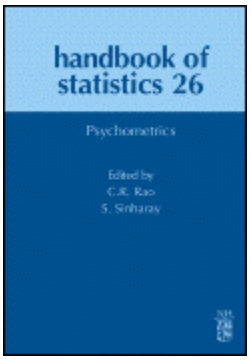
Electronic reprint of:

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.

HANDBOOK OF STATISTICS, 26 Psychometrics



C.R. Rao, The Pennsylvania State University, U. S. A.
Sandip Sinharay, MS and PhD in Statistics, Iowa State University, Ames, U.S.A., Educational Testing Service, Princeton, U.S.A.

Contents

13. Latent class analysis in psychometrics (C.M. Dayton and G.B. Macready)
14. Random-effects models for preference data (U. Bockenholt and R.-C. Tsai)

IRT models

15. Item response theory in a general framework (R.D. Bock and I. Moustaki)
16. Rasch models (G.H. Fischer)
17. Hierarchicam item response theory models (M.S. Johnson, S. Sinharay and E.T. Bradlow)
18. Multidimensional item response theory (M.D. Reckase)
19. Mixture distribution item response models (M. von Davier and J. Rost)
20. Scoring open ended questions (G. Maris and T. Bechger)
21. Assessing the fit of item response theory models (H. Swaminathan, R.K. Hambleton and H.J. Rogers)
22. Nonparametric item response theory and special topics (K. Sijtsma and R.R. Meijer)

Topics of special interst

23. Automatic item generation and cognitive psychology (S. Embretson and X. Yang)
24. Statistical inference for causal effects, with emphasis on applications in psychometrics and education (D.B. Rubin)
25. Statistical aspects of adaptive testing (W.J. van der Linden and C.A.W. Glas)
26. Bayesian Psychometric modeling from an evidence-centered design perspective (R.J. Mislevy and R. Levy)
27. Value-added modeling (H. Braun and H. Wainer)
28. Three statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licencing data (H. Wainer and L.M. Brown)
29. Meta-analysis (L.W. Hedges)
30. Vertical scaling: statistical models for measuring growth and achievement (R.J. Patz and L. Yao)
31. Cognitive Diagnosis - part I: (L.V. Dibello, L.A. Roussos and W. Stout) - part II: (S.J. Haberman and M. von Davier)

Operational

32. Marginal Estimation of Populalation characteristics: recent developments and future directions (M. von Davier, S. Sinharay, A. Oranje and A. Beaton)
33. Statistical procedures used in college admissions testing (J. Liu, D.J. Harris and A. Schmidt)
34. Integration of models (R.L. Brennan, D.R. Eignor, M.J. Gierl, J.P. Leighton, I. Lawrence, N. Kingston, P. Sanders and C. Schmeiser)

Bibliographic & ordering Information

Hardbound, 1190 pages, publication date: NOV-2006
ISBN-13: 978-0-444-52103-3
ISBN-10: 0-444-52103-8
Imprint: NORTH-HOLLAND

.

Handbook of Statistics, Vol. 26
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(06)26003-6

Validity: Foundational Issues and Statistical Methodology

*Bruno D. Zumbo**

1. Introductory remarks

From a statistical point of view, psychometricians are faced solely with an array of numbers whose elements are outcomes of random variables that are most often discrete or ordered categorical in nature. It is the ultimate aim of the psychometrician to use this array to make a variety of meaningful inferences about the examinees and the measurement instrument itself, which should be appreciated as a daunting task.

A long tradition in psychometrics has evolved (as it has in statistics) of concerning ourselves with decomposing observational values into a component that is *deterministic* and a component that is *stochastic* so that relationships between manifest and unobserved variables can be explicitly stated and uncertainty about model parameters can be estimated and used to qualify the inferences that are possible under a given model. In this light, many contemporary measurement inferences are model-based. Contemporary psychometric models involve latent variables, mixture distributions, and hyperparameterized parameter-driven models – all of which are, of course, descendants of a longstanding statistical tradition but are unique because they are located at the *intersection* of examinee and item spaces, *both* of which are typically of interest to measurement specialists (Zimmerman and Zumbo, 2001). As Zumbo and Rupp (2004) remind us, this is central to all major statistical theories of measurement. Specifically, classical test theory generically decomposes the observed score into a deterministic part (i.e., true score) and a stochastic part (i.e., error), generalizability theory further unpacks the stochastic part and redefines part of error as systematic components, and item response theory reformulates the two model components by inducing latent variables into the data structure. Structural equation models and exploratory as well as confirmatory factor analysis models decompose the covariance matrix of multivariate data into deterministic (i.e.,

*I would like to thank the following individuals for feedback and review of an earlier draft of this chapter (in alphabetical order): Greg Camilli, Craig Deville, Brian French, Anne M. Gadermann, Anita M. Hubley, Chris Richardson, Rick Sawatzky, Steve Sireci, and Cornelia Zeisser. Bruno D. Zumbo, Ph.D., is Professor of Measurement and Statistics as well as a member of the Institute of Applied Mathematics and the Department of Psychology at the University of British Columbia. His primary interests are in psychometrics, statistical science, and the mathematical and philosophical foundations of measurement. Send Correspondence to: Professor Bruno D. Zumbo, Department of ECPS, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4, E-mail: bruno.zumbo@ubc.ca, Fax: +1 604 822 3302.

reproduced covariance matrix) and stochastic (i.e., residual matrix) components, which is a model that can be equivalently written as a formulation involving latent variables.

As Zumbo and Rupp (2004) note, even though latent trait indicator values and observed composite scores are typically highly correlated, the injection of a latent continuum into the data matrix has given us the property of item and examinee parameter invariance. For perfect model fit across populations and conditions item and examinee parameter invariance has allowed us, for example, to define conditional standard errors of measurement similar to *generalizability theory*, and has opened up the road for adaptive testing through the use of item and test information functions. For a technical description of invariance and the effects of its violations please see Rupp and Zumbo (2003, 2004, 2006). Still, these advances have not come without a price. Improvements in the level of modeling and in quantifying measurement error have come at the expense of large sample sizes that are typically required for parameter estimation in both frequentist and Bayesian frameworks (see Rupp et al., 2004).

Throughout this chapter, 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). Furthermore, depending on the context (e.g., structural equation modeling) items or tasks may be referred to as indicators or manifest variables. In the remainder of this chapter, I will contextualize the terminology as much as possible. 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.

At this point, the following quotation that is often attributed to the philosopher John Woods may be appropriate "My paper is like a domestic wine, it's not particularly robust or important but I think you may be amused by its pretensions" – with all the various meanings of pretension. With this humorous (yet appropriate) remark in mind, the purpose of this chapter is to shine a spotlight on some foundational and statistical issues in validity theory and validation practice. Due to space limitations, relative to the breadth and scope of the task at hand, for some issues I will provide details whereas in others more general integrative remarks. The chapter is organized as follows. Section 2 discusses several foundational issues focusing on several observations about the current state of affairs in validity theory and practice, introducing a new framework for considering the bounds and limitations of the measurement inferences, and briefly discussing the distinction between measures and indices. Section 3 deals with two statistical methods, variable ordering and latent variable regression, and introduces a methodology for variable-ordering in latent variable regression models in validity research. Section 4 closes the chapter with an overview in the context of remarks around the question "When psychometricians speak of validity what are they really saying?"

Wherever possible, an example will be used throughout this chapter to make matters concrete and motivate several validity questions.

1.1. Example

The domain of validity can be very abstract and philosophic so let me introduce an example using the Center for Epidemiological Studies Depression scale (CES-D) to concretize matters and ground the discussion. The CES-D is useful as a demonstration because it is commonly used in the life, behavioral, health, and social sciences. The CES-D is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population (Radloff, 1977). The CES-D prompts the respondent to reflect upon his/her last week and respond to questions, such as "My sleep was restless", using an ordered or Likert response format of "rarely or none of the time (less than 1 day)", "some or a little of the time (1–7 days)", "occasionally or a moderate amount of time (3–4 days)", and "most or all of the time (5–7 days)" during the last week. The items typically are scored from zero (less than 1 day) to three (5–7 days). Composite scores therefore range from 0 to 60, with higher scores indicating higher levels of depressive symptoms.[1] The data presented herein is a sub-sample of a larger data set collected in northern British Columbia, Canada. As part of a larger survey, responses were obtained from a convenience sample of 600 adults in the general population – 290 females ranging in age from 18 to 87 years (mean of 42 years), and 310 males ranging in age from 17 to 82 years (mean of 46 years).

