CHAPTER 4

VALIDITY AS CONTEXTUALIZED AND PRAGMATIC EXPLANATION, AND ITS IMPLICATIONS FOR VALIDATION PRACTICE

Bruno D. Zumbo

ABSTRACT

This chapter has two aims: provide an overview of what I consider to be the concept of validity and then discuss its implications for the process of validation. I articulate an explanation focused view of validity that centers on a contextualized and pragmatic view of explanation—in essence, a contextualized and pragmatic view of validity. In the closing section of the chapter I describe the methodological implications of this view in terms of not assuming homogeneity of populations (from the Draper-Lindley-de Finetti framework) and allowing for multilevel construct validation, as well as the overlap between test validity and program evaluation.

The Concept of Validity, pages 65–82

> We are as sailors who are forced to rebuild their ship on the open sea, without ever being able to start fresh from the bottom up. Wherever a beam is taken away, immediately a new one must take its place, and while this is done, the rest of the ship is used as support. In this way, the ship may be completely rebuilt like new with the help of the old beams and driftwood—but only through gradual rebuilding. Otto Neurath (1921, pp. 75–76)

The philosopher Neurath's now famous nautical image in the quotation above is an important place to begin our voyage. There has been much discussion in the philosophy of science literature about the interpretations and implications of Neurath's analogy of scientific verification as the construction of a ship which is already at sea, but it certainly does highlight for us that over the nearly century of measurement work we, as a discipline, have built, rebuilt, re-visioned and otherwise restored and restocked the good ship *Validity* at sea. In this light, *The Maryland Validity Conference*, as it has now come to be called among many with whom I collaborate and correspond, and the proceedings for which this chapter is written, is a high mark in the nearly century-old history of *Validity*'s journey.

I have also chosen to open with Neurath's (1921) nautical quotation because I believe its message of ongoing building and rebuilding while at sea is one of the defining (and most complexifying) features of not just the concept of validity but of measurement validation. In short, it has been long recognized that activities of measurement validation are inextricably tied to theory building and theory testing so that one needs measures to help develop and test theory, but one cannot wait for the establishment of validity before one can get to the business of developing and testing theories. Likewise, almost by definition, measurement and testing are used ultimately for means such as the assessment of individuals for the ultimate aim of intervention or feedback, for decision-making, or for research and policy purposes. It is rare that anyone measures for the sheer delight one experiences from the act itself. Instead, all measurement is, in essence, something you do so that you can use the outcomes, and hence one cannot wait for validation to be completed before one gets to the matter of the use of the test and measurement outcomes. In short, the measurement enterprise is as close to Neurath's ship as one can imagine. That is, at the heart of validity and of validation is the matter of scientifically constructing, verifying, and appraising test score meaning as an on-the-fly activity that is conducted while the system is in operation.

I will consider the concept of "validity" for any kind of test or measure in social, behavioral, educational or health research, testing, or assessment settings. I believe that there is much more in common, than unique, among the various uses of tests and measures, that there is much to be gained by exploring this commonality, and that I wish to be a countervailing force to the creation of the various new disciplinary measurement sub-fields which act

as silos. This general objective has me focusing on a meta-theory of validity rather than a tailored context for only, for example, cognitive, educational, language, health, policy, or behavioral measures. My aim is to think broadly so as to embrace and show the relations among many of the prominent views of validity, with an eye toward an articulation of a novel framework.

With this broad objective in mind, the terms "item" and "task" will be used interchangeably. Furthermore, the terms "test," "measure," "scale," and "assessment" will be used interchangeably, even though "tests" are, in common language, used to imply some educational achievement or knowledge test with correct and incorrect responses or partial credit scoring, and "assessment" typically implies some decisions, actions, or recommendations from the test and measurement results and implies a more integrative process involving multiple sources of information. Finally, in the parlance of day-to-day social and behavioral researchers, clinicians, and policy specialists, tests may be referred to as valid or invalid, but it is widely recognized that such references are, at best, a shorthand for a more complex statement about the validity of inferences made about test scores with a particular sample in a particular context and, more often, are inappropriate and potentially misleading.

