VALIDITY THEORIES, FRAMEWORKS AND PRACTICES IN USING TESTS AND MEASURES
An Over-The-Shoulder Look Back at Validity While Also Looking To The Horizon

January 19, 2023

UNIVERSITÀ CATTOLICA DEL SACRO CUORE
Dipartimento di Psicologia
FORmazione MEtodologica (FORME)

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

Vancouver, Canada

Citation:

Zumbo, B.D. (2023, January 19). *Validity Theories, Frameworks And Practices In Using Tests And Measures: An Over-The-Shoulder Look Back at Validity While Also Looking To The Horizon* [Invited Address]. Ciclo Formazione Metodologica (FORME), Dipartimento di Psicologia, Università Cattolica Del Sacro Cuore, Milano, Italy. URL: https://brunozumbo.com/?page_id=31

Topics in Today's Webinar

1. Introduction
2. Concepts and histories of validity
3. Bridging concepts and Practices
   - 3A. My perspective on construct theories
   - 3B. Explanation centered validity and validation practices – On the many ways of being human [OPTIONAL, IF TIME ALLOWS]
4. Concluding remarks [OPTIONAL, IF TIME ALLOWS]

References and end material

Section 1

INTRODUCTION

[SOME COMMON LANGUAGE AND UNDERSTANDING ABOUT MEASURES, TESTS, AND ASSESSMENTS]

1) General remarks
2) An example to motivate our discussion
3) Items as building blocks
4) Transition to section #2 - 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 may be meaningless.

- Throughout this presentation, the terms measure, instrument, test, assessment, questionnaire, 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 to name but a few of the common applications.

• 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)

Example of a psychological measure

- A psychological test or measure may be viewed as a set of self-report questions (also called “items”) whose responses are then scored and aggregated in some way to obtain a composite score.

- In many psychological measures (e.g., attitudinal measures), there are not “correct” or “incorrect” responses, per se., rather we are dealing with compelled self-report responses.

Example of a psychological measure

This is the ***Center for Epidemiologic Studies Depression Scale (CES-D)***, which is a measure of depressive symptomology- an index of current feelings of general depression.

The higher the score on the measure, the greater the level of depressive symptomology.

- Note that items 4, 8, 12, and 16 need to be 'reverse coded' before one can compute the total score.

For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)
1 = some or a little of the time (1-2 days)
2 = occasionally or a moderate amount of time (3-4 days)
3 = most or all of the time (5-7 days)

	not even 1 day	1-2 days	3-4 days	5-7 days
1. I was bothered by things that usually don't bother me.	0	1	2	3
2. I did not feel like eating; my appetite was poor.	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3
4. I felt that I was just as good as other people.	0	1	2	3
5. I had trouble keeping my mind on what I was doing.	0	1	2	3
6. I felt depressed.	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
8. I felt hopeful about the future.	0	1	2	3
9. I thought my life had been a failure.	0	1	2	3
10. I felt fearful.	0	1	2	3
11. My sleep was restless.	0	1	2	3
12. I was happy.	0	1	2	3
13. I talked less than usual.	0	1	2	3
14. I felt lonely.	0	1	2	3
15. People were unfriendly.	0	1	2	3
16. I enjoyed life.	0	1	2	3
17. I had crying spells.	0	1	2	3
18. I felt sad.	0	1	2	3
19. I felt that people dislike me.	0	1	2	3
20. I could not get "going".	0	1	2	3

Note: Items 4, 8, 12, and 16 were reverse coded.

Example of a psychological measure

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

- The essential features therefore are:
  - a series of questions to which an individual responds, and
  - a composite score that arises from scoring the responses to these questions.

- In short, we are usually talking about a set of questions whose responses are aggregated into a composite or overall score.
  - The key point here is that the composite (i.e., scale) score is not depression itself but rather an observable indicator of depression -- or more accurately, the composite score is an indicator of depressive symptoms.

Items are the building blocks

- Tests and measures **come in various forms and lengths** and measure many psychological phenomena, such as **knowledge** of some domain or **psychological characteristics (attributes)** of test takers.
  - Despite their varied forms and lengths, all assessments share the property of being **composed of a series of items, tasks, or questions** to which an individual responds.
- Simply stated, **items are the building blocks** of an assessment.

- Item analysis can be used in the test development process to aid in item revision and later to help understand why a test shows specific levels of reliability and validity.

Transition to Section 2

- Our primary goal today is to describe theories and 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:

Methodology ↔ Method

Validity ↔ Validation

Transition to Section 2

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.

Transition to Section 2

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 & HISTORIES OF VALIDITY

1) Validity: An over-the-shoulder look back
2) Four periods of historical focus
3) Eight conceptualizations 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

1900 → 2000 …

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 (dominance during the **1960s to end of 1990s**, *but peaked in the mid 1970s, still on-going)*

- 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, Zumbo, Maddox, & Care, in press)

Section 3

BRIDGING CONCEPTS & PRACTICE

- Construct theories
- Explanation centered validity and validation practices

Section 3A

A DEEP DIVE INTO CONSTRUCT THEORIES

1) Describing the encounter of a person and an item (task) (main message)
2) Tests, Items, and Constructs
3) Growing prominence of argument-based approaches
4) Explicit synthesis of construct theories and argument-based approaches

Describing the encounter of a person and an item (task)

Putting the psychology back in psychometrics.

Main Message: *Describing the **encounter** of a person and an item (or task)*

- Today we will will continue to focus on statistical ideas and reasoning, but mixed methods are continuing to grow in prominence (e.g., Benítez, Van de Vijver, & Padilla, 2022; Padilla & Benitez, 2014)
- As much as possible we will rely on research methods and data tools motivating psychological interpretations.

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144.

Benítez, I., Van de Vijver, F., & Padilla, J. L. (2022). A Mixed Methods Approach to the Analysis of Bias in Cross-cultural Studies. Sociological Methods & Research, 51(1), 237–270.

