involving two key tasks: (a) to find a set of indicators (items, scales, tasks, performances, or more generally referred to as measurement opportunities) that we believe that the latent variable will imply, and (b) to find a methodology for constructing a summary measure or scalar measure of the latent variable from these indicators.

- Denoting the set of indicators by

$$x = (x_1, x_2, \ldots, x_q)$$

the second part of the problem is to find a function

$$\varphi(x)$$

so that the numerical value of φ can be regarded as an appropriate scalar measure of the unobserved or latent variable.

- In this light, it is important to keep in mind that the main goal of modeling test data should always be to make valid inferences about the examinees but inducing latent variables into the data structure cannot mechanically increase the validity of these inferences.
- No matter how sophisticated the psychometric model, the statement of φ, and estimation routines have become, a test with poor validity will always remain so

General Remarks

- Much of what travels under the umbrella of validity theory and validation practices is the methodology of measurement (testing, assessment, and surveys).
  - How should one go about doing and evaluating measurement validity?
  - How should one go about doing and evaluating measurement, testing, and assessment?

- The question is one of the methodology of methodology, i.e., of metamethodology.

General Remarks

- The issue at hand is that one needs to make an inference from a score about the state or status of an observational unit, whether it is something that is at first glance self-evident or objective (e.g., a score on a math test) or subjective (e.g., self-reported well-being or stress).

- My point is that whether one is measuring knowledge reflected in a math test or subjective well-being, one has just one (or more) of an extensible set of indicators (survey questions, items or tasks) of a construct of interest.

- Importantly, the score on the indicator (survey questions, items or tasks):
  - is not equated with the construct it attempts to reflect,
  - and often nor is it considered to define the construct as in strict conventional operationism.

General Remarks

- In some descriptions of validity, what is at issue is whether:
  - the constructs are based on an elaborated theory and hence are considered as having the status of being explicitly theoretical (e.g., the narrow domain of a knowledge test defined on a test blueprint),
  - or whether the constructs are merely embedded in a network of expected or hypothesized empirical relationships (e.g., self-report psychosocial measures).

- Therefore, for self-report psychosocial measures, the validation and substantive theory development are inextricably intertwined, whereas this is less of a case for the former where, for example, test specifications may arise from a prescribed narrow curriculum.
  - [We see “content validity evidence” as the central driver from some of these instruments that are thought of as “objective.”]

- To organize the diverse themes and yet reflect current thinking in validity, you will see that we have separated theory and methodology.
  - This is not to imply that theory and method are disjoint, but rather it is meant to highlight one of the central concepts in contemporary validity theory – validity is not simply a technique or method

General Remarks

- Our primary goal today is to describe some new methods for validation.
- However, we believe that one needs to articulate what they mean by “validity” to go hand-in-hand with the process of validation. So, we need to delve into the “foundations”.
- To begin with, it is important to note that there is a parallel between:

$$\text{Methodology} \leftrightarrow \text{Method}$$

$$\text{Validity} \leftrightarrow \text{Validation}$$

General Remarks

We want to consider “validity” and “validation” for any kind of test or measure in educational, social, behavioral testing, or assessment settings.

- This general objective focuses on a meta-theory of validity rather than a tailored context for only, for example, cognitive, educational, language, or behavioral measures.
- Our aim is to think broadly to embrace and show the relation between many of the prominent views of validity with an eye toward some synthesis.

17

General Remarks

In what follows we reflect on the state of the praxis and theorizing in validity and validation in general:

> *… where it has been, where it is now, and where we think it is, and should, be going.*

Along the way we intend to integrate and summarize major trends in the validity literature, provide some organizing principles that allow one to catalogue and then contrast the various validation methods, and to shine a light on what we believe is the future of validity theory and the process of validation.

Section 2

THE CONCEPTS OF VALIDITY

Objective

Provide a brief historical overview of validity theory with an eye toward a description of recent work on the theory of validity and the process of validation.

Objective

- This section portrays the concepts of validity undergoing consolidation, debate, and re-conceptualization.
- We raise new questions and re-awaken long-standing debates that lie at the heart of empirical science and speak to our collective desire to formalize and better articulate the concepts and measures we employ.
  - As we are reminded in the philosophies of science, linking concepts to observations (in the history of validity, relying on nomological network) is a fundamental strategy to clarify the meaning of a measure.

Validity: An over-the-shoulder look back

Aristotle, in his *Metaphysics*, pointed out that "we understand those things best that we see grow from their very beginnings."

We thus begin our discussion of measurement validity with an over-the-shoulder look at the history of the idea and of procedures that were developed to aid in the validation process.

- The general aim is to trace the history of the concept of measurement validity and validation methods from their heuristic beginnings to the more statistically rigorous methods currently available such as IRT, structural equation models for multi-trait multi-method matrices etc..

Zumbo, B.D., & Padilla, J.L. (2020). The Interplay between Survey Research and Psychometrics, with a Focus on Validity Theory. In P.C. Beatty, D., Collins, L., Kaye, J.L. Padilla, G. Willis, and A. Wilmot, (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 593-612). Hoboken, NJ: Wiley.

Validity: An over-the-shoulder look back

In what follows we propose that we consider, what appears to be, four somewhat distinct time periods of validity praxis and theorizing.

Please note that we are not suggesting distinct historical periods and a natural linear step-wise progression toward our current thinking .. and not suggesting "evolution" to the best theories.

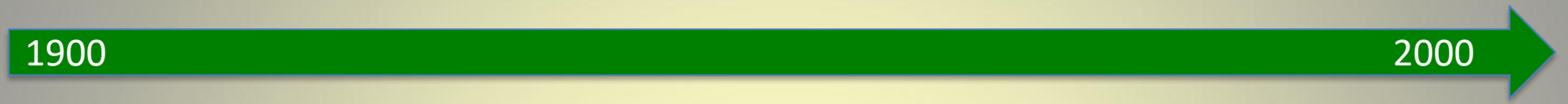
- Note: we are using "praxis" here to (a) convey a distinction between practice and theory, (b) highlight the application or use of the knowledge and/or skills, and (c) also reflect some of what is, in essence, the convention, habit, or custom of validity work of the time periods.

Shear, B.R., & Zumbo, B.D. (2014). What Counts as Evidence: A Review of Validity Studies in Educational and Psychological Measurement. In Bruno D. Zumbo, and Eric K.H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 91-111). New York: Springer.

Validity: An over-the-shoulder look back

1. The early- to mid-1900s: dominated by the criterion-based model of validity, with some focus on content-based validity models.
2. The mid-1930s to the late 1960s saw the introduction of, and move toward, the construct model with its emphasis on construct validity; a seminal piece being Cronbach and Meehl (1955).
3. The period post Cronbach and Meehl, mostly the 1960s to end of 1990s, saw the construct model take root and saw the measurement community delve into a moral foundation to validity and testing by expanding to include the consequences of test use and interpretation (Messick, 1975, 1980, 1988, 1989, 1995, 1998)
4. A period since about 2000 to date in which the debate about validity and validation has started up again after a quiet time post Cronbach's and Messick's programs of research.

Validity: An over-the-shoulder look back



Early 1900-1930's the criterion view

The key element being validity as correlation or prediction, involving either: an objective measure of that which the test is used to measure, a criterion, or anything for which it correlates.

The mid-1930s to the late 1960s

The proliferation of the multiple "types" of validity, and that we are validating the measures themselves in the psychological literature and in the early versions of the APA/AERA/NCME *Standards*.

1960s to end of 1990s

The "types of validity" talk is still dominant: discriminant validity, convergent validity, face validity, etc., as well as the methodological developments beyond the simple "validity coefficient" (a correlation) to patterns among planned validation studies in the multi-trait multi-method matrix.

Validity: An over-the-shoulder look back

Constructs take root and construct validity as the accumulation of evidence (the 1960s to end of 1990s , but peaked in the mid 1970s)

- The landmark paper in this tradition is Cronbach and Meehl (1955) and the description of *construct validity* and the explicit use of the *nomological network* to establish meaningfulness of the measure.
- Construct validity based on accumulation of research results: formulate hypotheses, test hypotheses. (APA/AERA Standards, 1974)
- Cronbach's (1971) and later view of validation (and perhaps validity) as evaluation and, in some sense, a process of social rhetorical arguments.

The Concept of "Validity"

If one wants to advance the theorizing and practice of measurement we believe, that one needs to articulate what they mean by "validity" to go hand-in-hand with the process of validation. So, we need to delve into the foundations.

We need to exploit the parallel noted earlier:

Methodology ↔ Method

Validity ↔ Validation

Some Concept(s) of "Validity"

Eight conceptualizations of "validity" … some of which imply a particular process of validation.