The purpose of this chapter is to provide an overview of what I consider to be the concept of validity and then discuss its implications for the process of validation. Due to space limitations relative to the breadth and scope of the task at hand, for some issues I will provide details whereas for others more general integrative remarks.

AN EXPLANATORY-FOCUSED VIEW OF VALIDITY

To continue Neurath's analogy, when one is at sea one always keeps an eye on where one is going, and from where one has come. Even a cursory glance of the research literature (see, e.g., Hubley & Zumbo, 1996; Kane, 2006; Zumbo & Rupp, 2004; Zumbo, 1998) will reveal that validity theory and practices have changed over the last century. In brief, the early- to mid-1900s were dominated by the criterion-based model of validity, with some focus on content-based validity models (Sireci, 1998). This view is perhaps best seen in Anastasi's (1950) characterization in her highly influential paper in the leading measurement journal at the time: "It is only as a measure of a specifically defined criterion that a test can be objectively validated at all. . . . To claim that a test measures anything over and above its criterion is pure speculation" (Anastasi, 1950, p. 67).

The early 1950s saw the introduction of, and move toward, the construct model with its emphasis on construct validity with a seminal piece by Cronbach and Meehl (1955). Likewise, in another seminal paper, Loevinger

(1957) highlighted the important point that every test underrepresents its construct to some degree and contains sources of irrelevant variance, if for no other reason than it is a test and not a criterion performance. Clearly then, the early- to mid-1900s in the history of validity reflected Psychology's focus on observed behavior and theories of learning, as well as its relatively recent break from psychoanalytic and introspective methods. In the 1960s, the precursors to what we now call the cognitive revolution of the 1970s could be clearly seen. The period post Cronbach and Meehl, mostly the 1970s to the present, saw the construct validity model take root and saw the measurement community delve into a moral and consequential foundation to validity and testing by expanding to include the consequences of test use and interpretation (Messick, 1975, 1980, 1988, 1989, 1995, 1998).

It is worth noting that a subtle, but important, shift occurred with Cronbach and Meehl's (1955) publication wherein the dominant view of measures changed from being "predictive devices" to being "signs." Not all psychological phenomenon allow for a criterion; that is, some psychological phenomenon are abstract and do not necessarily have a "prediction." Suddenly, by the 1950s to early 1960s, it was safe and respectable, again, to talk in the language of unobservables (e.g., constructs) and hence the nature of tests and measures changed implicitly. In light of this, I believe that the operationalism that rests at the core of the predictive model (prior to the 1950s) was de-emphasized by Cronbach and Meehl in favor of the nomological network as supporting meaningfulness—i.e., the meaningfulness of the scores produced by tests/measures as reflective of an unobserved phenomenon, the construct. It is important to note that validity continues to be deeply rooted in the notion of "individual differences" or disposition theory, as dispositional theory has evolved over the decades.

Although it has been controversial, one of the current themes in validity theory is that construct validity is the totality of validity theory and that its demonstration is comprehensive, integrative, and evidence-based. What becomes evident is that the meaning of "construct validity" itself has changed over the years and is being used in a variety of ways in the current literature. Arguably in its most common current use, construct validity refers to the degree to which inferences can be made legitimately from the observed scores to the theoretical constructs about which these observations are supposed to contain information. In short, construct validity involves generalizing from our behavioral or social observations to the *conceptualization* of our behavioral or social observations in the form of the construct. The practice of validation aims to ascertain the extent to which an interpretation of a test is *conceptually* and *empirically* warranted and should be aimed at making explicit any ethical and social values that overtly or inadvertently influence that process (Messick, 1995).

The term "construct validity" has therefore evolved to be shorthand for the expression "an articulated argument in support of the inferences made from scores." I will argue later in this section that construct validity has, from its introduction, been focused on providing an explanation for test scores; that is, the argument in support of the inferences is a form of an explanation. As we all know, there are strong and weak forms of construct validity (Kane, 2001). 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 criterion (or convergent) variables. In contrast to the weak form of construct validity, the strong form is based on a well-articulated theory and well-planned empirical tests of that theory. In short, the strong form is theory-driven whereas the weak form implies that a correlation with some criterion (or convergent measure) is sufficient evidence to use the test as a measure of that criterion.