Of course, the composite scale score is not the phenomenon of depression, per se, but rather is related to depression such that a higher composite scale score reflects higher levels of the latent variable depression. Although I will return to this later in the chapter, it is useful to note in the context of the example that there are essential distinctions between (i) the observed items, and their observed score composite, (ii) the latent variable which is defined and estimated in a statistical model, and (iii) the phenomenon, attribute or construct, of depression. Cast in this way, two commonly found questions of test validation are of interest: (a) Given that the items are combined to created one scale score, do they measure just one latent variable? and (b) Are the age and gender of the respondents predictive of the latent variable score on the CES-D? The former question is motivated by psychometric necessities whereas the latter question is motivated by theoretical predictions and investigates known-group gender and age differences on CES-D scores. Although this example is used throughout the chapter, I will return to these two validity questions after discussing some foundational issues in Section 2.

2. Foundational issues

2.1. A series of observations on the current state of affairs in validity theory and practice

There is little question that validity theory and practices have changed over the last century. Angoff (1988), Hubley and Zumbo (1996), Jonson and Plake (1998), and Kane

[1] Some items need to be re-coded to match a larger score being equated to more depressive symptoms.

(2001) provide histories and overviews of validity theory. In brief, the early- to mid-1900s were dominated by the criterion-based model of validity, with some focus on content-based validity models. The early 1950s saw the introduction of, and move toward, the construct model with its emphasis on construct validity; a seminal piece being Cronbach and Meehl (1955). The period post Cronbach and Meehl, mostly the 1970s to date, 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).

Measurement or test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals from a given sample and in a given context. The concept, method, and processes of validation are central to constructing and evaluating measures used in the social, behavioral, health, and human sciences because, without validation, any inferences made from a measure are potentially meaningless, inappropriate and of limited usefulness.

The above definition highlights three central features in current thinking about validation. First, it is not the measure per se that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from scores obtained from a scale may appear subtle at first blush but, in fact, it has significant implications for measurement and testing because it highlights that the validity of the inferences one makes from test scores is bounded by place, time, and use of the scores resulting from a measurement operation.

The second central feature in the above definition is the clear statement that inferences made from all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an "objective" educational, economic, or health indicator such as number of students finishing grade 12, or a more psychological measure such as a self-report depression measure, one must be concerned with the validity of the inferences (Zumbo, 1998). It is interesting to note that in my experience in working with health researchers the much exalted medical measures such as blood pressure readings or diagnostic assessments generally have no better psychometric and validity properties than psychosocial or educational measures.

The final central feature in the above definition is that validity depends on the interpretations and uses of the test results and should be focused on establishing the inferential limits (or bounds) of the assessment, test, or measure. In short, invalidity is something that distorts the meaning of test results for some groups of examinees in some contexts for some purposes. Interestingly, this aspect of validity is a slight twist on the ideas of test and item bias (Zumbo and Hubley, 2003). That is, test and item bias aim analyses at establishing the inferential limits of the test – i.e., establishing for whom (and for whom not) the test or item score inferences are valid.

It is instructive at this point to contrast contemporary thinking in validity theory with what is commonly seen in many introductory texts in research methodology in the social, behavioral, health, and human sciences with a series of observations about the state of the field of validity. The purpose here is not to be exhaustive of all aspects of

current thinking in validity but rather to shine a light on some points that I believe are key to contemporary validity theory and validation practice.

Observation 1

Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum.

Observation 2

Whereas the quantification of error of measurement aids in the inference from the observed (composite) score to the true score or latent variable, validity theory aids us in the inference from the observed score to the construct of interest, via the latent or unobserved variable. Although there is sometimes confusion even in the technical measurement literature, it is important to note that the construct is not the same as the true score or latent variable, which, in turn in practical settings, is not the same as the observed item or task score. An obvious and popular distortion of these concepts is nearly ubiquitously seen in the use of the term "construct comparability" in, for example, cross-cultural measurement settings. What is often referred to as "construct comparability" is, at best, the equivalence of latent variables. Construct comparability is more than the equivalence of latent variables.

Observation 3

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. In this sense, 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 hidden ethical and social values that overtly or inadvertently influence that process (Messick, 1995).

Observation 4

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 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 is sufficient evidence to use the test as a measure of that criterion. In my view (e.g., Zumbo, 2005), 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 will return to this point later.

Observation 5

Although it was initially set aside to elevate construct validity, content-based evidence has gained momentum again in part due to the work of Sireci (1998). In addition, as is shown by Sireci, content-based evidence is also growing in influence in the test development process with the use of subject matter experts early in the test development process.

Observation 6

As one can see in Messick (1989) and in Zumbo's (1998) volume on validity, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall (unified) validity framework for evaluating test interpretation and use came the consideration of the role of value implications and social consequences as validity evidence contributing to score meaning and use. This movement has been met with some resistance. In the end, Messick (1998) made the point most succinctly when he stated that one should not be concerned simply with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of "normal" test use. Social consequences and value implications, of course, also feed back into theory development (Hubley and Zumbo, 1996).

It is my interpretation that, by opening measurement and testing to a larger social and cultural interpretation, we are recognizing the social and cultural forces involved in test construction and test use. This, I believe, is a natural consequence of the recognition that test scores are impacted by the situation and living conditions of an examinee and that their test scores are, in part, a result of what the examinee makes of those living conditions. Zumbo and Gelin (in press), for example, build on this idea and expand the realm of test score explanation and validity beyond the cognitive to include psychosocial and cultural variables. In so doing, one embraces Messick's perspective on test consequences and acknowledges the importance of a consequentialist moral tradition such as that of the philosopher John Stuart Mill, who noted that our moral obligation as human beings is to try to make the greatest good – in our case the greatest good from our tests.

The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results being used to determine financing and staffing for educational institutions. Likewise, we will see the matter and role of consequences of measurement become prominent as we see health outcome measurement, for example, used to determine financing as well as staffing in public institutions in the United States, Europe, the Asia-Pacific and Canada.

Observation 7

Due to the number of international and cross-cultural efforts in testing and measurement, test adaptation and translation is growing in importance and centrality in the test validation literature and practice (see, for example, Hambleton and Patsula, 1998; Kristjansson et al., 2003). In this context, the matter of measurement invariance is also on the rise in importance and centrality.

Observation 8

There has been some debate as to whether reliability is a necessary but not sufficient condition for validity. It seems to me that this issue is better cast as one of measure-

ment precision so that one strives to have as little measurement error as possible in their inferences. Specifically, reliability is a question of *data quality*, whereas validity is a question of *inferential quality* (Zumbo and Rupp, 2004). Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter.

Observation 9

The use of cognitive models as an alternative to traditional test validation has gained a great deal of momentum in the last 15 years. One of the limitations of traditional quantitative test validation practices (e.g., factor-analytic methods, validity coefficients, and multi trait-multi method approaches) is that they are descriptive rather than explanatory. The cognitive models, particularly the work of Embretson (1994), Nichols (1994), and Mislevy (1996), have laid the groundwork to expand the evidential basis for test validation by providing a richer explanation of the processes of responding to tests and hence promoting a richer psychometric theory-building. A similar push for explanatory power has also taken place in the area of differential item functioning where attitudinal, background, and cognitive variables are used to account for differential mediating and moderating variables (Zumbo and Gelin, in press; Zumbo and Hubley, 2003).