1) A test is a predictive device or a short-hand. Therefore, validity is about establishing whether a test is a good predictive device or short-hand.
   - The correlation coefficient determines the validity (Hull, 1928). Validity is the correlation of test scores with some other objective measure of that which the test is used to measure (Bingham, 1937). (primary validation evidence is criterion correlation and prediction).

Zumbo, B.D., & Padilla, J.L. (2020). The Interplay between Survey Research and Psychometrics, with a Focus on Validity Theory. In P.C. Beatty, D., Collins, L., Kaye, J.L. Padilla, G. Willis, and A. Wilmot, (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 593-612). Hoboken, NJ: Wiley.

Some Concept(s) of "Validity"

2) Garrett's (1937) statement that validity is the extent to which the test measures what it purports to measure. (does not imply an process of validation)
3) Cronbach & Meehl (1955) and the logical empiricist influenced "nomological network" and "construct validity". Important because it signaled that tests changed from just being "predictive devices" to being "signs" of an underlying attribute. (validation: empirically establishing the nomological network)

Some Concept(s) of "Validity"

4) Messick (1970s to 1999) and reflected in the AERA/NCME/APA (1999) *Test Standards*
Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

(validation: It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. Multiple sources of validity evidence; consideration of consequences of test use.)

Some Concept(s) of "Validity"

5) Embretson's (e.g.,1983, 2007) work on construct representation versus nomothetic span, and a universal system for construct validity to illustrate how diverse evidence is relevant to measurement claims. (validation: well-suited for *formal* cognitive modeling)
6) Borsboom, Mellenbergh, and Van Heerden (2004) who argue that a test is valid for measuring an attribute if and only if the attribute exists and variations in the attribute causally produce variations in the outcomes of the measurement procedure. (validation: well-suited for *formal* cognitive modeling)
7) Lissitz & Samuelson (2007) validity is content representation (validation: content validity evidence)

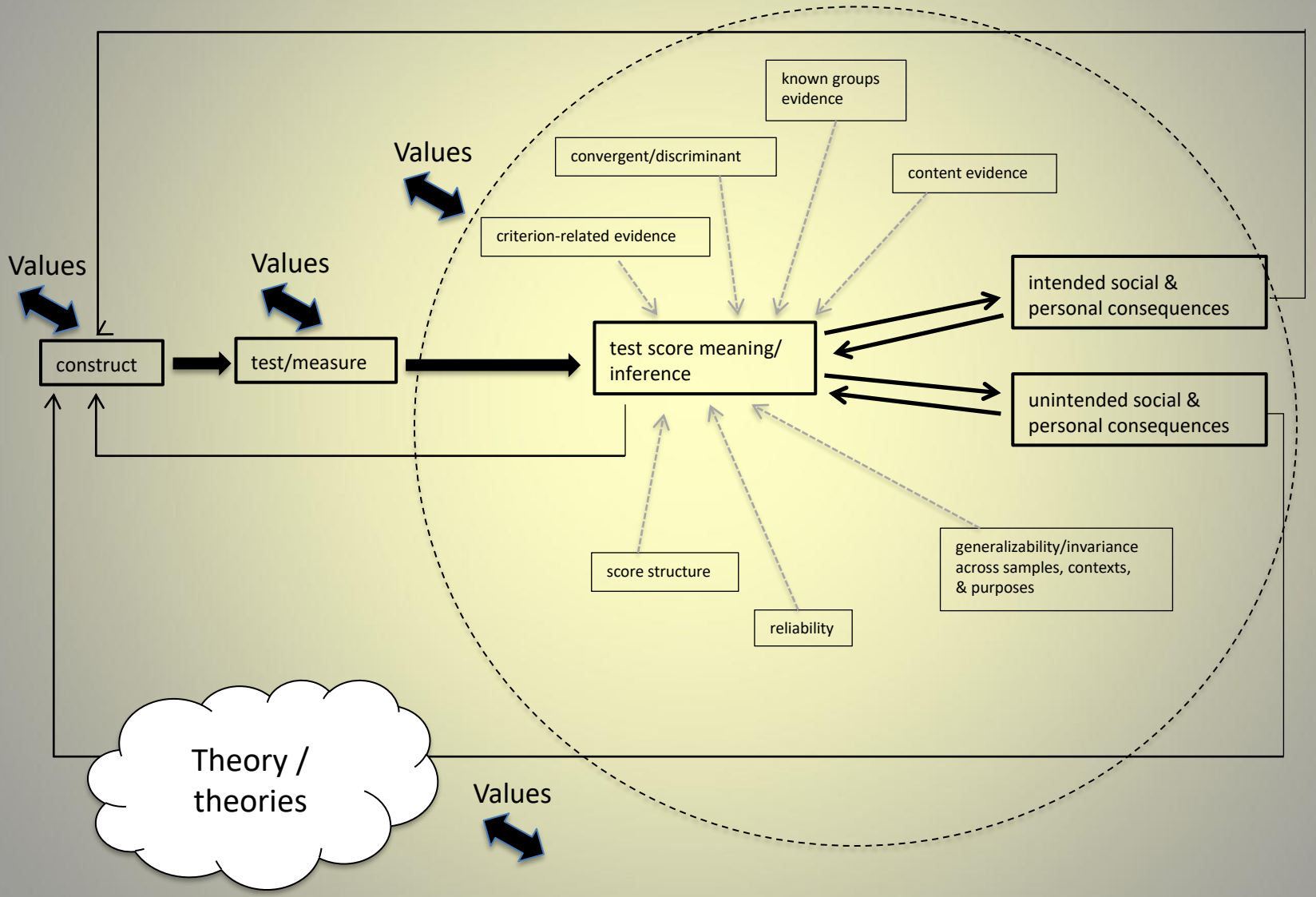
Some Concept(s) of "Validity"

8) Zumbo (2005, 2007, 2009, 2017) has taken the view of "validity" as the explanation for the item and test score variation, and "validation" as the process of developing and testing the explanation. *Contextualized pragmatic* explanation.

(particularly well-suited as a foundation for cognitive and statistical modeling of item response and test score data; also, for Zumbo's Draper-Lindley-DeFinnetti (DLD) methods; foundation for studies of heterogeneity)

- Zumbo (2007) envisioned a "judicial or courtroom" metaphor where all the evidence comes together and is judged, cases are made, evidence (witnesses) come forward and a reasoned body judges the evidence (weighing different aspects) for validity of the inferences made from a test or measure. In Zumbo (2009) I moved to a "cognitive integration" approach.
- Zumbo & Forer (2011), multilevel validation of multilevel construct for health and social policy measures.
- Response processes are important in the explanatory-focused approach (e.g., Zumbo & Hubley, 2017; Zumbo, 2017)

Hubley and Zumbo's (2011) revised view of validity and validation

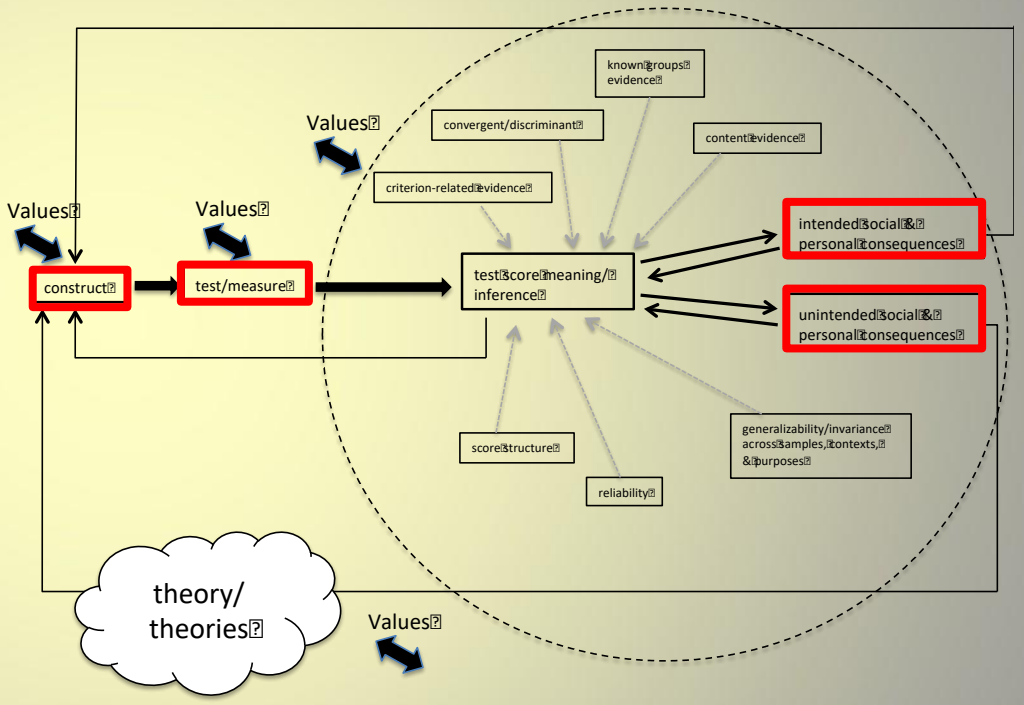


Hubley & Zumbo (2011): Five Points

First, at the core one can envision that from constructs one develops tests/measures, to which one ascribes test score meaning and inference.