In my view (e.g., Zumbo, 2005, 2007a), 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. I share the view with other validity theorists that validity is a matter of inference and the weighing of evidence; however, in my view, explanatory considerations guide our inferences. 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.

In essence, I see validation as a higher order integrative cognitive process involving every day (and highly technically evolved) notions like concept formation and the detection, identification, and generalization of regularities in data whether they are numerical or textual. 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. This is, in its essence, a form of inference to the best explanation.

Figure 4.1 depicts the four core elements of the integrative cognitive judgment of validity and the process of validation: validity, psychometrics, 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 psychometrics, but validity is shaped by these. Furthermore, the inferences are justified by the psychometric, social consequences, and utility but validity is something more because it requires the explanation.

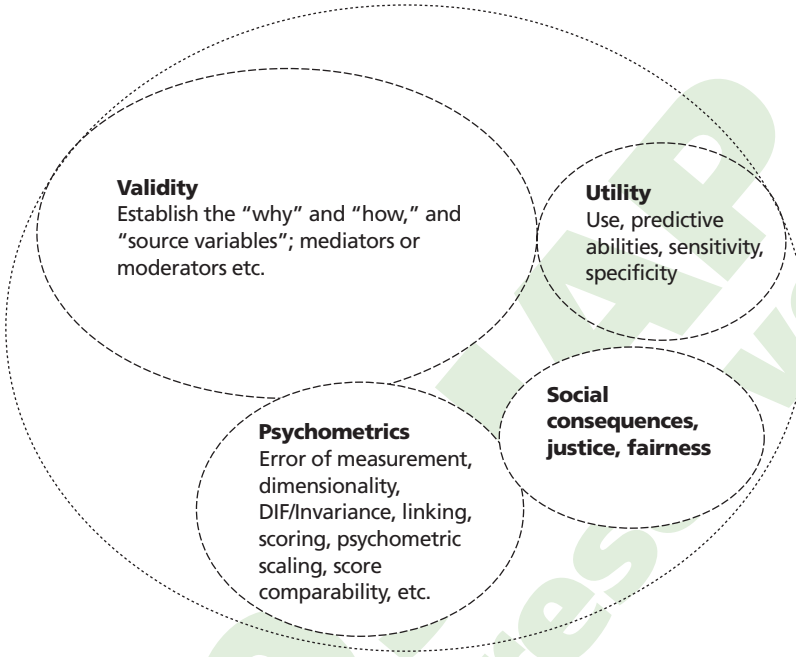

Figure 4.1 A depiction of the integrative cognitive judgment in the contextualized and pragmatic explanation view of validity and validation.

In short, Figure 4.1 shows that explanation is the defining feature of validity and hence supports the inferences we make from test scores. In terms of the process of validation, we can see in Figure 4.1 that the process of validation is distinct but is, itself, shaped by the concept of validity. The process of validation involves consideration of the statistical methods, as well as the psychological and more qualitative methods of psychometrics, work to establish and support the inference to the explanation—i.e., 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. The process of validation also includes the utility and evidence from test use such as sensitivity and specificity of the decisions (e.g., pass/fail, presence/absence of disease) made from test scores and predictive capacity (e.g., predictive regression equations), as well as the fourth element of social consequences. This latter element in the cognitive process depicted in Figure 4.1 has me clearly aligned with Messick (e.g., Messick, 1998) in that empirical consequences of test use and interpretation constitutes validity evidence in the validation process.

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. 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. 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. Quite to the contrary, the activities of the process of validation must serve the definition of validity. My aim is to refocus our attention on why we are conducting all of these psychometric analyses: that is, to support our claim of the validity of our inferences from a given measure. For example, as Zumbo (2007b) highlighted conducting test and item bias is not just about protecting a test developer or test user against lawsuits; it 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. 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 tests or items scores and hence promoting a richer psychometric theory-building.

Rereading Foundational Papers on Validity from the Explanatory-Focused View

Placing explanation as the driving element of validity is an interesting meta-theoretical place from which to reread some classic papers in validity with an eye to further explicating my view of validity as contextualized and pragmatic explanation.