Items, test scores, and constructs (again)

- Tests and measures **come in various forms and lengths** and measure many psychological phenomena, such as **knowledge** of some domain or **psychological characteristics (attributes)** of test takers.
  - Despite their varied forms and lengths, all assessments share the property of being **composed of a series of items, tasks, or questions** to which an individual responds.
- Simply stated, **items are the building blocks** of an assessment.

- Item analysis can be used in the test development process to aid in item revision and later to help understand why a test shows specific levels of reliability and validity.

Items, test scores, and constructs

Individuals' responses to items do double duty

An individual's **responses to the items on an assessment are used to make inferences** about the individual's level of the psychological attribute being measured, most commonly through creating a score reflecting the individual's level of the psychological characteristic (or of the knowledge).

- What makes things particularly thorny is that an individual's **responses to the items on an assessment are also used to make inferences** about the about the quality of the series of items, tasks, or questions to which an individual responds.

Items, test scores, and constructs

- It should be noted that many authors refer to construct validity as the most important characteristic of a test, but it is seldom defined.
- A clear statement of what a construct is and the logic of construct validation was presented by Cronbach and Meehl (1955).

Items, test scores, and constructs

Cronbach and Meehl (1955) wrote:

> *A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct.*
>
> *We expect a person at any time to possess or not possess a qualitative attribute . . . or to possess some degree of a quantitative attribute . . . Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability).*
>
> *The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or in a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test . . .* (p. 247)

Items, test scores, and constructs

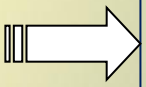
- Please note that in test develop or revision our research efforts are directed toward providing an evidential basis (i.e., empirical support) for making at **least two kinds of inferences from our test scores**.

Item 1. ….. score (0, 1, 2, 3)
Item 2. ….. score (0, 1, 2, 3)
Item 3. ….. score (0, 1, 2, 3)
⋮
Item k ….. score (0, 1, 2, 3)
Observed score _____

The observed score is often the sum or average of the item scores.
Total score is also called the observed test score, observed score, or composite score.

The target value of the observed test score is called the true score in (CTT), the universe score in generalizability theory, and the latent variable in item response theory (IRT) or factor analysis.

In all cases, the item responses are manifestations of an unobserved variable in common among the items

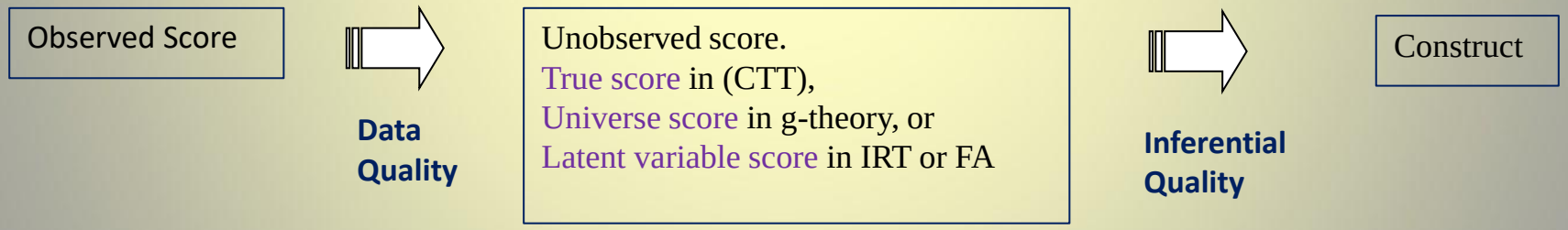


Construct or attribute

Construct is meaningfully related to other constructs to form a theory of what we intend to measure. And it can be verified by empirical studies.

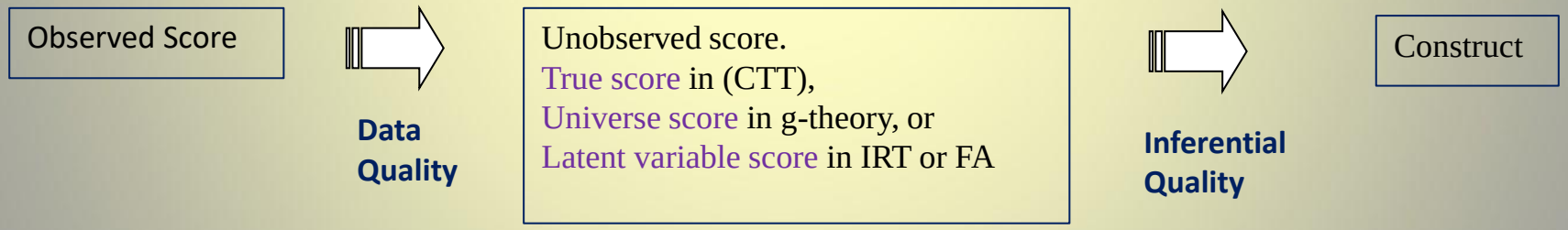
Items, test scores, and constructs

- The direction of the arrow are meant to suggest the direction when we are validating the inferences (claims) we wish to make from test scores.



Items, test scores, and constructs

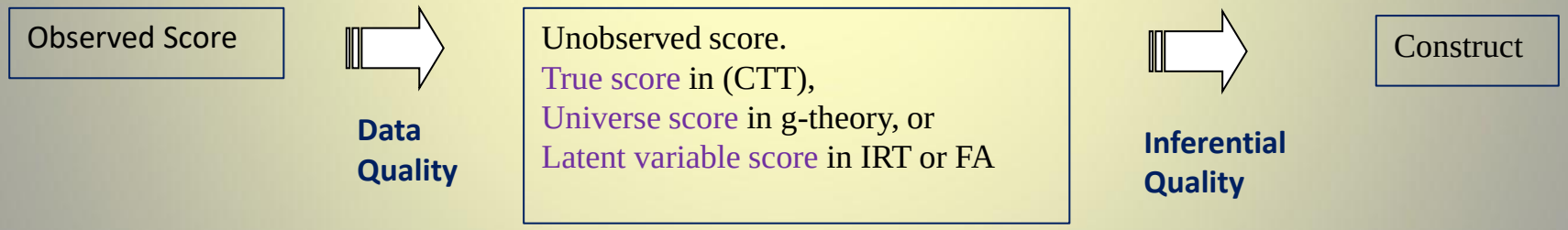
- As a by-product of the statistical models of psychometric theory, there is no such thing as a predicted score on a construct.
- The relationship is not the kind where you can predict a construct score.