From this test score meaning and inference emerges (a) intended social and personal consequences, and/or (b) unintended social and personal side effects.

Very importantly, these consequences and/or side effects (either personal and/or social) may also influence test score meaning and inference.

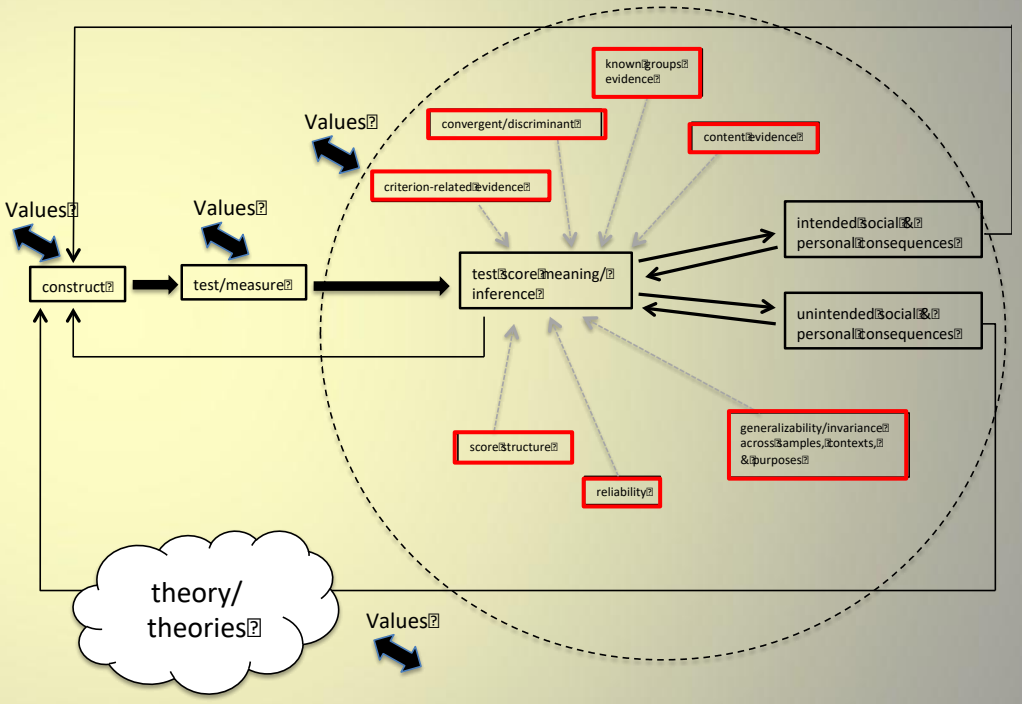


Hubley, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, *103*(2), 219-230.

Hubley & Zumbo (2011): Five Points

Second, test score meaning and inference is affected and shaped by several forms of validity evidence, including but not necessarily limited to:

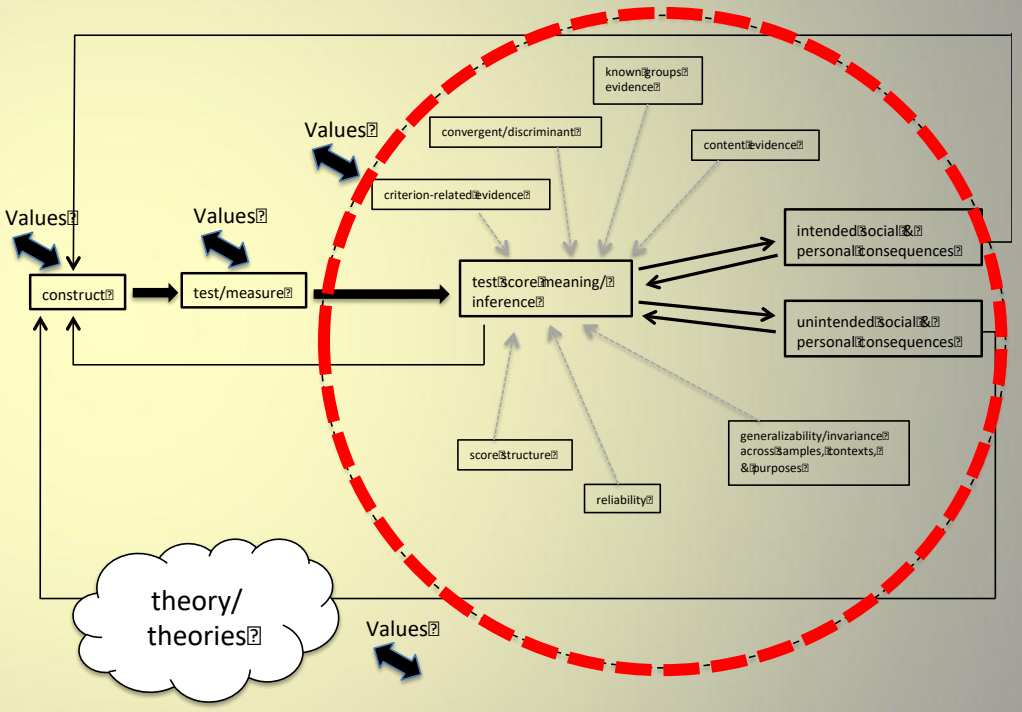
- criterion-related,
- convergent/discriminant,
- known groups,
- content,
- reliability,
- score structure and
- generalizability/invariance evidence.



Hubley & Zumbo (2011): Five Points

Third, the dashed circle encompasses what we could consider construct validity – containing within it the process of validation that provides the various sources of validity evidence and includes consequences and side effects.

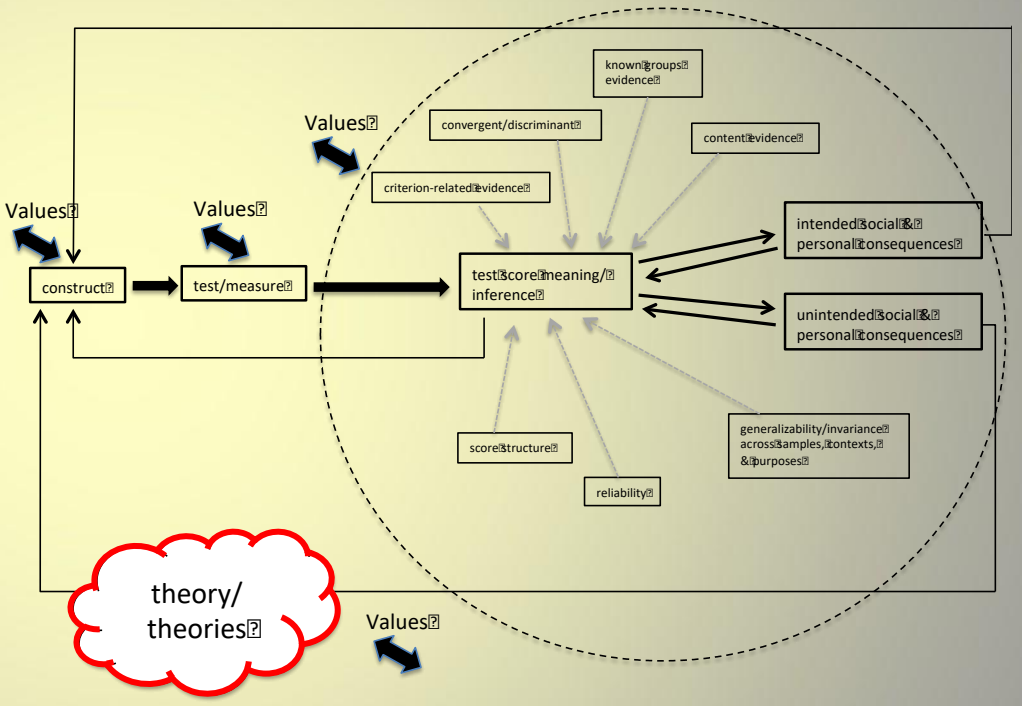
The centrality of the large (dashed) circle is meant to signify that construct validity is at the core of a unified view of validity and validation.