From my point of view, Cronbach and Meehl (1955) were also focused on providing explanation; however, reflecting the individual differences psychological focus of the time period, the construct and the nomological network was the explanation. Not only were Cronbach and Meehl focusing on explanation but, as suggested by Cronbach himself, they were presenting a variant on the then relatively recently introduced "covering law mod-

el," also called the deductive-nomological (DN), approach to explanation. It is should be noted that Cronbach and Meehl did not wholly adopt a strict DN form of explanation; however, its DN essence and purpose are very clear. Cronbach (1971, p. 481) acknowledged the influences of the (logical positivist) DN approach and stated that, in particular, Hempel's work in the 1950s and 1960s, was the clearest description of the philosophical bases of construct validation, as articulated in Cronbach and Meehl. Cronbach (1971, p. 481) went on to state that Ernst Nagel's characterization of theoretical entities, what Cronbach and Meehl called 'constructs', as instrumental tools (rather than descriptive or realist) is essentially the position taken by Cronbach and Meehl in advocating construct validation of tests.

From its earliest form, the DN approach (Hempel & Oppenheim, 1948) is an idea that has lingered from the logical positivist tradition and has been shown to be problematic—see for example, Suppe (1977) as well as work by Scriven (1962), and many others. Borsboom et al. (2004) provided an excellent description of this point. The nomological network was essential to Cronbach and Meehl's (1955) view and provided a variation on the so-called covering laws needed for the DN approach to be useful.

It is noteworthy that the "hypothetical construct" position that is at the root of Cronbach and Meehl (1955) importantly also offers an alternative to operational definitions, or other such correspondence rules. Therefore, Cronbach and Meehl's nomological network can be thought of as a sidestep around operational definitions in the so-called "problem of theoretical terms" in philosophy. In essence, however, a significant weakness of Cronbach and Meehl's DN variant is that the common use of nomological networks in empirical social sciences is more in line with a concept-map than a system of laws relating the theoretical terms to each other and to observations. So, even if one were to accept Cronbach and Meehl's variant on the covering law view of explanation, the "nomological networks" of typical social science do not suffice to meet the necessary conditions for the explanation.

Furthermore, and most importantly from my point of view, the fundamental problem with Cronbach and Meehl's nomological network approach is that it attempts, like its DN forefather, to characterize explanation as context free. From my own perspective, the core of meaning-making of empirical data, and hence measurement validity, is the explanation of the observed score variation. My view is clearly in line with the essence of Cronbach and Meehl (1955), but I focus on the importance that the context provides in the explanation.

As one might imagine, in philosophy there have been competing ideas about what is and qualifies as an explanation. As an alternative to covering law views, explanation has also been associated with causation; an explanation is a description of the various causes of the phenomenon, hence to

explain is to give information about the causal history that led to the phenomenon. Salmon (1984, 1990, 1998) did a wonderful job of discussing and describing various views of scientific explanation. I will not attempt to go into the details of the various views, but suffice it to say that alternatives have been offered to the DN approach.

In this context of causation as explanation, it is important to acknowledge the seminal paper by Borsboom, Mellenbergh, and Van Heerden (2004). Although, in its core concepts, Borsboom and his colleagues' views share a lot in common with the view of validity I have espoused, I differ from their view on several important philosophical and methodological features. For example, Borsboom and his colleagues 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. Philosophically, this is, as the authors themselves acknowledge, a very tidy and simple idea that has a currency among researchers because it may well be implicit in the thinking of many practicing researchers. From my explanatory-focused view, relying on causality is natural and plausible and provides a clear distinction between understanding why a phenomenon occurs and merely knowing that it does—given that it is possible to know that a phenomenon occurs without knowing what caused it. Moreover, their view draws this distinction in a way that makes understanding the variation in observed item and test scores, and hence validity, unmysterious and objective. Validity is not some sort of super-knowledge of the phenomenon one wishes to measure, such as that embodied in the meta-theoretical views of Messick, Cronbach and Meehl, and myself, but simply more knowledge: knowledge of causes.