Observation 10

As Zumbo and Rupp (2004) and Zumbo and MacMillan (1999) note, it is important to move beyond simple "cold" cognitive models of the sorts found in the cognitive revolution of psychology in the 1960s and 1970s, to more contextualized and social cognitive models that recognize that the examinee functions and lives within a social world that shapes and influences cognition. That is, in many assessment situations, researchers use the word 'cognition' to loosely refer to any process that is somehow grounded in our minds. These psychometric cognitive models seek to explicitly represent the cognitive processes that examinees engage in when responding to items via parameters in mathematical models, which typically consist of augmented IRT models, classification algorithms based on regular IRT models, or Bayesian inference networks that have IRT models as a central component. Developments in cognitive models have often taken place primarily in educational achievement and psychoeducational assessment contexts. An exception was Zumbo et al. (1997) in personality assessment in which they studied the relation of the abstractness and concreteness of items to the psychometric properties of a personality measure. Other advances are being made in the development of simulation-based assessment software that emphasizes a deeper and richer understanding of the cognitive processes required for performing certain tasks in which data are analyzed through Bayesian networks (Mislevy et al., 1999).

Observation 11

The basic idea behind the cognitive models and other explanatory approaches is that, if one could understand why an individual responded a certain way to an item, then that would go a long way toward bridging the inferential gap between test scores (or even latent variable scores) and constructs. Building on this observation, Zumbo (2005) notes it is important to separate the concept of validity from the process of test validation. According to this view, validity per se, is not established until one has an explanatory model of the variation in item responses and the variables mediating, moderating, and

otherwise affecting the response outcome. This is a tall hurdle indeed but I believe that it is in the spirit of the earliest work on validity such as that of Cronbach and Meehl (1955) interpreted as a strong form of construct validity; and points to the fact that by focusing on the validation process rather than the concept of validity we have somewhat lost our way as a discipline. 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. Quite the contrary, rather it points to the fact that the information from the validation process needs to be directed toward supporting the concept of validity and not is the end goal itself. My aim is to re-focus our attention on why we are conducting all of these psychometric analyses: to support our claim of the validity of our inferences from a given measure. For example, 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 thus establishes the inferential limits of the test.

Observation 12

There should be more discussion about "models" and "modeling" in validity, and their varieties of uses, meanings, and intentions (Zumbo and MacMillan, 1999). As Zumbo and Rupp (2004) note, in validity research, the issue is less about a lack of models for new kinds of test data but rather a lack of awareness in the applied world that these models exist along with a mismatch of assessment instruments and modeling practice. In other words, if test developers are interested in providing examinees and institutions with richer profiles of abilities and developmental progress, the nature of the assessment methods has to change to provide richer data sets from which relevant information can be more meaningfully extracted. What is meant by "more meaningful" will, of course, in the end, depend on the use of the assessment data but, in general, authorities in the field today are beginning to agree that we need more than simple test responses scored 0 and 1 to validate the inferences that are being made from the test data. It has been my view that the key to useful cognitive models is that they need to be explanatory and not just another set of descriptive models in cognitive terms rather than mathematical terms (Zumbo and MacMillan, 1999). It is important to note that not all cognitive models are explanatory. Put differently, a change of terminology in psychometrics from mathematical terminology to cognitive terminology is insufficient to claim true advances in gathering more meaningful and weighty validity evidence.

In this context, it is important to acknowledge the seminal paper by Borsboom et al. (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.[2] Philosophically this is a very tidy, and as the

[2] A variation on this idea, with a different focus, can be seen in the work of Bollen and Lennox (1991) in their distinction between "measures" as opposed to "indices" in the form of reflective as opposed to formative indicators in the context of structural equation modeling and factor analysis. I discuss these concepts more fully in a section below.

authors' themselves acknowledge, 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, since 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 we are wishing to measure, such as that embodied in the meta-theoretical views of Messick and others I describe above, but simply more knowledge: knowledge of causes.

I take a different position in the present chapter and elsewhere (e.g., Zumbo, 2005) than Borsboom and his colleagues. My view is that validity is a matter of inference and the weighing of evidence, and that explanatory considerations guide our inferences. I am not fond of the exclusive reliance on "causal" model 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, and that it is too weak or permissive, that it undermines our explanatory practices. My current leanings are toward inferences to the best explanation. In short, I believe that explanation is key to validity (and supporting the inferences we make from test scores) and that causation as the sole form of explanation is too narrow for the broad scope and context in which measurement and testing is used. 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 – 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.

Observation 13

There is one main point I want to make about the various uses of the term "model" as they relate to validity. In working with test developers and psychometricians, I have come to realize that "model" is a favorite word of measurement specialists. We use it in referring to "statistical models", "psychometric models", "IRT models", "cognitive models", etc. Sometimes, unbeknown to us, we use it with various shades of meaning. As Zumbo and MacMillan (1999) note, often times we use it to convey the sense of a: (i) mathematical model, (ii) model in the wider philosophic sense, (iii) explanatory model, (iv) descriptive model, (v) stochastic or random variable model, (vi) logical model, and (vii) computational model, to list but a few. What complicates matters is that these uses of "model" are not mutually exclusive (nor exhaustive) but they do have essential, though subtle, distinctions. Having had measurement specialists point out that they find there to be a confusing array of models, I have become fascinated that a word that is so commonly used and is the source of so much tension among measurement specialists has yet, to my knowledge, to be adequately defined and discussed in the psychometric literature. Even with all of these vague references and multiple uses, I am struck by how many of us are like Pygmalion and fall in love with our models, in good part we fall in love with what we want our model be. In some cases we are very much

like the fictional character Pygmalion in that we believe that our particular model of interest is so realistic that we fall in love with it.

Clarity will come from knowing how we are using the word "model" and what we are implying in its use. Zumbo and MacMillan (1999), in very broad strokes, described two different kinds of models as well as their purposes, types, and uses. I found that even this broad discussion of models helps clarify some of the confusion and clear some of the (battle) ground regarding the models used in applications of measurement. For example, it has helped me understand the philosophic basis and practical difference between Rasch and one-parameter IRT by clarifying the Rasch model's epistemological difference. Statistically, the Rasch and one-parameter IRT models are the same but the differences in epistemology between Rasch and IRT modelers leads to rather different validity practices and test data claims.

Observation 14

The function of the psychometric model in validity research is to step in when the data are incomplete. In an important sense, we are going from what we have to what we wish we had. If we had available the complete data or information, then we would know the true score, or theta in IRT models, and no statistics beyond simple summaries would be required. There would be no need for complex models to infer the unobserved score from the observed data and, hence, no need to check the adequacy and appropriateness of such inferences.

The occurrence of complete data or full information, as I describe it, is not commonly encountered, if ever, in the practice of measurement. Naturally, this leads to the common experiences that are the defining characteristics of what we call modeling:

(a) The data you have is never really the data you want or need for your attributions, recommendations or decisions.
(b) No matter how much data you have, it is never enough because without complete information you will always some error of measurement or fallible indicator variable. We get around data and information limitations by augmenting our data with assumptions. In practice, we are, in essence, using the statistical model to creating new data to replace the inadequate data. The most common data augmentation assumption in psychometrics is that the correlation among items is accounted for by an unobserved continuum of variation – of prominence in latent variable and factor analysis models. The key words here are "unobserved" and "continuum", as opposed to observed and categorical.
(c) In contemporary measurement practice, which is heavily-laden with model-based measurement practices, the inferences, in part, come from the model itself. In short, the validity statements rest on the measurement model. In fact, as we noted above, given that one begins with an array of numbers denoting responses to items or tasks for each examinee, it could be argued that the psychometric model "provides" the inferences one can make by being the vehicle for going from what we have to what we wish we had. Therefore, the measurement model is not neutral in validation and, not surprisingly, one's test score interpretations may change depending on the psychometric statistical model being used.

2.2. How the observations on the current state of validity speak to the example

In a broad sense, then, validity is about evaluating the inferences made from a measure. All of the statistical methods discussed in this handbook (e.g., factor analysis, reliability, item analysis, item response modeling, regression, cognitive models and cognitive psychology in testing) are directed at building the evidential basis for establishing valid inferences. In this light, I am bewildered by colleagues who make claims that classical test theory ended with the seminal work of Lord and Novick (1968) or perhaps even earlier with the publication of Gullicksen (1950). A more accurate statement would be that the systematic development and articulation of classical test theory began with these publications – for a description of recent advances in reliability and classical test theory see Zumbo and Rupp (2004).