Hubley & Zumbo (2011): Five Points

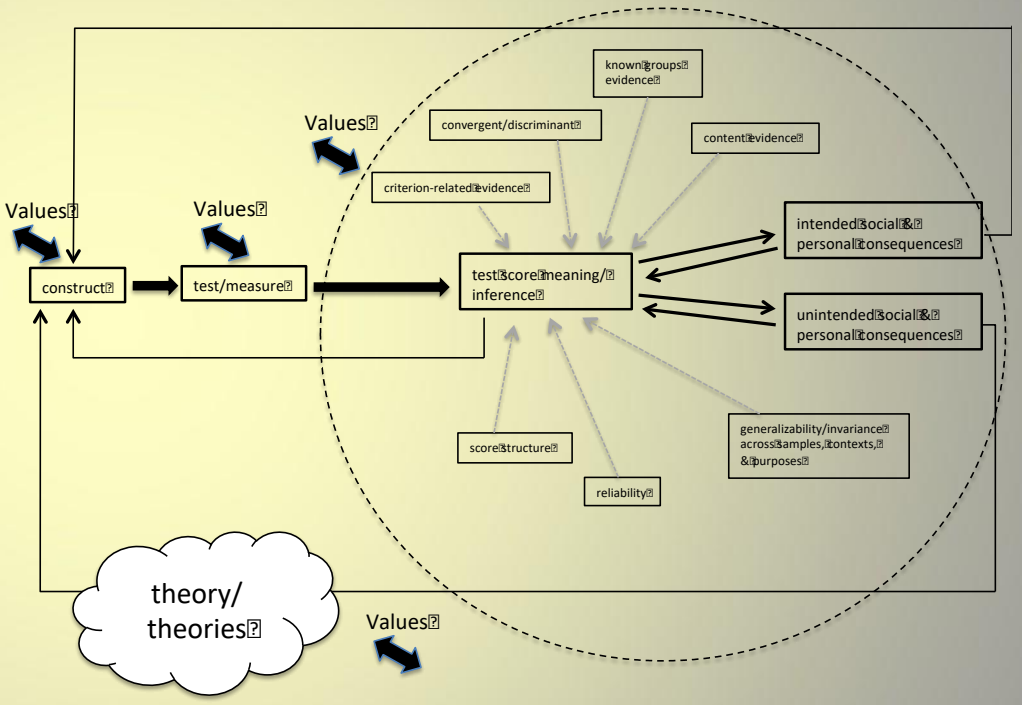
Fourth, one sees that theory or theories influence the construct, the test/measure, and construct validity.

The 'theory / theories' we are referring to include the theory related to the construct, theories related to the sample and context, and psychometric theory and models.



Hubley & Zumbo (2011): Five Points

Finally, we can see that the effect of values is pervasive and includes the impact of theory/theories (broadly defined), the construct itself, test/measure, and the impact of values on construct validity as well as validation choices and decisions.



Brief Summary of Views about Validity

We take a position herein and elsewhere that validity is a matter of inference and the weighing of evidence, and that explanatory considerations guide our inferences (Zumbo, 2005, 2007, 2009).

My current leanings are toward inferences to the best explanation- early influences from Bill Rozeboom and later by Brian Haig's and Paul Thagard works, I lean toward abductive methods.

Haig, B.D. (2018). *Method Matters in Psychology: Essays in Applied Philosophy of Science*. [In the series, Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer, Press.
Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10*, 371–388.
Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research, 40*, 303–329.

Validity

- In our view, in terms of the process of validation (as opposed to validity, itself):
  - the statistical methods, as well as the psychological and more qualitative methods of psychometrics, work to establish and support the inference to the best explanation.
- This best explanation is "validity" itself; so that validity is the explanation, whereas the process of validation involves the myriad methods of psychometrics to establish and support that explanation.
  - This is an interesting meta-theoretical place from which to re-read some classic papers in validity and to try and synthesize various views of validity.

What the view of validity and validation implies

- It is important to highlight that, as Kane (2001) reminds us, there are strong and weak forms of construct validity.
- The weak form is characterized by any correlation of the test score with another variable being welcomed as evidence for another "validity" of the test.
- That is, in the weak form, a test has as many "validities" and potential uses as it has correlations with other variables.
  - In contrast to the weak form of construct validity, the strong form is based on a well-articulated (explanatory) theory and well-planned empirical tests of that theory.

Validity/validation

In our view, the strong form of construct validity should provide an *explanation* for the test scores, in the sense of the theory having explanatory power for the observed variation in test scores.

- We share the view with other validity theorists that validity is a matter of inference and the weighing of evidence; however, in this view, explanatory considerations guide our inferences.
- Importantly, however, explanation acts as a regulative ideal; validity is the explanation for the test score variation, and validation is the process of developing and testing the explanation.

What the view of validity and validation implies

In short, the strong-form is theory-driven (à la Cronbach & Meehl, 1955) whereas the weak form implies that a correlation with some criterion is sufficient evidence to use the test as a measure of that criterion.

In our view, the strong form of construct validity should provide a contextualized pragmatic *explanation* for the test scores (Zumbo, 2009).

- Pragmatic view of explanation, emphasizing the context of explanation.

Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Validity/validation

In essence, we see validation as a higher order integrative cognitive process involving everyday (and highly technically evolved) notions like concept formation and the detection, identification, and generalization of regularities in data whether they are numerical or textual.

Validity/validation

From this, after a balance of possible competing views and contrastive data, comes understanding and explanation.

- What I am suggesting is a more technical and more data-driven elaboration of what we do on a day-to-day basis in an open (scientific) society; we are constantly asking why the things are the way we find them to be, answer our own questions by constructing explanatory stories, and thus come to believe some of these stories based on how good are the explanations they provide.

Validity/validation

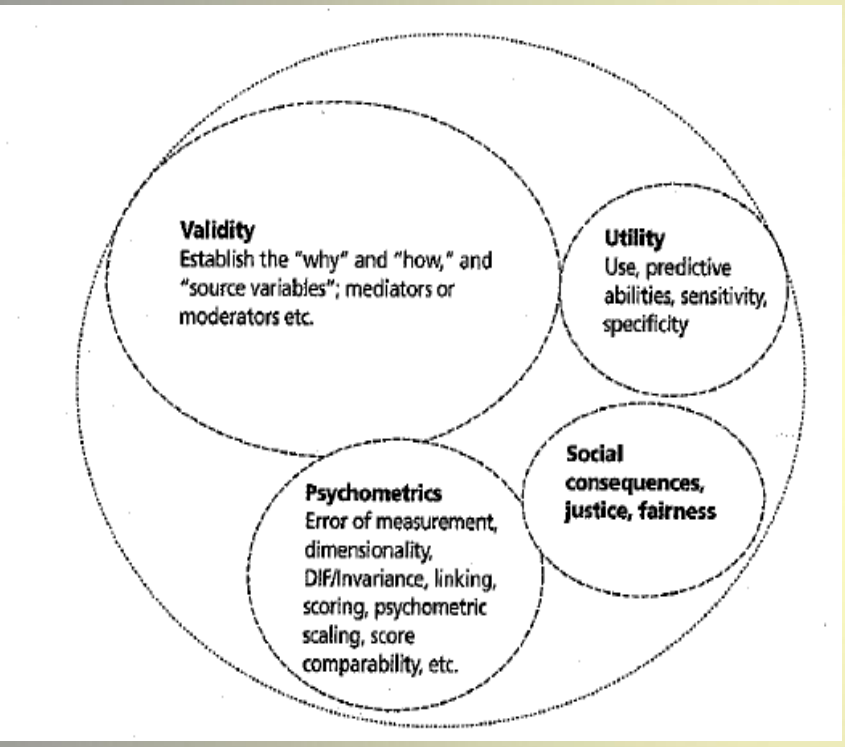


Figure 1 From Zumbo (2009).

Figure 1 depicts the four core elements of the integrative cognitive judgment of validity and the process of validation: validity, statistics, social consequences, and matters of utility – all of which are tightly packed in the Figure close to each other and hence influence, and shape, each other.

We can see that validity is separate from utility, social consequences, and the statistics, but validity is shaped by these.

Furthermore, the inferences are justified by the statistics, social consequences, and utility but validity is something more because it requires the explanation.