I am not fond of the exclusive reliance on "causal" models of explanation of the sort that Borsboom and his colleagues suggest. Their causal notions give us a restricted view of measurement because of the well-known objections to the causal model of explanation—briefly, that we do not have a fully adequate analysis of causation, there are non-causal explanations, that it is too weak or permissive, and that it undermines our explanatory practices. Also, like the covering law approaches, causal notions of explanation are typically aimed at context free explanations, which I do not accept as adequate for measurement purposes.

In addition to covering laws and causal views of explanation, there is a third broadly defined view of explanation that is often called the pragmatic approach and whose major proponents are, for example, Scriven and van Fraassen. According to Scriven (1962), in terms of context, all questions (and particularly all "why" questions) make presuppositions about what is known, and it is these presuppositions that supply the context of the answer. Therefore, an explanation is a body of information that implies that the phenomenon is more likely than its alternatives, where the information

is of the sort deemed "relevant" in that context, and the class of alternatives to the phenomenon are also fixed by the context. This approach highlights the importance of context to explanation.

Both Scriven and van Fraasen (1980) agree that scientific explanations are just specific kinds of explanations. Scriven offers criteria for a good explanation: they must be accurate/correct (i.e., are true), complete/adequate (e.g., give the appropriate causal connection), and relevant/appropriate/proper (i.e., cite the appropriate context). Van Fraasen's explanatory view makes explanation out to be what I refer to above, surrounding my description of Figure 4.1, as the overall cognitive evaluation, and what Scriven might refer to as unified communication. I also agree with Scriven and van Fraasen that scientific explanations (in our case, explanations as measurement validity) are just explanations wherein context is just as important in the science of measurement as it is in ordinary, day-to-day situations. As has been shown by several counterexamples in the philosophical literature (e.g., Scriven's explanation of the stain in a carpet is nothing more than "I knocked over the ink bottle," or van Fraasen's tower shadow), explanation is not merely a matter of logic, and nor, by extension, can it be simply a matter of causal explanation, but that it is a matter of pragmatics. Pragmatics refers to the aspects of language that reflect the practical circumstances in which we put it to use and, hence, the conditions or contexts that make some statements appropriate or meaningful.

These distinctions among the various views of explanation may appear subtle, but are important differences that play themselves out in both what is considered validity and the process of validation wherein certain methods, approaches and strategies are more naturally affiliated to one view than the other. For example, Cronbach and Meehl (1955) with their construct as covering law focus are closely aligned to multitrait-multimethod whereas Borsboom and colleagues are well suited with cognitive approaches, whereas I emphasize "why" questions and am more ecological, sociological and contextual in orientation. Also, because I do not rely so heavily on "constructs" and "dispositions" but also focus on situational and contextual elements, my approach more easily and naturally focuses on the multi-level notions in the next section of this chapter.

Therefore, the strength of Cronbach and Meehl's (1955) work is that they conceptualized validity as explanation rather than the prediction/correlation approach that dominated the first half of the 1900s. This is important because, in its essence, statistical prediction on its own does not necessarily impart understanding. Our ability to give explanations precedes any scientific knowledge. However, over and above the concern for the "nomological network" notion as really often being seen as simply a concept map, the major limitation of Cronbach and Meehl's contribution is that, like its

covering model explanatory parents, it treats explanations as context free. In so doing, it makes validity just about impossible to use because all measurement and testing are context bound. Instead, in the model I describe above in Figure 4.1, I wish to offer a context-bound sense of explanation and, hence, a context-bound view of validity.

As a side note, going back to the opening quotation from Neurath, one may ask: 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? The answer to this question is the same as the one for all the approaches to validity as explanation (e.g., it applies also to Cronbach and Meehl); that is, explanation is a regulative ideal therefore one can start (cautiously) using the measure as one gains a deeper understanding and explanation, but that the stakes for the measurement use should guide this judgment. What I am suggesting is that psycho-social, policy, and health studies research use the framework I describe surrounding Figure 4.1 to take on a robust and integrative research agenda in which the bounds and limitations of the inferences we can make from scores becomes a core task of the research agenda with an aim to providing a contextualized and pragmatic explanation of the test and/or item score variation.