Items, test scores, and constructs

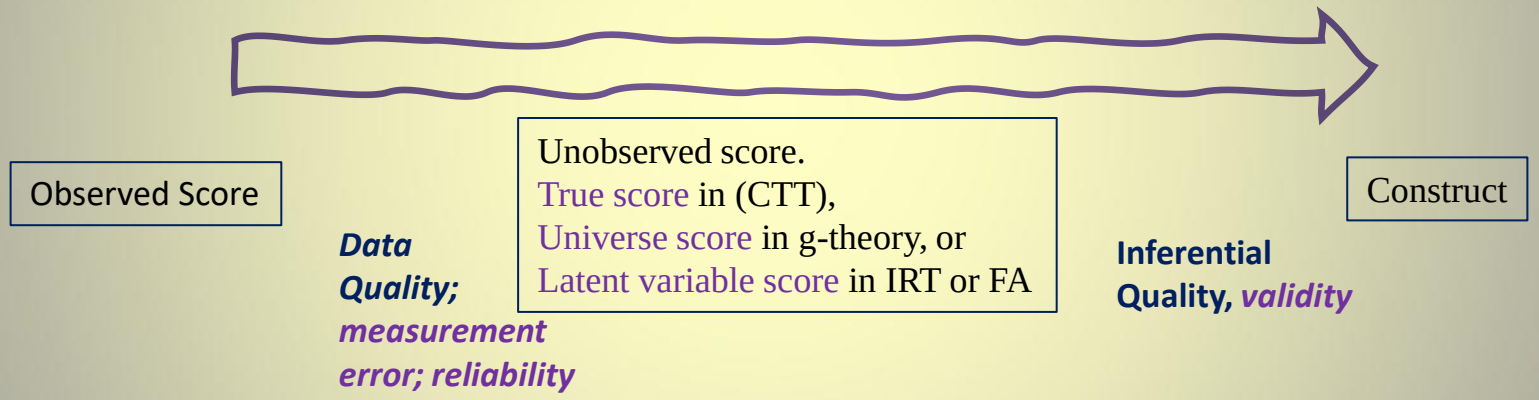
– One can imagine two parts stages:

1. The data quality stage (depicted by the first arrow on the left) is from the observed variables to latent unobserved variables.
2. The inferential quality stage (depicted by the arrow on the far right) is for the inference from the latent variable to claims about the construct.



Items, test scores, and constructs

– Validity: If we follow this arrow, you can see that when we use test scores we are, in essence, traveling from the observed score to claims about the test taker in terms of the intended construct (or attribute).



It should be clear at this point that the data quality (reliability) has an effect on the validity. An unreliable assessment will have limited validity.

Items, test scores, and constructs

Because constructs are not the same as the unobserved (true, universe, or latent) variables we can have a case of **construct underrepresentation** or **construct irrelevant variance**.

Construct **underrepresentation** occurs when a test does not adequately measure all aspects of the construct of interest. ***Narrowing the construct has impact on the reliability and validity of the operationalized measurement***.

Construct **irrelevant variance** as a source of invalidity

- A guiding question for construct irrelevant variance is: to what extent are we measuring our attribute of interest with the test or assessment?

Items, test scores, and constructs

As a source of invalidity, **construct underrepresentation** negatively affects the soundness of the test score interpretation and any inferences or claims made from the test score.

- … may also impact on how relevant the test is for a target population of test takers and/or it may have value implications.
- When construct underrepresentation is found to contribute to social consequences of test use, the construct and/or the test may need to be revised or adapted.
- For example, in cross-cultural comparisons, it is crucial to ask whether a new cultural group conceives of or values the construct in the same manner as the original test development group. The answer to this question reveals how much the obtained scores reflect construct underrepresentation and/or construct irrelevancy

Items, test scores, and constructs

Construct irrelevant variance as a source of invalidity

– A guiding question for construct irrelevant variance is: to what extent are we measuring our attribute of interest with the test or assessment?

Response: The psychometric models and approaches I have developed embody statistical and psychometric models, and an ecological model of item and test performance (Zumbo et al., 2015).

– By observing the testing situation, we hope to identify clues about the way the test is constructed, understood and performed as a social occasion. Bringing the psychology back into psychometrics!

Items, test scores, and constructs

Construct irrelevant variance as a source of invalidity

- The central question driving the study of construct irrelevant variance is:
  - To what extent might we be measuring, unintentionally, other (un)important constructs that are not meant to be included in our inferences of our attribute, such as, conformity to expected cultural norms (e.g., related to, for example, multiculturalism, ethnicity, gender identity, and gender roles)?
  - The gender bias of the "crying" item (#17) of the CESD is an example of how construct irrelevant variance may be a source of construct invalidity (Gelin & Zumbo, 2003) .

Gelin , M. N., & Zumbo, B. D. (2003). DIF results may change depending on how an item is scored: An illustration with the Center for Epidemiological Studies Depression (CES-D) scale. *Educational and Psychological Measurement, 63*, 65-74.

Emergence of argument-based approaches

The emergence of argument-based approaches to test validation

Argument-based approaches to test 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).

Arguments in Validity and Validation

- This 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.
  - 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 plausibility 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).