Taking the CES-D as a case in point to illustrate some of the conceptual features of the current view of validity, gender and age differences are often reported in the depression literature – that is, generally, females score higher than males on depression measures, and depression scores are higher in young adults (although high scores tend to also be found in the elderly). Therefore, any new measure of depression (or the application of a known measure with a new population of respondents) is expected to have, for example, higher depression scores for female respondents. This finding of gender differences is part of what reassures some researchers that they are, in fact, measuring depression. The gender differences are therefore embedded in the network of ideas defining the depression construct hence demonstrating the interplay of empirical findings, theoretical developments, and construct elaboration that is central to test validation. At the same time, the gender and age findings also illustrate the potential for value implications and social consequences. When examining the consequential basis for CES-D score interpretation, one needs evidence of construct validity and an appraisal of the construct label itself, the broader theory, and even broader ideologies (Hubley and Zumbo, 1996). For example, what is the implication of labeling the constellation of listed symptoms of depression in the CES-D as "depression"? What are the implications of our theories of aging for the recognition of depression in the elderly? And vice versa, what are the implications of our empirical results with current measures of depression for theory-building in the area of depression? Likewise, what are the implications of our theories of gender differences and the recognition of women being labeled depressed more often than men? And, for example, given our theories of aging is depression among the elderly viewed as a normal, abnormal, negative, or a realistic state? In addressing these questions, we are bringing the social consequences and value implications to the foreground of the validity discussion and shining a light on the appraisal of the hidden values of inferences made from the CES-D scores. This example of the CES-D vividly demonstrates (a) how one should not be simply concerned with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of "normal" test use, and (b) how social consequences and value implications feed back into theory development. For now, our discussion has been conceptual and theoretical (after all, validation does involve theorizing), but our focus will turn empirical when we return to the question of age and gender as predictors later in this chapter in the discussion of variable ordering and regression.

2.3. *Different strengths of test inferences: A new framework for measurement inferences*

One must keep in mind the purpose of the testing. The two main purposes for testing are: (a) descriptive: wherein one is assigning numbers to the results of observations for the purpose of obtaining a scale score in scientific or policy research, or (b) decision-making: using the original scale scores to categorize individuals or groups of individuals based on their responses to the test or measure. The latter purpose is the most common use of testing in, for example, the educational testing, diagnostic screening, certification testing and occupational and employment contexts. Two common examples are: (i) individuals are selected or not selected for a job, or (ii) individuals are screened for further testing, interviewing, or training.

The whole purpose of measurement and methodological techniques is to help in establishing the *degree of* validity of the *inferences* made from the scores obtained on a test or assessment battery. The key words in the previous sentence are "degree of" and "inferences". These two features are highlighted in nearly all current views of validity (see, for example, Hubley and Zumbo, 1996, or Zumbo, 1998). That is, as I described above, validity is not an all-or-none decision but rather a matter of degree. This, of course, implies that validation is an on-going process and is not simply something done once and involving only a correlation with a "gold standard". Furthermore, what is being validated are the inferences made from the measurement and not the measurement tool itself. Therefore, one can have a degree of evidential strength for validity and, in fact, as I will describe below, there is a continuum of validity statements that can be articulated in the process of validation.

Below I will introduce a framework that will be useful in the critical analysis of validity evidence.[3] Furthermore, this framework will help answer questions of sampling and representativeness and will aid us in being more explicit about the strength of our validity claims. It is imperative that we be explicit about the limitations or strengths of our validity claims. The following framework is modeled on Draper's (1995) approach to classifying causal claims in the social sciences and, in turn, on Lindley's (1972) and de Finetti's (1974–1975) predictive approach to inference. Although I will rely heavily on Draper's approach (including paraphrasing some of his very clear statements), Draper focused on the issue of causal inferences in quasi-experimental designs whereas I will discuss the matter of measurement validity. In recognition of its bases, I will refer to my new framework as the Draper–Lindley–de Finetti (DLD) measurement validity framework.

The main point to be taken from the DLD framework is the necessity to be explicit about the sorts of inferences one makes and that one is able to make. It is not that some inference is necessarily better than others (because this sort of value judgment needs to take the purpose of the measure into account), but rather that credible and defensible science requires one to be explicit about the sorts of inferences that are made and that can be made in a given context. As Draper states:

[3] An earlier version of this framework was introduced by Zumbo (2001) and then further expanded in Zumbo's invited address to the conference in celebration of the work of Professor Ross Traub upon his retirement, OISE/University of Toronto, 2002.

> *Within this approach, the only inferential elements with objective reality are data values X you have already observed and data values Y you have not yet observed. Inference is then the process of quantifying your uncertainty about future observables Y on the basis of things you know, including the X values and the context in which they were obtained. Informally, one might call X the data you have and Y the data you wish you had; with this terminology, a statistical model supporting your inference is a mathematical story that links these two data sets.* (p. 119)

In the context of validity theory, the data you have and the data you wish you had may differ from each other in two main ways: (1) problems of measurement, and (2) problems of sampling.

1. Problems of measurement

There are two common forms of measurement problems. One form arises when you are trying to quantify something elusive and you are not sure you have got it right. The technical terms for this are construct underidentification or construct misrepresentation. The other form occurs when you may have also included construct(s) irrelevant to the task at hand. The technical term for this is construct irrelevant variance. For example, you are measuring some knowledge domain such as reasoning skills, but the construct "writing ability" could be considered irrelevant because we may be interested in reasoning but it is confounded with writing ability. That is, because the answers are expressed in written form in essays and sometimes good writing masks poor reasoning skills and vice versa.

A useful way of thinking about the tasks or items in an assessment is that they are a sub-set of a (potentially well-defined) finite population of items or tasks from which one or more test forms may be constructed by selection of a sample of items from this population. Obviously, creating all possible items or tasks would be both economically and practically unfeasible. In some situations, one has a clearly articulated domain of items (perhaps even articulated in the form of a test blue-print) so that one would have a very good sense of the exchangeability of the items one has with those that are not in the assembled test. In other cases, one has very little information about the domain being measured and hence has little knowledge of the exchangeability of the items one has with those in the hypothetical item population (or domain). In this case, the data you have is the information on the sampled items or tasks in your assessment and the data you wish you had is the corresponding information on the unsampled items. In essence, this gets at how well one knows the construct or domain being sampled and how well one plans (e.g., task specifications) in the process of constructing the assessment.

Before turning to problems of a sampling nature, I would like to add that measurement error (quantified, for example, by reliability of measurement) is also a related but separate issue here. In the case of measurement error, the data you have are your "measurements" or "observations" which have measurement error or construct irrelevant variance, whereas the data you wish you had are the "true" or error-free scores.

2. Problems of sampling

The second way the data you have and the data you wish you had differ relates to problems of a sampling nature. This sort of problem arises, for example, when there is a

finite population of subjects of scientific or policy interest and, due to financial or other constraints, you can only get a subset of the whole population. In this case, the data you have is the information on the sampled individuals and the data you wish you had is the corresponding information on the unsampled people.

The problems of measurement and problems of sampling may, in turn, be thought of as special cases of a *missing data* problem. Also, it is important to note that much of the work in the predictive approach to inference described by Draper is founded on the notion of *exchangeability*. Exchangeability can be thought of as: without prior or additional/supplemental information, the data are similar – and hence, in a mechanical sense, exchangeable.

In part, the DLD framework is a natural extension of ideas in the 1940s and 1950s by Louis Guttman who wrote about generalizing from measures (or tasks) that we have created to more tasks of the same kind. Therefore, this framework is a merger of Guttman's ideas with some Bayesian thinking about inferences. There is also a tidy connection to measurement invariance that I will discuss in more detail below.

Now, when measurement and validation examples are examined from the DLD "predictive point of view", four kinds of inference of varying strength and scope of generality are discernible. Figure 1 depicts these various types of inference graphically. As much as possible I have tried to stay close to Draper's conceptualization, including the terms for the various forms of inference (i.e., calibrative, specific sampling, specific measurement, and general measurement).