IMPLICATIONS OF THE VALIDITY AS CONTEXTUALIZED AND PRAGMATIC EXPLANATION FOR VALIDATION PRACTICE

It is not sufficient, I believe, to just offer a new conceptualization of validity. Rather, one needs to explore the implications for day-to-day practice. In this final section, I aim to draw out the implications of my view of validity for the processes of validation. Note that my view of validity as contextualized and pragmatic explanation allows for all the methods currently used in validation research and also brings to the forefront, and out of the shadows, several interesting validation approaches that I wish to highlight. Due to space limitations, I will only be able to say a few words about each of these validation approaches, the Draper-Lindley-de Finetti framework and how it shines a light on modeling sample heterogeneity, a multi-level view of measurement, and the overlap between program evaluation and validation. See Zumbo (2007a) for a description of other statistical methods appropriate for the explanatory view of validation, and particularly the Pratt indices and variable ordering as tools in explanatory statistical modeling.

The Draper-Lindley-de Finetti Framework

The Draper-Lindley-de Finetti (DLD) framework of measurement validity (Zumbo 2007a) provides a useful overview of the assumptions that must be tested to validate the use of a psychometric tool. According to Zumbo's DLD framework, measurement problems and sampling problems must both be examined when assessing measurement validity. Measurement problems include those problems pertaining to the exchangeability of the observed and unobserved items (the items you have versus the items you wish you had) whereas sampling problems refer to the degree to which the measurement structure is appropriate for all respondents (i.e., equivalent across different sampling units in the target population). Although measurement problems can be examined by testing the extent to which a particular factor analysis solution fits the data of a sample as a whole, an examination of sampling problems involves determining the extent to which the factor analysis solution is appropriate for all respondents. Zumbo (2007a) referred to this as "the exchangeability of sampled and unsampled units (i.e., respondents) in the target population" (p. 59). This aspect of measurement validation relates to the degree to which individuals interpret and respond to items in a consistent and comparable manner. The exchangeability of sampling units is a necessary condition for the generalizability of inferences made about the measurement structure of a particular instrument.

The DLD framework brings the matter of sample homogeneity to the forefront. This is an important issue for all model-based measurement (and particularly item response theory). In essence DLD highlights that model driven applications, like item response theory and, perhaps, computer adaptive testing, require that the sample is homogeneous with respect to the measurement model. Therefore, for model-based measurement practices, the model assumptions (such as unidimensionality and sample homogeneity) are part of the validity concerns.

As Zumbo (2007a) highlighted, contemporary measurement theory based on latent variable models hold central the notion of invariance. Invariance implies that the model fit in all corners of the data (see Rupp & Zumbo, 2003, 2004, 2006). Invariance, therefore, is a guiding principle of, and an ideal for, much of contemporary model-based measurement theory. Invariance, in essence, carries with it the covering law, logical positivist, notion of context free measurement. However, from my contextualized and pragmatic view of validity, the aim should be to always take the context into account in measurement rather than to wash it away.

From a statistical point of view, this implies, then, that psychometric modeling should explore and allow for latent class and mixture models. My view of validity can be seen as a foundation and focus for Muthen's program of research which since at least the mid 1980s has been developing a class

of methods to model population heterogeneity. Muthen and his colleagues have created statistical theory and software (MPlus) to address the "why" question of validity highlighted in the description surrounding Figure 4.1, above (e.g., Muthen, 1985, 1988, 1989; Muthen & Lehman, 1985; Muthen, Kao, & Burstein, 1991). This class of approaches, which exploits, among other things, the multiple-indicators multiple causes structural equation model, and how this model relates to item response theory. As Zumbo (2007b) noted, one way of conceptualizing Muthen's work is that it is a merging of modeling item responses via contingency tables and/or regression models and item response theory frameworks. An essential feature of the Muthen approach, and one that is central to my view of validity, is that Muthen's approach explicitly and relatively easily allows the validity researcher to focus on sociological, structural community and contextual variables as explanatory sources of measurement invalidity (Zumbo & Gelin, 2005). Sawatzky, Ratner, Johnson, Kopec, and Zumbo (in press) provide a detailed example of using factor mixture models in validiation research from an explanatory point of view. In short, in my view of validity, measurement is not just the purview of psychology but must expand its view to be, as a start, more sociological and ecological in its orientation.