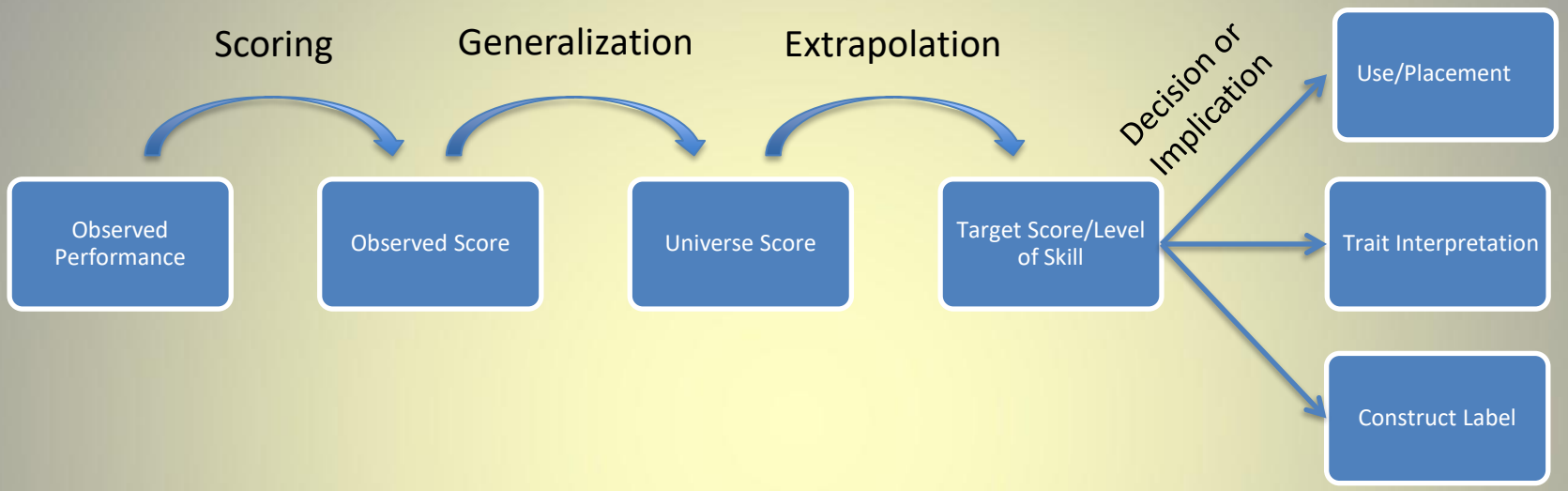
Arguments in Validity and Validation

Cronbach (1988), Kane (1992, 2006), Shepard (1993) and others advocate using *argument* 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.

Kane's Argument-based Approach to Validation



Notes:

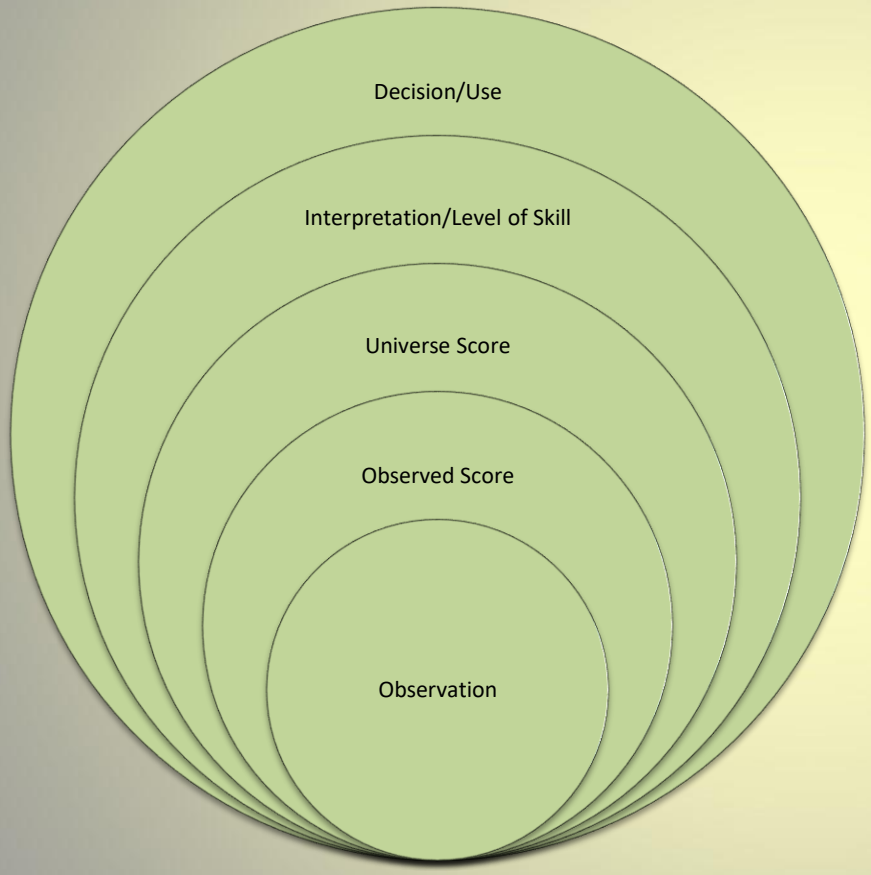
Different forms of interpretive arguments.

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



Please note:

Influence of psychometric G-theory.

Connection to Zumbo's DLD Framework (Zumbo, 2007)

Competency vs construct.