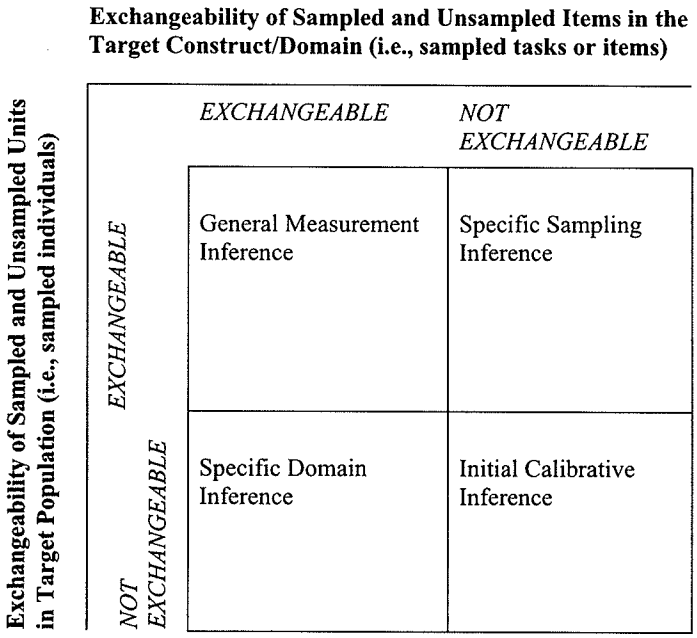


Fig. 1. The various forms of measurement inference.

- Initial Calibrative Inference (calibration inference). This provides the lowest level of inferential strength and generalizability of results. The term "calibrative" is being used here to convey the sense that a measure is simply being applied in an exploratory fashion so that one can observe the measure in action with data. The case could be made that this is not useful measurement at all because item exchangeability is not there and hence not justifiable. However, although I do share this general sentiment, I prefer that the statements that one is limited to be simple calibration statements about whether items or tasks can be comprehended and completed in the most elemental way by some group of individuals.
- Specific sampling inference. This sort of inference shares the domain sampling limitations of calibrative inference except that one has a very good sense of who is being measured – in the sense that the sample is exchangeable to the target population of respondents. Obviously, the statements that one is limited to are specific calibration statements about whether the measure can be used at all for some well-defined population of individuals.
- Specific domain inference. This type of inference is simply of the sort where one can make strong claims (in terms of, for example, reliability and other measurement properties) about what is being measured but the statement is not generalizable because the exchangeability of sampled and unsampled respondents (or examinees) is not justifiable. In fact, this is a very common form of inference in research settings where claims of reliability and validity are limited to the sample at hand.
- General measurement inference. This kind of inference is ideal because it has both exchangeability of items and exchangeability of sampling units. That is, like specific domain inference one has exchangeability of sampled and unsampled items (or tasks) in the target domain *and* one also has the justification for assuming exchangeability of the sampled and unsampled subjects. Therefore, strong validity and measurement claims can be made.

In terms of inferential strength, I conclude that both initial calibrative and specific sampling are less strong than specific domain, which in turn is less than general measurement inference. Clearly, general measurement inference is the strongest. I will return to this later.

Note that Figure 1 depicts the four corners of a matrix, which most certainly has several gradients in between these four corners. Also, the exchangeability of sampled and unsampled items is yet another way of thinking of a limited form of construct validity. As the DLD framework shows, however, construct validity needs to be considered in the light of the exchangeability of sampled and unsampled units (i.e., respondents) in the target population.

What the DLD framework highlights is that you need to consider the sample in two distinct ways. First evidence for the validity of inferences made from task performance needs to be provided with each new sample examined. This is why the process of validation is on going. The inferences you make from your task may be valid for one population (e.g., males) but not another population (e.g., females). Or the inferences you make may be valid for males in one setting (e.g., a large North American city) but not in another setting (e.g., in rural communities in Eastern Europe or perhaps even large cities in Asia). Second, one needs to consider whether the sample with whom you

are working (whether in a validation, standard-setting, or normative study) is exchangeable with your population of interest. For example, is the sample of respondents to the CES-D used to validate the inferences representative of target population of interest to users of the CES-D (that is, with respondents from that same population of interest that were not in your sample)?

The objective in this classification and ordering of inferences in the DLD framework is to encourage researchers to be explicit about the types of inferences they can make. That is, we need to be explicit about the information we have about the level of exchangeability of individuals and tasks or items used in our validation study. This explicitness will go a long way toward creating credible scientific measurement evidence. For example, it is sometimes stated in the context of personnel testing that job simulation tasks are used so that one does not have to make any inferences from the test scores; a reliance on a form of what is called face validity. Clearly, from the above framework one is still making inferences in "simulation" tasks because you still do not have the individual doing the job, per se. The simulation may be more realistic, hence shortening the inferential distance from simulation to real job performance, than a job knowledge test, but an inference is still being made from those job simulation tasks to other tasks like them. One should be cautious, however, not to let the DLD framework be interpreted in a weak form, hence allowing the same slippage as is noted above with weak forms of construct validity. This caution also applies to other simulation-like (i.e., performance) tasks such as some language tests.

Note that the DLD framework puts importance on the sampling units (the respondents) and their characteristics, something that this not highlighted enough in the conventional discussions of psychometrics and of validity evidence. Most, but not all, validation studies in the research literature give little time to the exchangeability of the sampled and unsampled units. In addition, there is little discussion in the psychometric literature of matters of complex sampling. Survey data now being collected by many government, health, and social science organizations around the world have increasingly complex structures precipitating a need for ways of incorporating these complex sampling designs into our psychometric models. For example, many surveys entail very complex sampling designs involving stratification and clustering as the components of random sampling. In addition, there can be complexities due to the patterns of nonresponse – either planned or unplanned nonresponse. A common psychometric problem is to compute variances for psychometric models that incorporate both the stochastic modeling as well as the survey design. Advances are being made on item response and latent variable modeling, including item response theory, with complex surveys by, for example, Asparouhov (2005), Asparouhov and Muthen (2005), Cyr and Davies (2005), Kaplan and Ferguson (1999), Mislevy (1991), Muthen and Satorra (1995), Thomas (2001), Thomas and Cyr (2002), and Thomas and Zumbo (2002). It is worth noting that the DLD framework shines a spotlight on the person sampling aspect of measurement, which has mostly held a secondary place to item or domain sampling in psychometrics.

Now, when inferential examples of testing and measurement are examined from the predictive point of view, four kinds of test inferences – of varying strength and scope of generality are discernable. I have found it useful to help one think of the various possibilities from the combinations of the four kinds of test inferences, whether they happen

regularly or not. Examining the various combinations of the kinds of test inferences also helps one detail the range of conditions under which generalizations and inferences are expected to hold.

The notions of generalizations and inferences are intimately tied to the notion of invariance in measurement models, and hence to validity. Simply put, measurement invariance allows us model-based generalization and inferences – noting, of course, that model-based inferences are inferences from assumptions in the place of data, as described above. Much has been said in the item response theory literature about invariance, and Rasch specialists make much hay of that model's invariance and specific objectivity properties with, often, implicit and explicit suggestions that one would have greater measurement validity with the Rasch model[4] (Bond and Fox, 2001; Wright, 1997). This suggested supremacy of the Rasch model is an overstatement. In fact, in several applications of the Rasch model one hears the claim that simply fitting the Rasch model gives one measurement item and person parameter invariance, without mention of any bounds to this invariance. The suggestion is that if one fits the Rasch model to data, then one can apply those item parameters to all individuals (imbuing a universality that seems remarkable). Furthermore, if one fits the Rasch model to data, the person parameters (i.e., the theta scores) can be adequately predicted by any items; again, without reference to bounds on this generality. There many advantages to Rasch (or 1-parameter IRT) models in test applications but rarely are they the advantages that advocates of the Rasch model present – one advantage is the ease of IRT test equating in a one-parameter IRT model.