Validity/validation

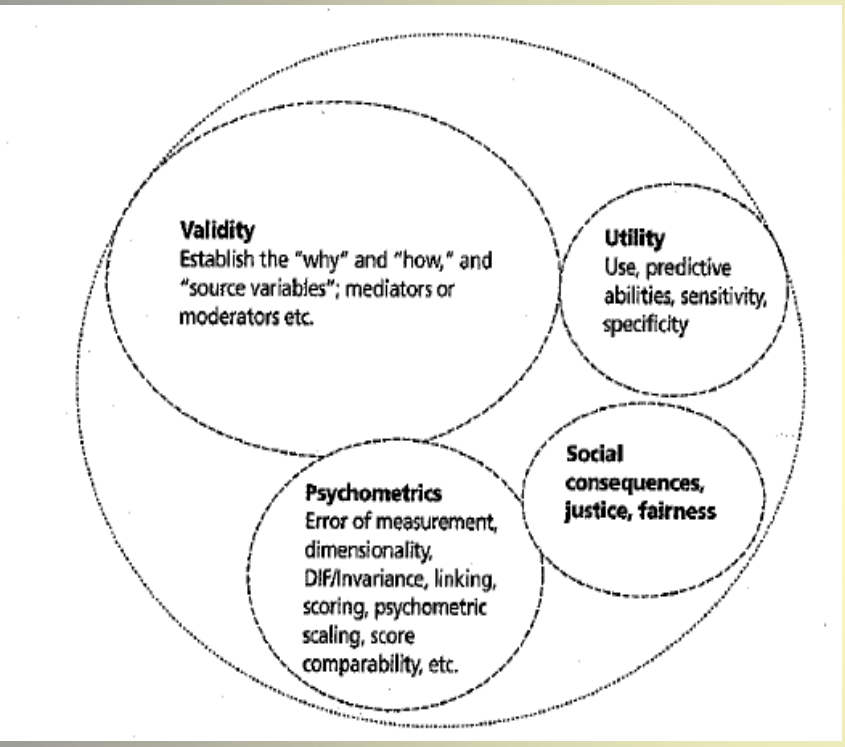


Figure 1 From Zumbo (2009).

The basic idea underlying my explanatory approach is that, if one could understand the variation in an indicator, then that would go a long way toward bridging the inferential gap between test scores and the constructs.

According to this view, validity per se, is not established until one has an explanatory model of the variation in test (item) scores and the variables mediating, moderating, and otherwise affecting that observed variation – recall tht its is a regulative ideal.

This is a tall hurdle indeed. However, I believe that the spirit of Cronbach and Meehl's (1955) work was to require explanation in a strong form of construct validity.

Section 3

BRIDGING CONCEPTS & PRACTICES

Transitioning from the concept of validity to research and validation with the aid of an argument-based approach

Arguments in Validity and Validation

- In this section we begin to transition from more conceptual or theoretical considerations to the applied practice of validation. There is an oft-cited gap between validity theory and the practice of validation, which many trace to the theory of construct validity and difficulty of implementing such a theory (Messick, 1988; Shepard, 1993; Kane, 2004).

- Grows out of a notion that we validate inferences and uses rather than tests. We must clearly state the inference and assumptions that move us from observed performances to proposed interpretations regarding a construct or uses.
  - In particular, Kane describes an interpretive argument, which clearly states the assumptions and inferences that move us from an observation to a final interpretation or decision. Then, in a separate process, called a validity argument, we evaluate the plausability of the inferences and assumptions we have proposed.

First proposed by Cronbach (1988); more systematically elaborated by Kane (1992, 1999, 2001, 2002, 2004, 2006, 2009).

Arguments in Validity and Validation

Cronbach (1988), Kane (1992, 2006), Shepard (1993) and others advocate using argument as a way to frame or focus validation efforts and to clarify intended interpretations and uses.

> "The main advantage of the argument-based approach to validation is the guidance it provides in allocating research effort and in gauging progress in the validation effort" (Kane, 2006, p. 23).

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Arguments in Validity and Validation

What does an interpretive argument look like?

- What follows are three examples of how we might go from an observed performance to a final interpretation or decision for an individual.
- Note that in some cases we make a final decision and in others we arrive at a description about a person.
- Various studies may evaluate or investigate one or more of these inferences and gather evidence to support (or refute) them. This would be done as part of the validity argument.

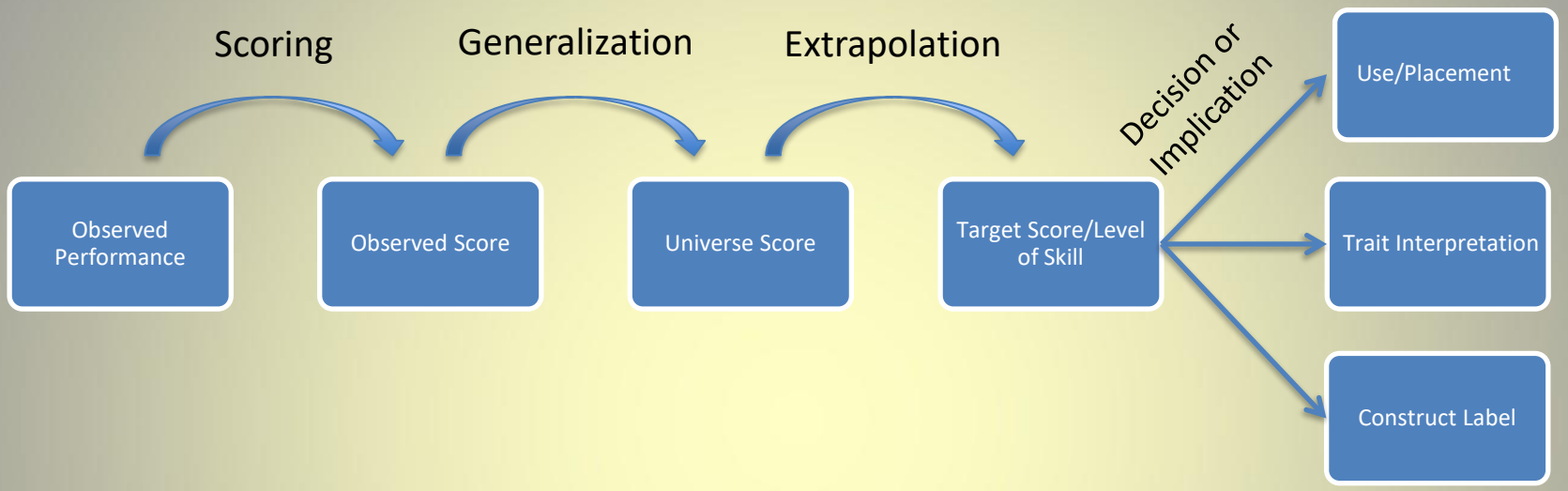
"The interpretive argument is to provide a clear statement of the inferences and assumptions inherent in the proposed interpretations and uses of test results, and these inferences and assumptions are to be evaluated in a series of analyses and empirical studies." (Kane, 2006, p. 23)

"While the interpretations discussed in Sections 3 to 5 are evaluated in terms of their coherence and plausibility, decisions are evaluated in terms of their outcomes, or consequences." (Kane, 2006, p. 51)

- Highlight that these are essentially different forms of interpretive arguments (examples taken from Kane, 2006) rather than different arguments per se.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Kane's Argument-based Approach to Validation



Notes:

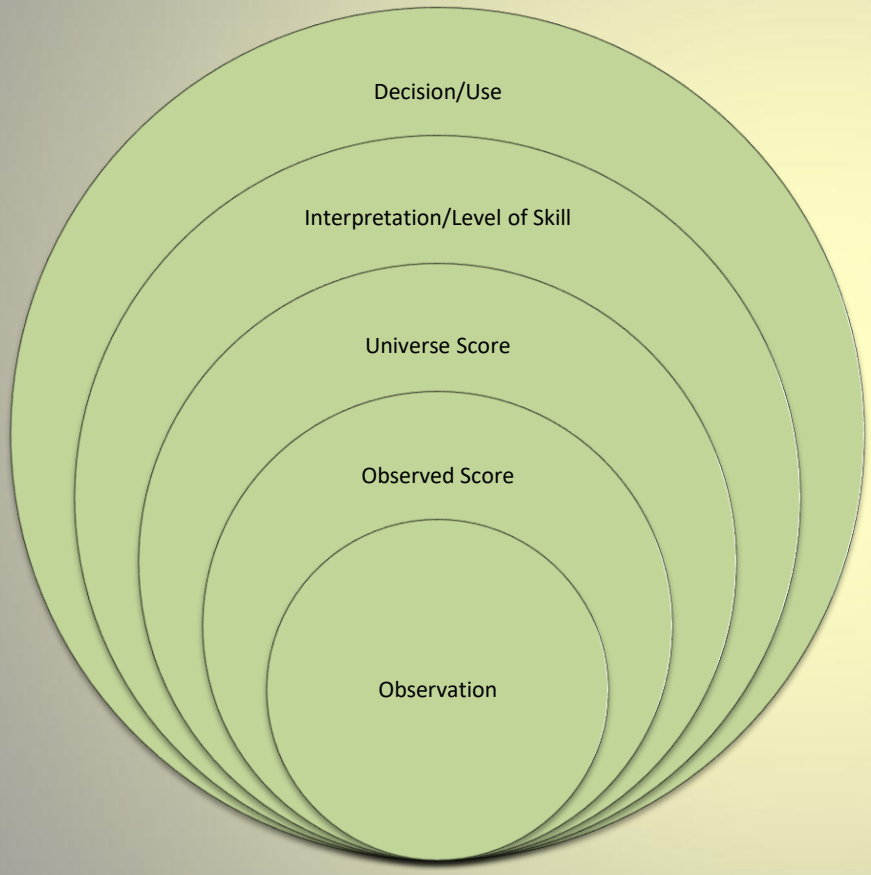
Different forms of interpretive arguments.