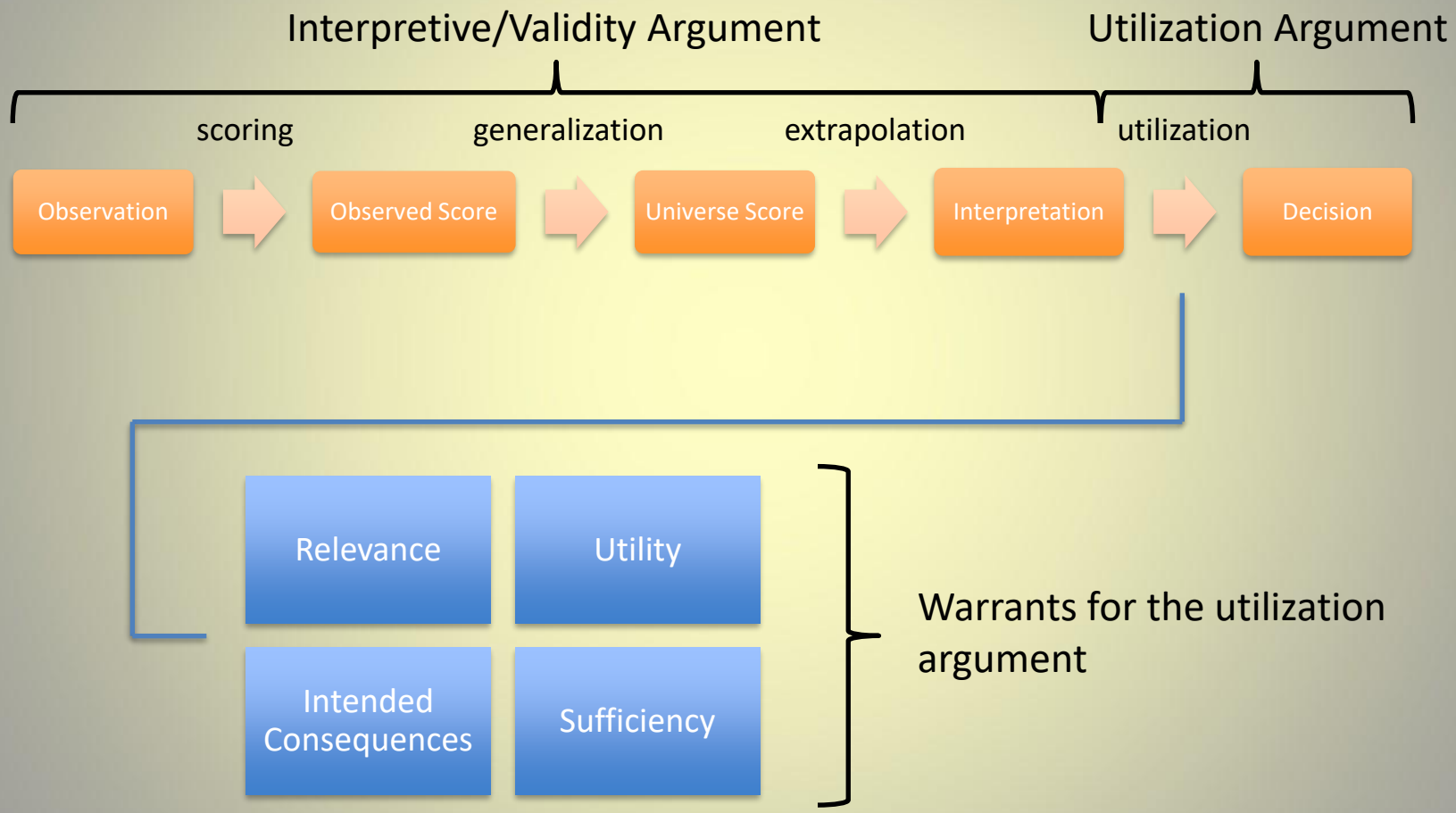
Zumbo, B.D., & Shear, B. (2011) *The Concept of Validity and Some Novel Validation Methods.* Presentation to the 42nd Annual Conference of the Northeastern Educational Research Association (NERA), Connecticut, USA.

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.

Synthesis of construct theories and argument-based

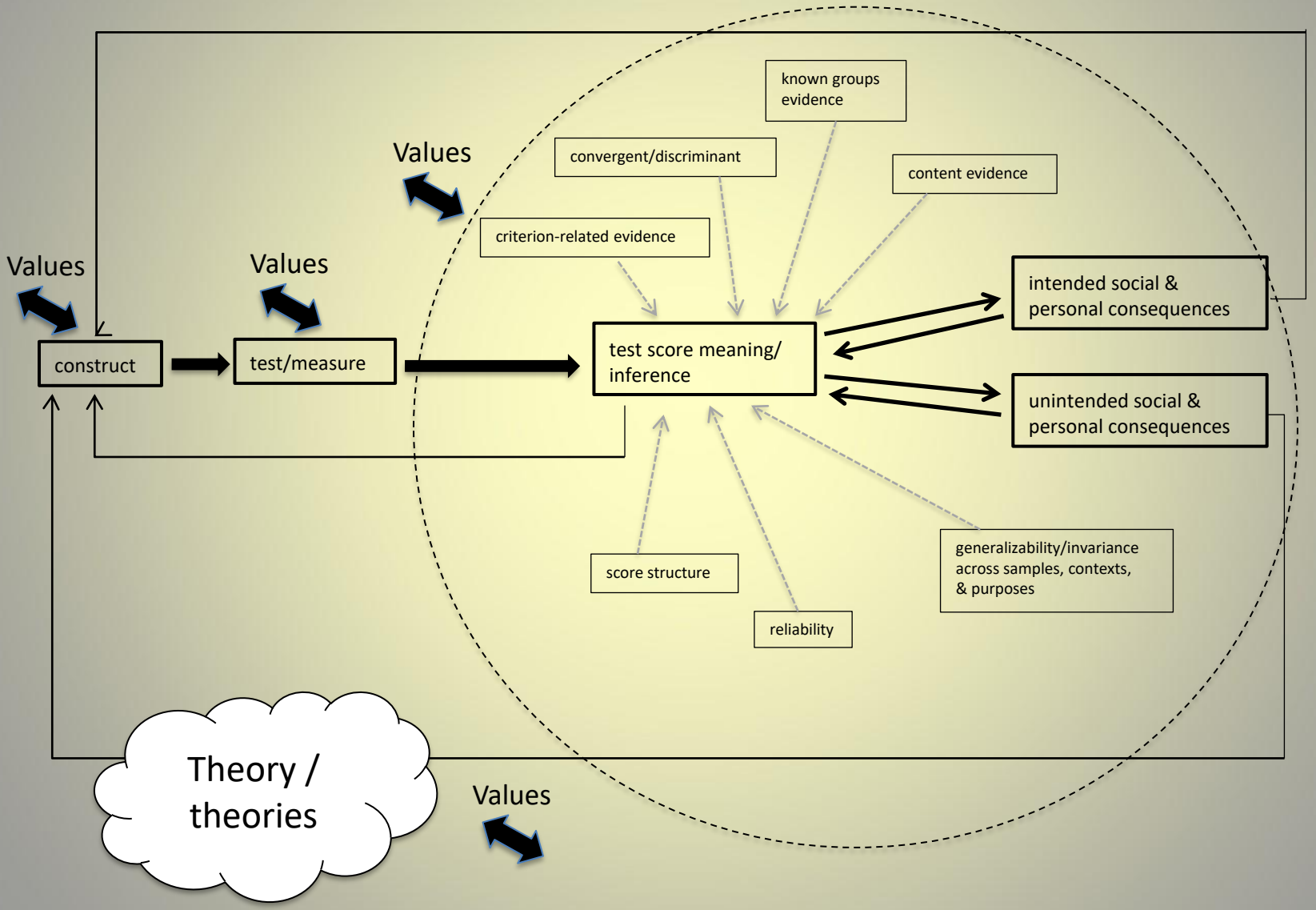
(More explicit) Synthesis of construct theories and argument-based approaches to validation