The DLD framework is meant to establish limits, bounds, and conditions for the inferences. It is unfounded to believe that fitting a model to data gives one such unbounded and unbridled universality. Simply put, the main goal of item response modeling should always be to make valid inferences about the examinees. However, inducing latent variable(s) into the data structure, by, for example, fitting a Rasch model to your data cannot mechanically increase the validity of these inferences. In contrast, the DLD framework helps one detail the range of conditions under which invariance is expected to hold. There may be situations (such as those found in a type of calibrative measurement or assessment context) wherein we do not have any sense of a population or sub-population and, in those contexts, we are, in essence, not concerned with invariance. See Rupp and Zumbo (2003, 2004, 2006) for detailed descriptions of invariance in item response theory. In more common contexts, however, the aim is to use a statistical measurement model to draw inferences from calibration samples to the respective populations from which these were drawn. An additional focus is on the range of possible conditions under which invariance is expected to hold.

Simply going about fitting an item response model to data does not necessarily give you measurement invariance. Believing that going about fitting a model to data guarantees you measurement invariance is simply magical thinking! In short, invariance requires that the model be correct (true) in all corners of the data and is an empirical

[4] A similar claim is also often made by advocates of the Rasch model that it has fundamental measurement properties, such as additive item response functions, that make it a superior model that produces "ratio" levels of measurement.

commitment that must be checked. One needs to focus, with the aid of the DLD framework, on the range of possible conditions under which invariance is expected to hold. It depends, then, on the type (or strength) of inferences one wants to draw. Empirically, one goes about investigating the possible conditions and bounds via psychometric methods such as differential item functioning and test level invariance (for recent overviews please see Maller et al., in press; Zumbo and Hubley, 2003). This is a different spin on the use of these psychometric methods; they are not just about fairness issues but rather about empirically investigating the range of possible conditions and hence informing the conclusions (bounds and limitations) of the DLD framework.

It is evident from the DLD framework that invariance is nonsensical without a clear statement about the range of possible conditions of sampled and unsampled items (tasks) and/or sampling units (respondents or examinees). Likewise, invariance is not just a property of item response theory models. Zimmerman and Zumbo (2001) showed that some of the properties and quantities from classical test theory hold for the entire population of individuals, as well as any subpopulation of individuals. Furthermore, many of the examples of invariance we prove in the Zimmerman and Zumbo (2001) paper are of the variety seen in all classical mathematical statistics. For example, we proved that measurements that are parallel in a given population are also parallel in any subpopulation. This was a point also emphasized by Lord and Novick (1968). We provide other new results of the flavor seen in the sentence above about parallel tests. It, of course, should be appreciated that there are quantities in observed score classical test theory that are population specific (and lacking invariance). These include variance of the true scores and, in turn, reliability because it is bound to the variance of the true scores.

It is important to acknowledge that, in some validation settings, there will be an inherent tension that is difficult to balance between the (i) exchangeability of sampled and unsampled items in the target construct/domain (i.e., sampled tasks or items), and (ii) exchangeability of sampled and unsampled units in the target population (i.e., of individuals). This sort of tension is probably less likely in large-scale testing organizations than in research settings with one-time and small-scale measurement. In practice, what should be the trade-off in this tension? Which axis has precedent? In some important senses, the more exchangeable your items or tasks are, perhaps via a longer test, the more difficulty one faces obtaining large numbers of sampling units. Simply put, it is more difficult to obtain large numbers of respondents with very long tests due to the time demands. In this sense, there is usually (but not necessarily always) a trade-off that needs to be balanced between the two forms of exchangeability. In my opinion, the balance between exchangeability of items and respondents should have item exchangeability as the primary concern. There are easily many counter-examples to this advice but, generally, one needs to first know what they are measuring and then find the limits of the inferential bounds on respondents (i.e., sampling units). That is, one should always have at least an adequate amount of item or task exchangeability and then focus on sampling exchangeability. Quite simply, first establish what you are measuring and then try to generalize it. If you do not have some semblance of item and task exchangeability, then you really have nothing to generalize. By this I am suggesting that by establishing item exchangeability one will have to know what they are measuring.

Considering our example with the CES-D, we do not have the inferential strength for general measurement inference because the sample is a convenience sample of respondents. Depending on the exchangeability of the sampled and unsampled items, we are limited to either specific domain or initial calibrative inferences. The question of the exchangeability of the sampled and unsampled items of the CES-D has, to my knowledge, never been directly addressed in the research literature. For example, I know of no content validation studies of the CES-D, which would go a long way in establishing the exchangeability of the items. My sense, however, is that there is moderate support for the item exchangeability and hence our study is limited to specific domain inferences, with the limitations and strengths that this form of inference provides as described above. In addition, although some applied measurement books and articles state otherwise, fitting a Rasch model would not, in and of itself, increase my claims to invariance, generalizability, or validity.

2.4. Tests/measures and indices: Effect (reflective) versus causal (formative) indicators, respectively, and validity

In validity research, it is important to distinguish between measures and indices, also referred to as reflective versus formative measures, respectively. For my purposes a measure is defined as a latent variable for which a shift in the value of the latent variable leads to an expected shift in the value of the indicator (i.e., the items are effects indicators). An index is defined as a latent variable for which a shift in the value of the indicator variable (the observed or manifest variable) leads to an expected shift in the value of the latent variable. Consider an example provided by Bollen and Lennox (1991) wherein responses to a series of items on a math test should reflect a student's quantitative ability. An index, on the other hand, has the observed variables (e.g., self-report items) as causal indicators. The typical example, as described by Bollen and Lennox, is the index socio-economic status. They go on to suggest that socio-economic status might be better viewed as formative (an index) but is often treated as reflective.

There are few empirical tests of whether a latent variable is a measure or index, the exception is the vanishing tetrads test of Bollen and Ting (2000). Please note that computing a principal components analysis (PCA) is not sufficient evidence that one has an index, nor does fitting a factor analysis model provide sufficient evidence to claim one has a measure – as you will see below both models could fit the same data. Bollen and Lennox (1991), suggest that a good place to start, and often the only thing available, is a literal thought experiment.

I might add that one can also supplement this thought experiment with a content validation study wherein one asks subject matter experts to consider and rate whether the items (or indicators) are effects or causal; that is, whether the variable is a measure or index, respectively. One can build on the methodologies described for content validity in Sireci (1998) by incorporating questions about whether an item should be considered a causal or effects indicator using methodology in content validity including the coefficients, designs, etc. Also, one could investigate the source of the decision of effects versus causes by talk-aloud protocols and/or by conducting multidimensional scaling of the subject matter experts' judgments. These approaches aid the validity arguments

of whether one has effect or causal indicators. What I am suggesting is an extension of Bollen and Lennox's thought experiment to include data from subject matter experts.

My purpose for raising the matter of measures versus indices is that the validity evidence, particularly the empirical validity evidence, may, as Bollen and Lennox suggest, be quite different. The DLD framework, described above, however, will apply to both measures and indices. Therefore, the central matter is that most, if not all, validity research assumes reflective measures; that is, effects indicators. More thought and empirical support as I describe above, needs to be given to deciding on whether one has a measure or index.

Furthermore, it can be shown that the structural equation model for indices (causal indicators) is, in essence, principal components analysis (PCA) whereas the same model for measures (effect indicators) is factor analysis. Factor analysis and PCA are not the same model statistically nor conceptually. Furthermore, PCA is conceptually inappropriate for measures (effects indicators). There may be some statistical reasons for favoring PCA (for example, one may want to avoid the indeterminacy of factor analysis) but in validity research factor analysis is the standard. One should not (mis)use the fact that (a) there are circumstances (e.g., when the variances of the error terms in the factor model are roughly equal or are small) under which PCA is a good approximation to factor analysis, and (b) it is this good approximation that provides the basis for the widespread use of PCA statistics (e.g., number of eigenvalues greater than one, scree plot, or parallel analysis) in the preliminary stages of exploratory factor analysis to choose the number of factors, as reasons for arguing that PCA is the same as factor analysis.

It should be noted, however, that the "causal" perspective of factor analysis should not be interpreted in a narrow sense of the word "cause". The history of factor analysis is replete with debates about the nature of the factors that are determined/uncovered/discovered by factor analysis. I believe it would be too generous of a reading of Bollen and Lennox (1991), as well as the histories of sociology and of factor analysis, to interpret the causal language in a narrow sense. In this light, there is much we can learn about causation from a well-known book on the logic of causal inference by Cook and Campbell (1979) as well as work by Freedman (e.g., Freedman, 1987). Also see Shadish et al. (2002), which has among other things a useful discussion of manipulationist as opposed to non-manipulationist ideas of causation. Also, of course, there is much to be gained from the elegant statistical model for causation proposed by Neyman (1923). In short, one should not interpret the "arrows" in the graphical depictions of structural equation models of factor analysis in a narrow and strict sense of causation.