Multi-Level View of Measurement

As Zumbo and Forer (in press) noted, there are a growing number of testing and assessment programs in which one gathers individual person measures but, by design, makes inferences or decisions not about individual people but rather for an aggregate, such as a school district, neighborhood, or state. We called such measurement practices "multilevel measurement." In striking contrast to multilevel measurement, however, our widely-used measurement and testing models (including our psychometric and validation models) are, by historical precedent, geared to individual differences, as are our constructs and construct validation work.

The National Assessment of Educational Progress (NAEP) is an example of a multi-level measurement system. Educational testing and assessment in the domains of science and mathematics, for example, are focused on assessment *of* learning (i.e., summative) or even assessment *for* learning (i.e., formative) but, in both cases, the student's individual learning or knowledge is the focus. Contrary to our conventional individual differences use of such tests, however, NAEP is neither designed for, nor provides any, feedback to individual students or examinees, nor to paraprofessionals to provide feedback or planning for individual students. That is, NAEP is not used for individual decision-making but rather is used to inform policy and

perhaps assess the impact of community-scale interventions and changes in the educational and social support system.

Instead of individual differences constructs, NAEP involves what Zumbo and Forer (in press) called "multilevel constructs" that have emerged at the confluence of multilevel thinking (and ecological perspectives) with psychology, health, and social policy. A multilevel construct can be defined as a phenomenon that is potentially meaningful both at the level of individuals and at one or more levels of aggregation, but the construct is interpreted and used only at the aggregate level. While all constructs reside in at least one level, an organizational setting like formal education is inherently multilevel, given the natural nesting of students within classes within schools within school districts. Having to deal with multilevel issues should be assumed when studying phenomena in these multilevel settings (e.g., Klein, Dansereau, & Hall, 1994; Morgeson & Hofmann, 1999).

The essential feature is that these multilevel measures are not conventional educational achievement or psychological measures because they have been designed to only provide aggregate level information, such as tracking how a state is performing on a mathematics or science assessment. This aggregate level information is in contrast to the typical use of educational and psychological measures that are used for assessment of individual differences. This essential feature is easily accommodated in my view of validity as contextualized and pragmatic explanation.

From my explanation focused point of view, the central messages and their implications, are that multilevel constructs are different in purpose and scope than individual differences constructs, although they still carry high stakes for the individual test taker. Likewise, multilevel constructs necessitate multilevel measures. Implied in my view is that solely applying traditional individual differences psychometric methods (e.g., correlation with another math score at the child level) and/or most cognitive assessment approaches is insufficient evidence for the support of multilevel validation inferences. In fact, as Zumbo and Forer (in press) noted, these individual differences methods are susceptible to cross-level inferential fallacies such as the ecological fallacy or atomistic fallacy.

Multilevel measurement and testing arise when one has a multilevel construct; an individual level measure (or assessment) and aggregating it to make inferences at a higher level. Historically, multilevel constructs have not been a widespread issue in measurement and validation because testing and measurement have been immersed in, and emerged from, an individual differences psychological school of thought. Given the move to the increased policy usage of assessment results, and the shift in educational and psychological theorizing toward ecological and sociological views of our phenomenon, I fully expect to see more multilevel constructs in the coming years.

The Overlap between Test Validity and Program Evaluation

I will only briefly explore using validity as a way of looking at program evaluation. Other theorists have approached validity from an evaluation point of view, but Ruhe and Zumbo (2009) modified Messick's (1989) view of validity and approached evaluation from that framework in what they called the Unfolding Model. The term "program," in program evaluation, has been defined as a set of resources and activities directed toward one or more common goals. By this definition, a test or measure is a program. Therefore, measurement validity and program evaluation share a common conceptual core, which involves determining the worth and merit of goal-oriented activities. Ruhe and Zumbo showed that Messick's framework is an omnibus model for program evaluation. In fact, Messick treats tests as if they were programs, and the categories of his model overlap with categories commonly used for evaluating programs (e.g., cost-benefit, relevance, values and unintended consequences).