Synthesis of Construct Theories with Argument-based Approach

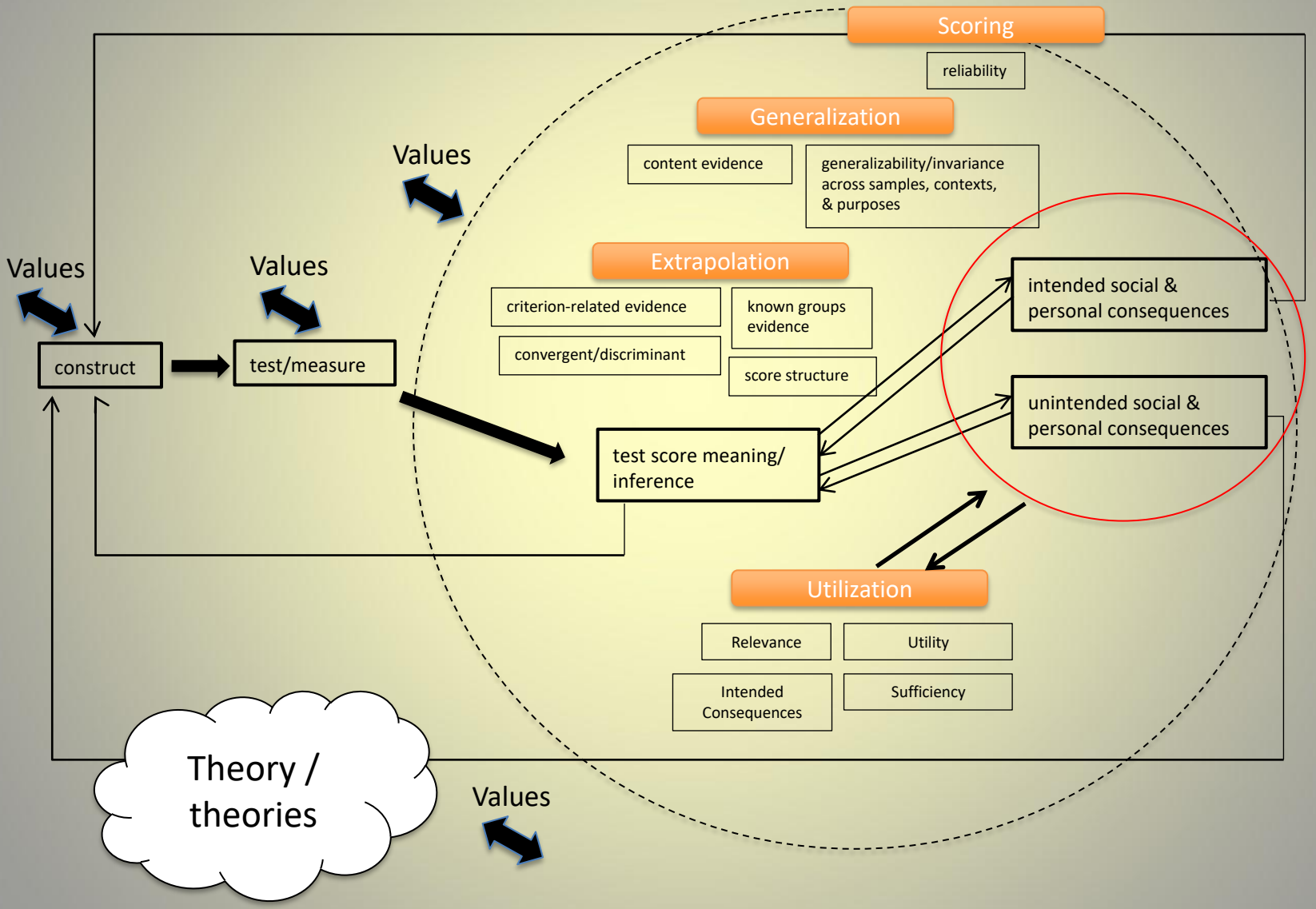
- After considering these argument-based approaches, we can now return and consider how they fit with the construct theory approaches to validity.
  - Argument-based approaches have embraced certain construct theories, but they foreground competencies.
  - Let us consider where different forms of evidence might be used to support the interpretive argument.
  - Notice that this ends with an interpretation, rather than decision, but as this still raises issues about the consequences involved.
  - Adding the utilization aspects discussed by Lyle Bachman (2005) 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.

Let us start with Hubley and Zumbo's (2011) construct theory approach to validity & validation

Hubley and Zumbo's (2011) construct theory approach to validity & validation



Explicit Synthesis of Construct Theories & Argument-based Approaches



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 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 3B

EXPLANATION CENTERED VALIDITY AND VALIDATION PRACTICES

Bridging concepts and practices

1) A guiding principle: the many ways to being human
2) Situating the challenges within my view of test / assessment validity and validation practices?
3) Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice
4) Fairness & Equity: Ecological Model of Item and Test Responding

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. (2022, November 9). *Repositioning construct validity theory: From nomological networks to pragmatic theories, and their evaluation by explanatory means*. https://doi.org/10.31234/osf.io/k54b6

Haig, B. (2012) From Construct Validity to Theory Validation, Measurement: Interdisciplinary Research and Perspectives, 10:1-2, 59-62, DOI: 10.1080/15366367.2012.681975

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10*, 371–388.