Interpretive argument followed by the validity argument.

Descriptive vs. decision-based interpretations.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane's Argument-based Approach to Validation



Notes:

Presence/influence of G-theory.

Connection to DLD Framework.

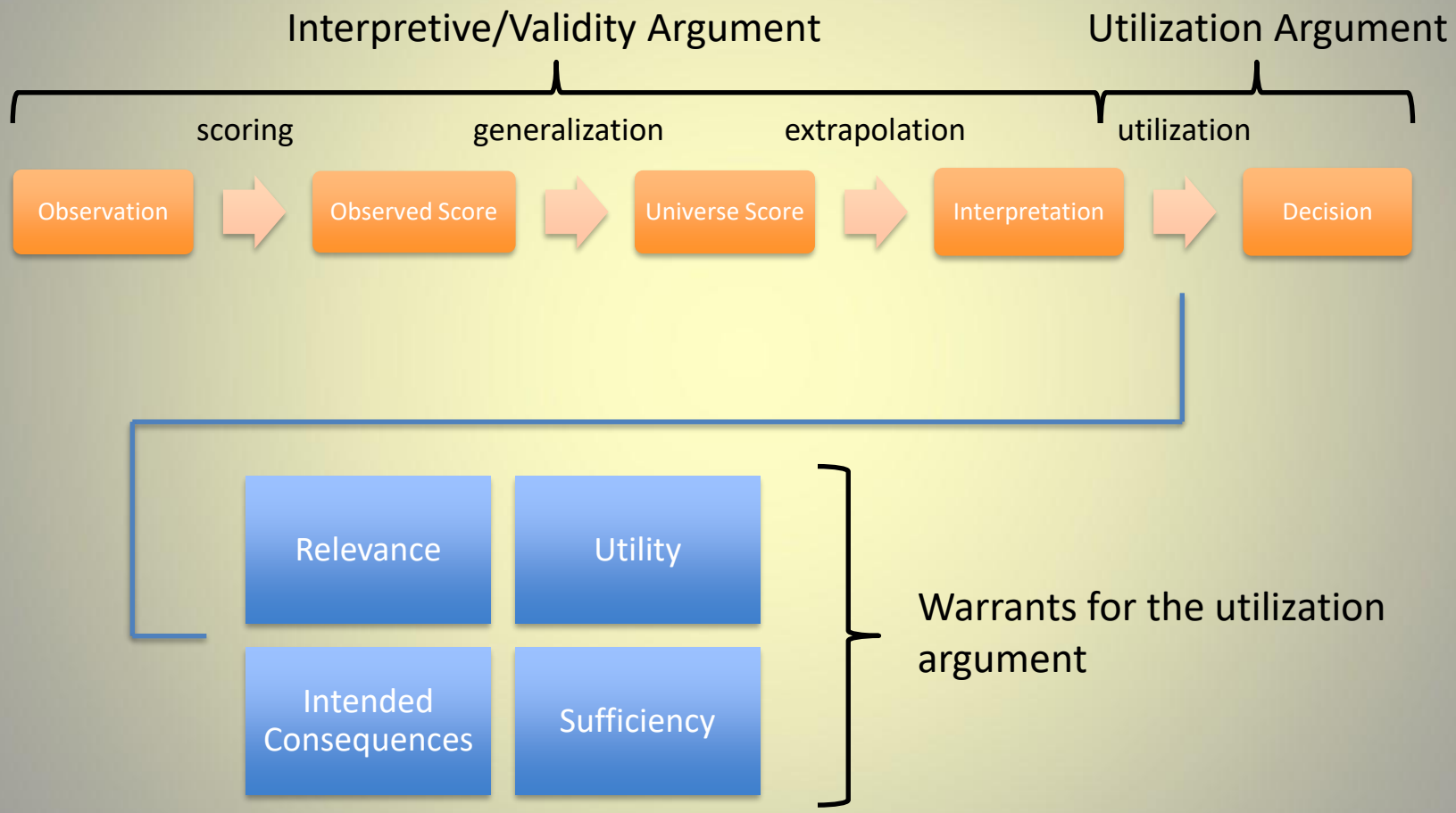
Competency vs construct.

Bachman, supporting a case for test use

- Lyle Bachman differentiates between arguments that lead toward a description versus those that lead towards a particular decision.
- For example, Bachman differentiates between making an inference about a potential candidate's language ability in certain tasks from the subsequent decision about whether to hire that person.
  - He feels there is not enough systematic attention focused on supporting the decision as compared to stating the interpretation.
  - He proposes the following framework, the creates a separate argument for those cases in which we are also evaluating a particular use, not only an interpretation or description of observed performance.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

Bachman's Assessment Use Argument (AUA)

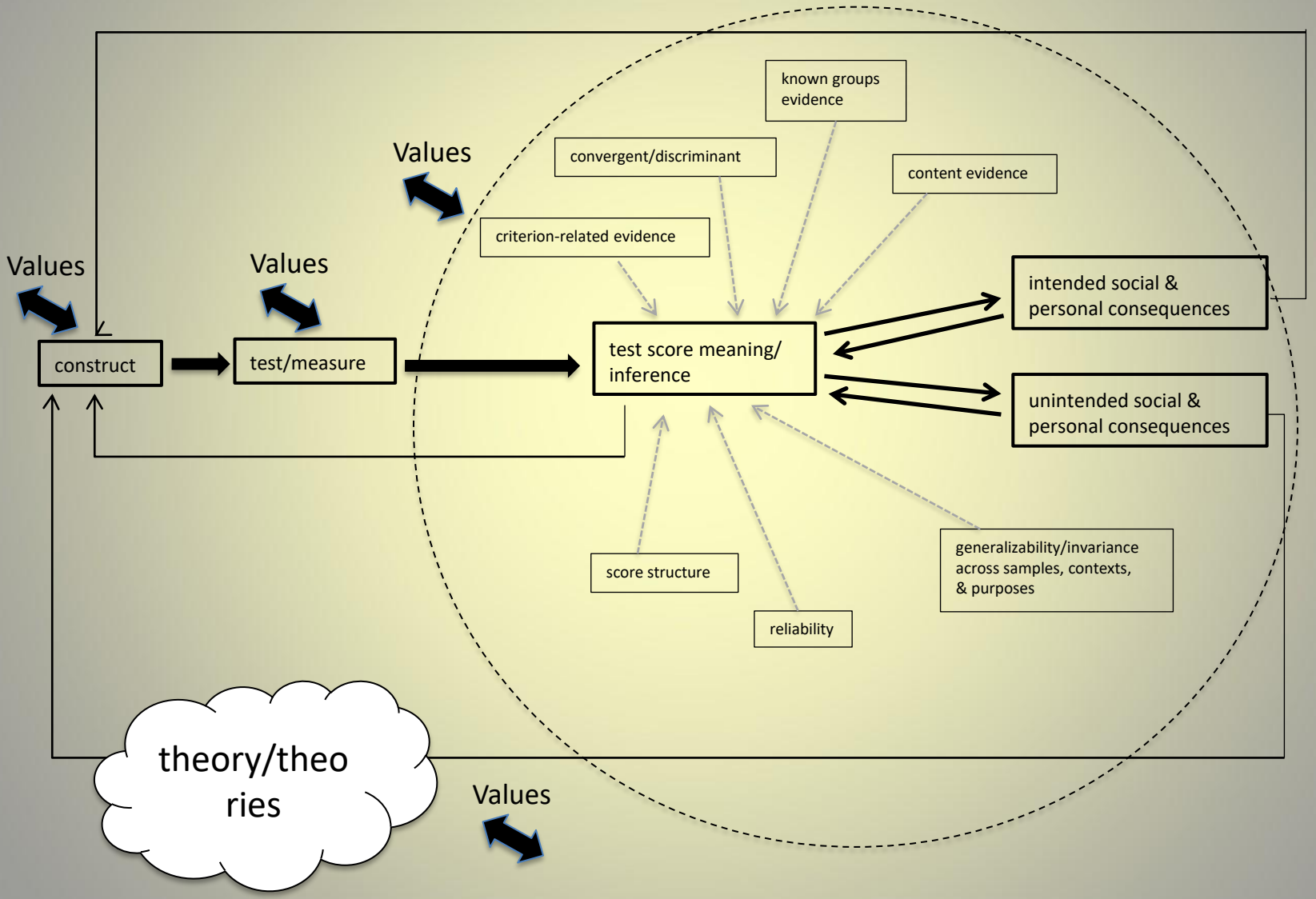


Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34.