In adapting Messick's (1989) framework into an evaluation model, are Ruhe and Zumbo (2009) implying that test validity and program evaluation are the same thing? Not exactly. Fifty years ago, when the fields of program evaluation and assessment were based on (quasi) experimental methodologies, there was substantial overlap between them. However, with the adoption of qualitative methodologies and the proliferation of new approaches to program evaluation, assessment and program evaluation later emerged as distinct fields. Even so, these two fields share a common conceptual core, which is determining the worth and merit of educational and/or social policy activities. Therefore, Messick's framework can be used to evaluate both standardized tests and educational programs. Because Ruhe and Zumbo's Unfolding Model is based on Messick's framework, it is a program evaluation model grounded in the science of test assessment and educational measurement. The key to the unfolding model, like Messick's validity model and the contextualized and pragmatic view of validity I describe above, is that it brings into the forefront several features (e.g., the role of values, or of sample heterogeneity) that are largely ignored.

CLOSING REMARKS

I began this chapter by reminding the reader of Neurath's (1921) analogy of scientific verification with the construction of a ship that is already at sea. Validity was rebuilt, yet again, as Neurath highlights, one plank at a time. Wholesale changes at sea are impossible. In light of this, this chapter had two aims, to provide an overview of what I consider to be validity as

contextualized and pragmatic explanation and then discuss its implications for the process of validation. By building on the iconic works of Cronbach and Meehl (1955) and Messick (1989) and contrasting my view of validity as contextualized and pragmatic explanation, I was able to better explicate the subtleties of my own view. In the closing section of the chapter I described the implications of this view in terms of not assuming homogeneity of populations (from the point of view of the Draper-Lindley-de Finetti framework) and allowing for multilevel construct validation, as well as the overlap between test validity and program evaluation.

ACKNOWLEDGMENT

I would like to thank Professor Anita M. Hubley for feedback and for allowing for such a rich on-going discussion on validity and validation.

REFERENCES

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10,* 67–78.

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

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, DC: American Council on Education.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15,* 135–175.

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.

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.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19,* 195–229.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694 (Monograph Supp. 9).

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.

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). Amsterdam: Kluwer Academic Press.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review, 24*, 249–265.

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*, 121–132.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Lawrence Erlbaum.

Muthen, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 551–585.

Muthen, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1–22.

Muthen, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133–142.

Neurath, O. (1921). *Antispengler* (T. Parzen, Trans.). Munich: Callwey.

Ruhe, V., & Zumbo, B. D. (2009). *Evaluation in distance education and e-learning: The unfolding model*. New York: Guilford Press.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. [Theme issue in honor of Ross Traub] *Alberta Journal of Educational Research, 49*, 264–276.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether invariance holds for IRT models: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64*, 588–599. [Errata, (2004) *Educational and Psychological Measurement, 64*, 991]

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63–84.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Salmon, W. (1990) *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Salmon, W. (1998). *Causality and explanation*. New York: Oxford University Press.

Sawatzky, R. G., Ratner, P.A., Johnson, J. L., Kopec, J., & Zumbo B.D. (in press). Sample heterogeneity and the measurement structure of the Multidimen-

sional Students' Life Satisfaction Scale. *Social Indicators Research: International Interdisciplinary Journal for Quality of Life Measurement.*

Scriven, M. (1962). Explanations, predictions, and laws. In H. Feigl & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science* (Vol. 3, pp. 170–230). Minneapolis: University of Minnesota Press.

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). Amsterdam: Kluwer Academic Press.

Suppe, F. (1977). *The structure of scientific theories.* Chicago: University of Illinois Press.

van Fraassen, Bas C. (1980) *The scientific image.* Oxford: Clarendon Press.

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. Amsterdam: 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. (2007a). Validity: Foundational issues and statistical methodology. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam: Elsevier Science B.V.

Zumbo, B.D. (2007b). 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., & Forer, B. (in press). Testing and measurement from a multilevel view: Psychometrics and validation. In J. Bovaird, K. Geisinger, & C. Buckendahl (Eds.), *High stakes testing in education—Science and practice in K-12 settings [Festschrift to Barbara Plake].* Washington, DC: American Psychological Association Press.

Zumbo, B. D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5*, 1–23.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage Press.