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.

On the Many Ways of Being Human

- A key principle, as I see it, is that there are <u>many ways to be human</u>.
- Over the last 30 years my experience has been that the **field of psychometrics** has tended to go into a **moral panic over gender identity, gender expression, and aspects of cultural expression**.
  - At the core of my theorizing and the methods I develop and/or advocate aim to challenge that view and aim to honor the many ways of being human and capturing the human experience.

On the Many Ways of Being Human

To **what extent** are we **measuring our attribute or competency of interest** with the test, assessment, or survey in use?

Response: My approaches embody **statistical and psychometric** models, an **ecological model of item and test performance.**

> By observing the testing situation, we hope to identify clues about the way the test is constructed, understood and performed as a social occasion.

Addey, C., Maddox, B. & Zumbo, B.D. (2020) Assembled validity: rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs). *Assessment in Education: Principles, Policy & Practice, 27:6*, 588-606. DOI: 10.1080/0969594X.2020.1843136.

On the Many Ways of Being Human

To what extent might we be measuring, unintentionally, other (un)important constructs that are not meant to be included in our inferences of our attribute or domain of interest, such as, conformity to expected cultural norms (e.g., related to, for example, multiculturalism, ethnicity, gender identity, and gender roles)?

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4(2)*, 223–233.

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.

Addey, C., Maddox, B. & Zumbo, B.D. (2020) Assembled validity: rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs). Assessment in Education: Principles, Policy & Practice, 27:6, 588-606. DOI: 10.1080/0969594X.2020.1843136.

On the Many Ways of Being Human

As Fox (2003) points out, “From an ecological perspective, individuals do not exist as isolated units; rather they are dynamic, socially embedded, and defined by a network of relationships— perceived or actual—occurring in time” (p. 22).

Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing, 3(1)*, 21–48.

Section 4

SOME CONCLUDING REMARKS

Validity as Contextualized and Pragmatic Explanation Validation

- First, please note that I am separating "validity" from the "process or elements of validation" (Zumbo, 2007).
  - Validity involves establishing an explanation for the observed score variation. This is, of course, an old tradition in philosophy of science.
    - In the mid-1950s Cronbach and Meehl brought this idea most clearly to the measurement community by drawing on a form of Hempel-Oppenheim deductive nomological model of explanation for measurement validity in what Cronbach and Meehl called construct validity and a nomological network.

Validity as Contextualized and Pragmatic Explanation Validation

- My own work in the last approx. 20 years has challenged that view.
  - Part of the problem with the Cronbach and Meehl approach is that it was rooted in a neo-behaviorist tradition of conflating explanation and confirmation as well as suffering of Michael Scriven's later clearly articulated concern with the Hempel-Oppenheim deductive nomological model of explanation.

Validity as Contextualized and Pragmatic Explanation Validation

- I have espoused a different explanatory view. I have my roots more firmly in a pragmatic approach and particularly an inference to the best explanation like strategy.
- Second, in separating "validity", per se, from the validation process then I have a clearly sense of the role of social consequences, justice, and fairness in the validation process …. and separate from validity itself. This recognized measurement as a power tool in public debate and public policy.

Validity as Contextualized and Pragmatic Explanation Validation

- The basic idea underlying my explanatory approach is that, if one could understand why an individual responded a certain way to an item or scored a particular value on a scale, then that would go a long way toward bridging the inferential gap between test scores (or even latent variable scores) and constructs.

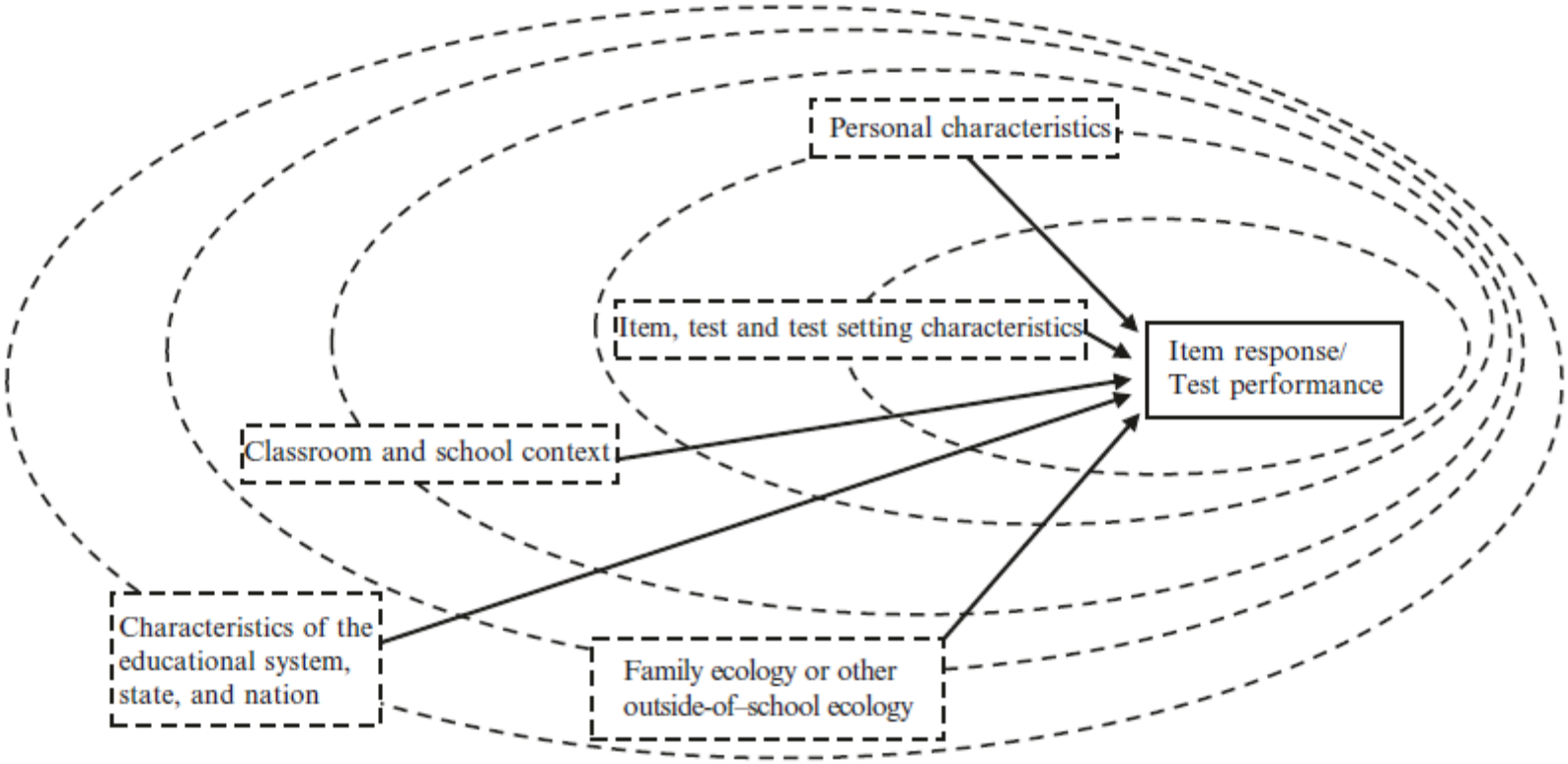
Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) The Concept of Validity: Revisions, New Directions and Applications, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Validity as Contextualized and Pragmatic Explanation Validation