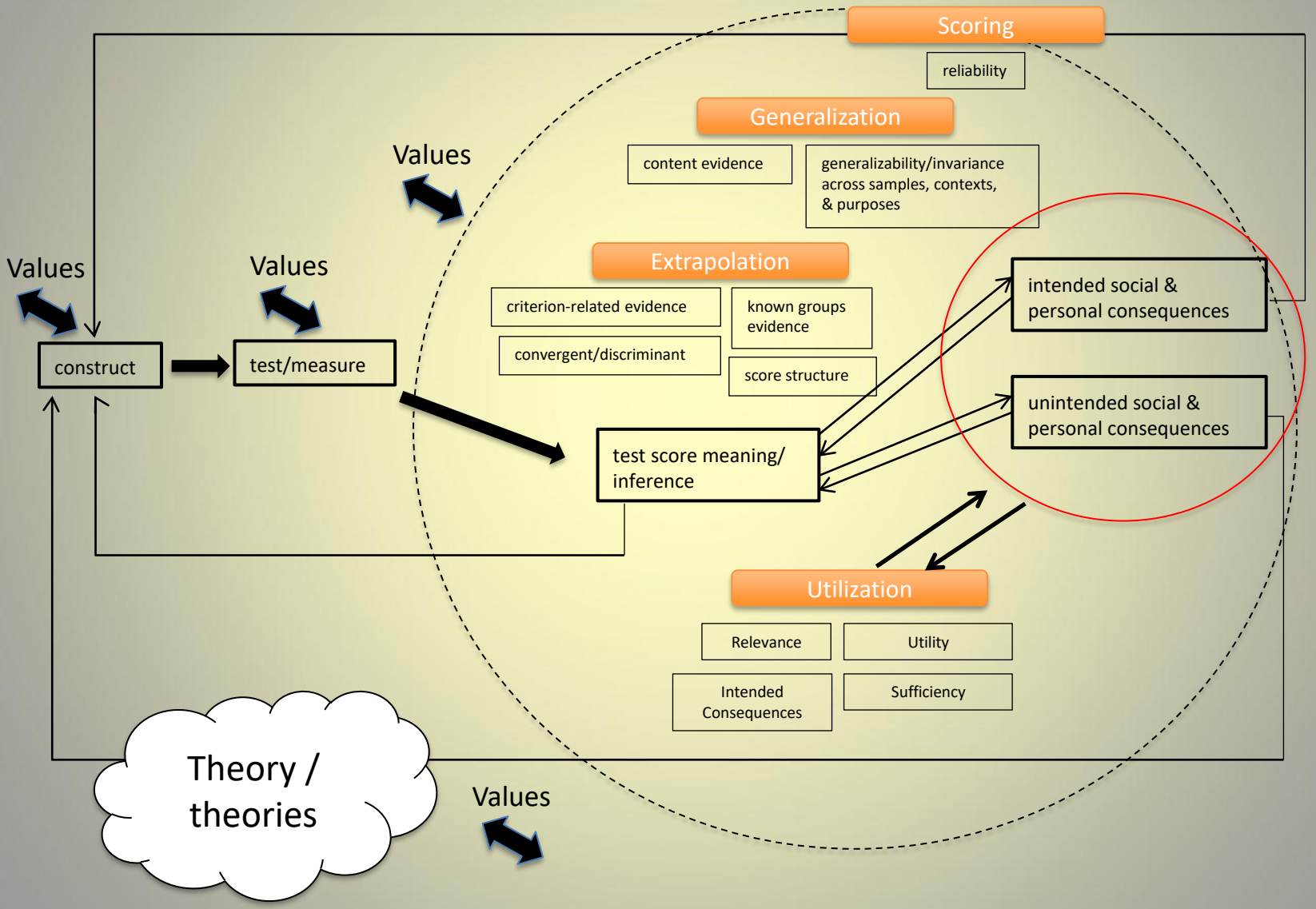
A Merging of Hubley & Zumbo with Arguments

- After considering these argument-based approaches, we can now return and consider how they fit with the concept of validity presented earlier.
  - First, we re-arrange the various sources of evidence that feed into our interpretations.
  - Then we can examine where different forms of evidence might be used to support the interpretive argument.
  - Notice that this ends with an interpretation, rather than decision, but as Zumbo has mentioned still raises issues about the consequences involved.
  - Adding the utilization aspects discussed by Bachman brings in a new set of evidence – here it is very clear that the consequences of using a particular test to make decisions needs to be considered or addressed.

Hubley and Zumbo's (2011) revised view of validity and validation



A Merging of Hubley & Zumbo with Arguments



Arguments and Explanations

At a more conceptual level, we might compare the argument-based approach and explanation-focused view by posing the following question…

Is an explanation an argument or is an argument an explanation?

Probably are multiple answers. Turning to logic, explanations are seen as types of arguments.

There are at least two types of arguments: justificatory and explanatory.

Sinnott-Armstrong, W. & Fogelin, R. (2010). *Understanding Arguments: An Introduction to Informal Logic*. United States: Wadsworth CENGAGE Learning.

Types of Arguments

Distinguished largely by purpose or use rather than form:

- Explanatory: provide an explanation of why or how something we agree about has happened; how did we arrive at a particular interpretation?
- Justificatory: provide reasons for belief; why should I accept the proposed interpretation?

Focusing on the purpose of the argument brings our attention to who the audience is. This may be important.

Interpretive argument as *explanatory*?

Validity argument as *justificatory*?

Arguments and Explanations

- These two sorts of arguments often have similar forms, moving through chains of inferences.
- But their purposes and the context in which we use them will often differ.
  - Please note that inference to the best explanation essentially combines these; first we formulate an explanation, then a justificatory argument to convince us it is indeed the best possible explanation.
- There is an interesting parallel here between focusing on the use of a test to guide validation work; similarly, we can focus on the use of the argument to guide our construction of the argument.

Types of Arguments

Although it is clear how the validity argument serves to evaluate the particular pieces of the interpretive argument, what standards ought to be used to judge whether the interpretive argument, in context, is complete or serves its purpose (Messick, 1995)?

> Perhaps by conceptualizing the interpretive argument as explanatory, we gain a new set of criteria (for explanations) by which to evaluate our interpretive argument.

Messick S. (1995). Validity of Psychological Assessment : Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749.

Types of Arguments

- By framing the two parts of the validity argument as explanatory/justificatory, we can leverage various frameworks for evaluating explanations in the service of developing our interpretive argument.

- In addition to Kane's clarity, coherence, plausibility of inference and assumptions…"Implicit assumptions can be particularly harmful because they may be left unexamined" (p. 29).

Types of Arguments

- Just as measures are fallible (hence the need for validation) so too are our arguments fallible. And some arguments may be solid in one context but not in another.
  - Hence, we need an analogous procedure to be sure our arguments are sufficient in a particular case, the same way we evaluate whether a test use or interpretation is sufficient in a particular context.

- Criteria for inference to the best explanations (think: selecting the best interpretive argument):
  - “In sum, a hypothesis provides the best explanation when it is more explanatory, powerful, falsifiable, modest, simple, and conservative than any competing hypothesis” (Sinnott-Armstrong & Fogelin, 2010, p. 262).

Section 4

SOME CONCLUDING REMARKS

What the view of validity and validation implies…

- An important issue:
  - When can we start using a measure? Or do we need to establish the "validity" (i.e., the explanation for the test and item response variation) before we can use the measure to make inferences and research conclusions?
    - Answer: **Explanation is a regulative ideal**.

- What I am suggesting is that assessment research research take on a robust and integrative research agenda in which the bounds and limitations of the inferences we can make from scores (and hence ferreting out invalidity) becomes a core task of the research agenda.

What the view of validity and validation implies…

- The demands are high, but we believe that they are in line with the desires spelled out in the seminal paper by Cronbach and Meehl (1955), read as a strong program of construct validity research.
- One thing that gets highlighted by Zumbo's DLD framework (2007) is that, in general, in psychometrics do not unthinkingly assume homogeneity.
  - Work, where possible, with multi-level and latent class models.
- In the tradition of inference to the best explanation (or abductive methods) the latent variables of factor analysis may take on an explanatory role.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) Handbook of Statistics, Vol. 26: Psychometrics, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

THE END!

Thank you for your time.

For a copy of these slides and/or the forthcoming papers please write to:

bruno.zumbo@ubc.ca

How this webinar fits into my broader program of research

The program of research on validity is organized around three themes:

1) Towards metamethodology for measurement and validity theory

Focus of today's Webinar.