In the view I am espousing one should instead seek "explanation", and hence factor analysis becomes an explanatory model. As I note above in my critique of Borsboom et al. (2004), I consider the core of validity as one of a quest for explanation for the variation in the observed task or item responses. In this view, factor analysis is a statistical tool to aid in my inference to the best explanation, more broadly defined than just causal explanations.

One should note, as highlighted by Bollen and Lennox, that what we would expect for the composite score for a measure and index to be different in terms of validity. That is, for a measure the indicators or items need to be correlated because it is the correla-

tion that is the source of the measure, in a factor analytic sense. Furthermore, the factor analysis aims to reproduce the covariance among the items therefore there needs to be substantial correlation for the factor analysis to be meaningful. For an index, however, there is no need to expect a correlation among the items or indicators. Also, for a measure one thinks of reliability of measurement as a quality in the internal consistency sense, but for an index that may not make sense in all settings – e.g., whether the indicators of SES are internally consistent is *not* something of concern because it is an index and not a measure. Therefore, the validity evidence (including the dimensionality and error of measurement aspects of validity evidence) could be quite different for an index versus a measure. The point, however, is that this chapter deals with the validity of measures, some of which will be relevant for indices (e.g., the DLD framework) but not all.

Returning to our example of the CES-D, the literature is somewhat mixed. Many, if not most, of the users of the CES-D treat it as a measure. Bollen and Lennox (1991) refer to their work creating some sort of hybrid latent variable with some effects indicators and other causal indicators, but this hybrid model is not, to my knowledge, widely used. I will follow the convention in the depression literature and consider the CES-D as a measure and hence the items as effects; leading us to later use of factor analysis; and not principal components analysis. Furthermore, I fully expect high correlations among the items and to be able to examine the internal consistency of the responses as a way of quantifying error of measurement.

3. Statistical methods

Having dealt with foundational issues, let us now turn to statistical methods for validation studies. There is an extensive and easily accessible literature on the statistical methods of validation research in most introductory and intermediate measurement and assessment texts. Much has been written in the area of educational and achievement testing. An important paper in the context of personality and attitude scales is Thissen et al. (1983).

There is, however, one class of methods that are particularly central to the validation process: structural equation models. These models are particularly important to test validation research because they are a marriage of regression, path analysis, and latent variable modeling (or factor analysis). Given that the use of latent variable structural equation models presents one of the most exciting new developments with implications for validity theory, this section discusses these models in detail. At this point, let me now turn to the two empirical validity questions listed earlier: (a) Given that the items are combined to created one scale score, do they measure just one latent variable?, and (b) Given a set of predictor variables of interest can one order them in terms of importance?

3.1. Factor analysis: Given that the items are combined to create one scale score, do they measure just one latent variable?

The question of whether the items measure just one latent variable is addressed by using factor analysis. In the typical confirmatory factor analysis (CFA) model, the score

obtained on each item is considered to be a linear function of a latent variable and a stochastic error term. Assuming p items and one latent variable, the linear relationship may be represented in matrix notation as

$$y = \Lambda\eta + \varepsilon, \tag{1}$$

where y is a $(p \times 1)$ column vector of *continuous* scores for person i on the p items, Λ is a $(p \times 1)$ column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, η is the latent variable score for person i, and ε is $(p \times 1)$ column vector of measurement residuals.

In the example with the CES-D, a CFA model with one latent variable was specified and tested using a recent version of the software LISREL. Because the CES-D items are ordinal in nature (i.e., in our case a four-point response scale, and hence not continuous), a polychoric covariance matrix was used as input for the analyses. Using a polychoric matrix is an underlying variable approach to modeling ordinal data (as opposed to an item response theory approach). For a polychoric correlation matrix, an underlying continuum for the polytomous scores is assumed and the observed responses are considered manifestations of respondents exceeding a certain number of latent thresholds on that underlying continuum. Conceptually, the idea is to estimate the latent thresholds and model the observed cross-classification of response categories via the underlying latent continuous variables. Formally, for item j with response categories $c = 0, 1, 2, \ldots, C - 1$, define the latent variable y^* such that

$$y_j = c \quad \text{if } \tau_c < y_j^* < \tau_{c+1},$$

where τ_c, τ_{c+1} are the latent thresholds on the underlying latent continuum, which are typically spaced at non-equal intervals and satisfy the constraint $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{C-1} < \tau_C = \infty$. It is worth mentioning at this point that the latent distribution does not necessarily have to be normally distributed, although it is commonly assumed so due to its well understood nature and beneficial mathematical properties, and that one should be willing to believe that this model with an underlying latent dimension is actually realistic for the data at hand.

Suffice it to say that an examination of the model fit indices for our example data with the CES-D, such as the root mean-squared error of approximation (RMSEA = 0.069, 90% confidence interval for RMSEA of 0.063 to 0.074), showed that the one latent variable model was considered adequate. This finding of unidimensionality is important because the single composite score of the CES-D items can, therefore, be easily interpreted as a continuum of depressive symptomology – an increasing score indicating more depression. As well, quantitative indices of measurement error, such as reliability coefficients, are also easily interpreted because of the unidimensionality.

3.2. Regression with latent variables, MIMIC models in test validation: Given a set of predictor variables of interest can one order them in terms of importance?

In the context of the CES-D example, are the age and gender of the respondents both important predictors of the latent variable score on the CES-D? This question is often addressed by using ordinary least-squares (OLS) regression to regress the observed

composite score of the CES-D onto age and the dummy coded gender variable. The problem with this approach is that the regression results are biased by the measurement error in the observed composite score. Although widely known among statisticians, this bias is unfortunately ignored in most day-to-day validation research.

The more optimal statistical analysis is to use SEM and MIMIC models. MIMIC models, first described by Jöreskog and Goldberger (1975), essentially posit a model stating that a set of possible observed explanatory variables (sometimes called predictors or covariates) affect latent variables that are themselves indicated by other observed variables. In our example of the CES-D, the age and gender variables are predictors of the CES-D latent variable, which itself is indicated by the 20 CES-D items. Our example highlights an important distinction between the original MIMIC models discussed over the last three decades and the most recent developments in MIMIC methodology. In the original MIMIC work, the indicators of the latent variable(s) were all continuous variables. In our case, the indicators for the CES-D latent variables (i.e., the CES-D items) are ordinal or Likert-type variables. This complicates the MIMIC modeling substantially and, until relatively recently, was a major impediment to using MIMIC models in validation research.

The recent MIMIC model for ordinal indicator variables is, in short, an example of the merging of statistical ideas in generalized linear models (e.g., logit and probit models) and structural equation modeling into a generalized linear structural modeling framework (Jöreskog, 2002; Jöreskog and Moustaki, 2001; Muthen, 1985, 1988, 1989). This new framework builds on the correspondence between factor analytic models and item response theory (IRT) models (see, for example, Lu et al., 2005) and is a very general class of models that allows one to estimate group differences, investigate predictors, easily compute IRT with multiple latent variables (i.e., multidimensional IRT), investigate differential item functioning, and easily model complex data structures involving complex item and test formats such as testlets, item bundles, test method effects, or correlated errors with relatively short scales, such as the CES-D.

A recent paper by Moustaki et al. (2004) provides much of the technical detail for the generalized linear structural equation modeling framework discussed in this chapter. Therefore, I will provide only a sketch of the statistical approach to motivate the example with the CES-D. In this light, it should be noted that these models can be fit with either Mplus or PRELIS-LISREL. I chose to use the PRELIS-LISREL software and hence my description of the generalized linear structural equation model will use Jöreskog's notation.