- According to this view, validity per se, is not established until one has an explanatory model of the variation in item responses and/or scale scores and the variables mediating, moderating, and otherwise affecting the response outcome.
- 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.

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.

Ecological Model of Item and Scale Responding



- 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.
- Chen, M.Y., & Zumbo, B.D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 53-68). New York, NY: Springer.

Validity as Contextualized and Pragmatic Explanation Validation

- Overlooking the importance of explanation in validity we have, as a discipline, focused overly heavily on the validation process and as a result we have lost our way.
  - This is not to suggest that the activities of the process of validation, such as correlations with a criterion or a convergent measure, dimensionality assessment, item response modeling, or differential item or test functioning, are irrelevant or should be stopped.

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 as Contextualized and Pragmatic Explanation Validation

- Quite to the contrary, the activities of the process of validation must serve the definition of validity.
  - My aim is to re-focus our attention on why we are conducting all these psychometric analyses: that is, to support our claim of the validity of our inferences from a given measure.

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 as Contextualized and Pragmatic Explanation Validation

- For example, as Zumbo (2007) highlighted conducting test and item bias is not just about protecting a test developer or test user against lawsuits.
  - Conducting test and item bias is also a statistical methodology that ferrets out invalidity that distorts the meaning of test results for some groups of examinees and thus establishes the inferential limits of the test.

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

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 as Contextualized and Pragmatic Explanation Validation

- One of the limitations of traditional quantitative test validation practices (e.g., factor-analytic methods, validity coefficients, and multitrait-multimethod approaches) is that they are descriptive rather than explanatory.
  - The aim of my explanatory approach is to lay the groundwork to expand the evidential basis for test validation by providing a richer explanation of the processes of responding to tests and variation in test or items scores and hence promoting a richer psychometric theory-building.

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.

Ecological Model of Item and Test Responding

We believe that these **richer ecological variables** have been **largely ignored** in relation to explanations for (and causes of) DIF because of the focus on test format, content, cognitive processes, and test dimensionality that is pervasive in the second generation of DIF.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4(2)*, 223–233.

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.

What my approach to fairness and equity implies

- Traditional views follow a "**social address**" model of criterion prediction and group differences.
  - This spills over in to test validation; group differences.
- In using the common "social address" approach to group comparisons, **classification into groups** might be **confused with fixed biological or ethnic classification**.

  As John Stuart Mill (1848) wrote:

> Of all the vulgar modes of escaping the consideration of the effect of social and moral influences on the mind, the most vulgar is attributing the diversities of conduct and character to inherent natural differences. (p. 319).

What my approach to fairness and equity implies

- In a series of chapters and papers from 1998 to 2021, I have made the case that the aim is: **identifying the determinants (or explanatory theory)** of task / item / test score variation … **the explanation is the basis of any strong validity claims**.

- I take an **ecological systems approach**

- Most research on response processes focuses on cognitive factors.
  - We have taken a **broader view of response processes** proposed by Zumbo & Hubley (2017) and embrace the notion of **assessment 'in vivo'** to shine a spotlight on test-takers' behaviour, stance, gesture, motivation, and affect besides cognition.

So… Now What?

- **Do not skirt these hurdles**.

- There are now **sophisticated methodologies to tackle these challenges**. (Complex models of the impact of the ecological model of item and test responding)
- A well thought out research program needs to be established and funded to address these issues.
- We can use assessment outcomes, but we need to get going on this research program to support the inferences we intend to make with them.

Implications for studies of fairness and equity in testing

- Examples of issues in a validation research agenda:
  - One needs to **investigate if and how race, gender, and culture as related to the object of measurement** (i.e., individual students or communities) may shape and alter the inferences one makes from testing or assessment outcomes.
  - One needs to **investigate the role of various levels of measurement (e.g., individual, community, neighborhood, city, province, region) in the inferences**. For example, is one measuring learning or perhaps different secondary (or primary) dimensions at the various levels of analysis. Including the predictive nature at these various levels. (see, Zumbo et al, 2017)

Zumbo, B.D., Liu, Y., Wu, A.D., Forer, B., & Shear, B.R. (2017). National and International Educational Achievement Testing: A Case of Multi-Level Validation Framed by the Ecological Model of Item Responding. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 341-362). New York, NY: Springer.

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.

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.