- Current developments contextualized in a history of science; history of validity theory
- Exploring a view of "validity" as the explanation for the test score variation, and validation as the process of developing and testing the explanation. Meta-theory being the focus (e.g., Zumbo, 2005, 2007a, 2007b, 2009; 2017; Woitschach, Zumbo, B.D., & Fernández-Alonso, 2019).

2) Statistical and methodological approaches and techniques:

- Focus on latent variable modeling (e.g., DIF, Pratt Indices, multi-group factor analysis, IRT invariance).
- Understanding and Investigating Response Processes in Validation Research (e.g., Zumbo & Hubley, 2017; Zumbo, 2017)
- Multi-level construct validation for assessment systems like NAEP and statewide assessments (e.g., Forer & Zumbo, 2011; Zumbo & Forer, 2011).
- A micro-simulation framework for validation; a sensitivity analysis framework.

3) The use of validity (Messick's work) as a framework for program evaluation in e-learning (book by Ruhe & Zumbo, 2009, Guilford Press).

Acknowledgements

This lecture has grown out of, and was shaped by feedback at invited addresses by Bruno Zumbo:

- Continuing The Legacy Of Cronbach & Meehl And Of Messick To Advocate For A Science Of Measurement Validation: New Psychometric Methods for Variable Ordering. (2019). Invited Colloquium at the Department of Psychology Colloquium Series, University of Manitoba.
- On New Methods That Support An Explanation Focused View of Test / Measurement Validity: Pratt Indices for Latent Variable Models. (2018). Address at the Centre for Research in Applied Measurement and Evaluation (CRAME), University of Alberta, Edmonton, AB.
- Language Testing: Impact of Technology and Placement Issues. (2018). Invited panel address at the 11th Conference of the International Test Commission, Montreal, Canada.
- Methodologies Used To Ensure Fairness And Equity In The Assessment Of Students' Educational Outcomes. (2018). AERA Presentational Symposium "Methodology and Equity: An International Perspective" at the Annual Meeting of the American Educational Research Association (AERA), New York, NY.
- Assessment and Validity 'In-Vivo'. (2017). Keynote Symposium address, the annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Prague, Czech Republic. [with Bryan Maddox, University of East Anglia, UK]
- The Interplay Between Survey Research and Psychometrics, with a Focus on Validity Theory. (2016). [with Jose-Luis Padilla, University of Granada, Spain]. Invited address at the American Statistical Association 2nd International Conference on Questionnaire Design, Development, Evaluation and Testing (QDET2), Miami, Florida, USA.
- Tides, Rips, and Eerie Calm at the Confluence of Data Uses, Consequences, and Validity. (2015). Plenary address, 'The Production of Data in International Assessments', Research Conference organised by the Laboratory of International Assessment Studies, Economic and Social Research Council (ESRC), University of East Anglia, Norwich, UK.
- Consequences, Side Effects and the Ecology of Testing: Keys to Considering Assessment 'In Vivo'. (2015). Invited plenary address, the annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Glasgow, Scotland.
- Inviting the Study of Consequences by Using Ethnographic-Psychometrics. (2015). [with Bryan Maddox] Plenary, Language Testing Research Colloquium (LTRC). Toronto, ON.
- Measurement Validity and Validation in the Social and Health Sciences: A Meditation on Where We Have Come From and the State of the Art Today. (October, 2014). Quantitative Methods Colloquium Series, Department of Psychology, York University, Toronto ON.
- Address at the Northeastern Educational Research Association, October 2011, Rocky Hill, CT. [with Ben Shear]
- Invited address at Catholic University of Milan, September 2011.
- Invited address at ETS, R&D Division, September 2010.
- Invited address at the 2010 International Conference on Outcomes Measurement (ICOM 2010), the US National Institutes of Health (NIH), Bethesda, MD September 1-3, 2010
- Invited address at the 2009 Firenze, Italy, meeting "Statistics, Knowledge and Policy: Understanding Societal Change", which was an Organisation for Economic Co-operation and Development (OECD) hosted Global Project in association with the Joint Research Centre (JRC) of the European Commission and the International Society for Quality of Life Studies (ISQOLS).
- The Messick career award address delivered in 2005.

Bibliography (incomplete listing)

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34. doi:10.1207/s15434311laq0201_1

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 4, 281-302.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement*, 2nd Ed., (pp. 443-507). Washington, D.C.: American Council on Education.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher, Vol. 36,* 449–455.

Forer , B., & Zumbo, B.D. (2011). Validation of Multilevel Constructs: Validation Methods and Empirical Findings for the EDI. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement, 103*, 231-265.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*, 207-215.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, *103*(2), 219-230.

Hubley, A. M., & Zumbo, B.D. (2013). Psychometric Characteristics of Assessment Procedures: An Overview. In Kurt F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology, Volume 1* (pp. 3-19). Washington, D.C.: American Psychological Association Press.

Bibliography (incomplete listing)

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Landy FJ. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.

Lissitz, R. W. (Ed.) (2009). *The Concept of Validity: Revisions, New Directions and Applications*. IAP - Information Age Publishing, Inc.: Charlotte, NC.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, 30*, 955- 966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

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.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Bibliography (incomplete listing)

Messick, S. (1998). Test validity: A matter of consequence. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: perspectives from the social and behavioral sciences* (pp. 35-44). Netherlands: Kluwer Academic Press. Special issue of the journal Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45 (1-3), 1-359. Netherlands: Kluwer Academic Press.

Sinnott-Armstrong, W., & Fogelin, R. (2010). *Understanding arguments: An introduction to informal logic* (8th ed.). United States: Wadsworth CENGAGE Learning.

Sireci, S. G. (1998). The construct of content validity. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: perspectives from the social and behavioral sciences* (pp. 83-117). Netherlands: Kluwer Academic Press. Special issue of the journal Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45 (1-3), 1-359. Netherlands: Kluwer Academic Press.

Sireci, S. G. (2009*).* Packing and Unpacking Sources of Validity Evidence: History Repeats Itself Again. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 19-38). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5-8, 13, 24.

Stone, J., & Zumbo, B.D. (2016). Validity as a Pragmatist Project: A Global Concern with Local Application. In Vahid Aryadoust, and Janna Fox (Eds.), *Trends in Language Assessment Research and Practice* (pp. 555-573). Newcastle: Cambridge Scholars Publishing.

Bibliography (incomplete listing)

Sinnott-Armstrong, W., & Fogelin, R. (2010). *Understanding arguments: An introduction to informal logic* (8th ed.). United States: Wadsworth CENGAGE Learning.

Woitschach, P., Zumbo, B.D., & Fernández-Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema, 31(2),* 194-203.

Zumbo, B. D. (Ed.) (1998). Validity theory and the methods used in validation: perspectives from the social and behavioral sciences. Special issue of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45 (1-3)*, 1-359. Netherlands: Kluwer Academic Press.

Zumbo, B. D. (2005). Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing (July 22, 2005). Samuel J. Messick Memorial Award Lecture, LTRC 27th Language Testing Research Colloquium, Ottawa, Canada.

Zumbo, B.D. (2007). Validity: Foundational issues and statistical methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of statistics, Vol. 26: Psychometrics*, (pp. 45-79). The Netherlands: Elsevier Science B.V..

Zumbo, B. D. (2009*).* Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Zumbo, B.D. (2014). What Role Does, and Should, the Test Standards Play Outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31-33.

Bibliography (incomplete listing)

Zumbo, B.D. (2017). Trending Away From Routine Procedures, Towards an Ecologically Informed 'In Vivo' View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives, 15:3-4*, 137-139.

Zumbo, B.D., & Chan, E.K.H, (Eds.) (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. New York: Springer.

Zumbo, B. D., & Forer, B. (2011). Testing and Measurement from a Multilevel View: Psychometrics and Validation. In James A. Bovaird, Kurt F. Geisinger, & Chad W. Buckendahl (Editors). High Stakes Testing in Education - Science and Practice in K-12 Settings, (pp.177-190) *[Festschrift to Barbara Plake]*. American Psychological Association Press, Washington, D.C..

Zumbo, B.D., & Hubley, A.M. (2016). Bringing Consequences and Side Effects of Testing and Assessment to the Foreground. Assessment in Education: *Principles, Policy & Practice, 23*, 299–303.

Zumbo, B. D., & Hubley, A.M. (Eds.). (2017). *Understanding and Investigating Response Processes in Validation Research*. New York, NY: Springer.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Astivia, O.L.O. & Ark, T.K. (2015). A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding. *Language Assessment Quarterly, 12,* 136-151.

Zumbo, B.D., & Padilla, J.L. (2020). The Interplay between Survey Research and Psychometrics, with a Focus on Validity Theory. In P.C. Beatty, D., Collins, L., Kaye, J.L. Padilla, G. Willis, and A. Wilmot, (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Te*sting (pp. 593-612). Hoboken, NJ: Wiley.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.