To write a general model allowing for predictors of the observed (manifest) and latent variables, one extends Eq. (1) with a new matrix that contains the predictors x

$$y^* = \Lambda z + Bx + u, \quad \text{where } z = Dw + \delta, \tag{2}$$

u is an error term representing a specific factor and measurement error, and y^* is an unobserved continuous variable *underlying* the observed ordinal variable denoted y, z is a vector of latent variables, w is a vector of fixed predictors (also called covariates), D is a matrix of regression coefficients and δ is a vector of error terms which follows a $N(0, I)$. Recall that, in Eq. (1), the variable being modeled is directly observed (and assumed to be continuous), but in Eq. (2) it is not.

Note that because the PRELIS-LISREL approach does not specify a model for the complete p-dimensional response pattern observed in the data, one needs to estimate the model in Eq. (2) with PRELIS-LISREL in two steps. In the first step (the PRELIS step), one models the univariate and bivariate marginal distributions to estimate the thresholds and the joint covariance matrix of y^*, x, and w and their asymptotic covariance matrix. In the PRELIS step, there is no latent variable imposed on the estimated joint covariance matrix, hence making that matrix an unconstrained covariance matrix that is just like a sample covariance matrix, S, for continuous variables. It can, therefore, be used in LISREL for modeling just as if y^* was directly observed using (robust) maximum likelihood, weighted least-squares estimation methods, etc. Of course, one can use unweighted least-squares and the parameter estimates are consistent, but the standard errors are inaccurate. Thus, the unweighted least-squares is a descriptive, rather than inferential, technique in this context – this will be become important later when we use (finite sample) variable ordering methods based on least-squares types of estimation.

3.3. Variable ordering: Pratt's measure of variable importance

Due to space limitations, I will only provide a sketch of Pratt's variable ordering measure (please see Pratt (1987) and Thomas et al. (1998) for the details). Pratt considered a linear regression of the form

$$y = b_0 + b_1 x_1 + \cdots + b_p x_p + u, \tag{3}$$

where the disturbance term u is uncorrelated with $x_1, x_2, \ldots, x_p$ and is distributed with mean zero and variance σ^2. It is important to note that the linear regression in Eq. (3) is rather generic and hence can apply when the predictors are latent variables as in Eq. (2), the MIMIC model.

The total (standardized) population variance explained by the model in Eq. (3) can be written as

$$R^2 = \sum_j \beta_j \rho_j, \tag{4}$$

where β_j is the usual standardized regression coefficient corresponding to x_j, and ρ_j is the simple correlation between y and x_j. Pratt justified the rule whereby relative importance is equated to variance explained, provided that explained variance attributed to x_j is $\beta_j \rho_j$ – a definition which is widely used in the applied literature (e.g., Green et al., 1978), but as documented by Pratt (1987) and Thomas et al. (1998), it has also been criticized. Pratt justified the measure using an axiomatic approach based largely on symmetry and invariance to linear transformation. He showed that his measure is unique, subject, of course, to his axioms. An additional feature of Pratt's measure is that it allows the importance of a subset of variables to be defined additively, as the sum of their individual importance irrespective of the correlation among the predictors. Other commonly used measures (e.g., the standardized (beta-)weights, the t-values, unstandardized b-weights) do not allow for an additive definition and are problematic with correlated predictor variables.

Thomas et al. (1998) gave a sample interpretation of Pratt's measure based on the geometry of least squares. We considered a sample of N observations fitted to a model

of the form (3), so that the observed variables $y, x_1, \ldots, x_p$ comprise vectors in R_N. We assumed, without loss of generality, that the variables are centered to zero,

$$y'1_N = x_1'1_N = \cdots = x_p'1_N = 0,$$

where 1_N is an $N \times 1$ vector of ones. In this case, $\hat{y}$, the fitted value of y, is the projection of y onto the subspace X spanned by the x_i, and has the representation

$$\hat{y} = \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p \tag{5}$$

where the $\hat{b}_j$'s are least squares estimates of the population regression coefficients, b_j, $j = 1, \ldots, p$. We defined the partition of R^2 of x_j, $j = 1, \ldots, p$, to be the signed length of the orthogonal projection of $\hat{b}_j x_j$ onto $\hat{y}$, to the length of $\hat{y}$. By definition, this ratio represents the proportion of R^2 and sums to 1.0. Furthermore, the partitioning is additive so that one could, for example, compute the proportion of R^2 attributable to various subsets of the explanatory variables, irrespective of the correlation among the explanatory variables.

One then can partition the resulting R^2 by applying

$$d_j = \frac{\hat{b}_j r_j}{R^2}, \tag{6}$$

where, as above, $\hat{b}_j$ is the jth standardized regression coefficient (the "beta"), r_j is the simple Pearson correlation between the response and jth explanatory variable in Eqs. (3) and (5).

Several points are noteworthy at this juncture. First, all of the components of the right hand side of Eq. (6) are from a least squares procedure. Second, one can easily investigate "suppressor" variables (also sometimes referred to as "confounders") through the geometric approach described above. By definition, suppressors are variables that are not individually related to the response variable, but do make a significant contribution to the statistical model in combination with one or more of the other explanatory variables. In practice, a suppressor is identified when one has a relatively small value for d_i and a standardized regression coefficient (in the multiple regression model) comparable to the values exhibited by explanatory variables whose d_i are appreciably larger. Third, the index in Eq. (6) can be negative, a result which is counter-intuitive and can be seen as a negative feature of the index. See Thomas et al. (1998) for a discussion of procedures for addressing negative d_i of large magnitude. Fourth, when referring to variable-ordering as a statistical procedure, we are referring to ordering the predictor (explanatory) variables in a regression analysis in terms of *relative importance* as predictors of some dependent (response) variable. Therefore, the relative importance, in our context, is defined for a particular model after having decided on the "best model". Therefore, this is not a procedure for model-building (like stepwise regression). Rather, it is useful in determining the relative importance of predictors after one has decided on the "best model" by the relative proportion of variation accounted for by the predictors. Please see Thomas (1992) and Thomas and Zumbo (1996) for a discussion of variable ordering in the MANOVA and discriminant analysis cases, as well as, Wu et al. (2006) for variable ordering in factor analysis, and Thomas et al. (2006) for variable ordering in logistic regression, which would all be of use in validity studies.

Turning to the CES-D example, the validity researcher may be interested in the question of whether age and gender are predictive of CES-D scale scores. I will conduct the regression two ways: (a) the conventional way with the dependent variable being the observed total score across the 20 items, and (b) using the MIMIC model described above with the latent depressive symptomology variable as the dependent variable. In the latter case, I will use unweighted least-squares[5] estimation in LISREL so as to be able to apply the Pratt index for variable ordering. In both cases, the predictor variables are the respondents' age and gender. Note that the respondents' gender was coded 0 for male and 1 for female respondents. The correlation of age and gender is, as expected from the descriptive statistics of age for each gender, negative.

The ordinary least-squares results, as conventionally performed in validation research, results in an R-squared of 0.013, a value that is small but not unusual in this area of research given that we have only two predictor variables. The standardized b-weights (i.e., the beta weights) for gender and age are 0.064 and −0.084, respectively. The correlations between the observed composite score over the 20 CES-D items are 0.077 and −0.094 for gender and age, respectively. The corresponding Pratt indices computed from Eq. (6) are 0.385 and 0.615 for sex and age, respectively. Therefore, 61.5% of the explained variation (i.e., the R-squared) in the observed CES-D total scale score is attributable to age of the respondents; this makes age the more important of the two predictors. Note that the reliability of the 20 item scale was 0.91 for our sample, as estimated by coefficient alpha. Although introduced herein as a validation research tool for the first time in the research literature, I have already helped implement Pratt indices in an operational validity research program for the General Management Admissions Test (GMAT®) school-based validation reports.

In contrast to the observed score regression, one can estimate the regression of the latent variable (presumably of depressive symptomology) measured by the 20 CES-D items on to age and gender via the MIMIC model in Eq. (2). One can then use the unweighted least-squared parameter estimates (including the ULS model R-squared) and compute the Pratt indices. The Pratt indices can be applied to Eq. (2) because Pratt's original work, and the later geometric work by Thomas et al. (1998) are valid for a very generic regression model described in Eq. (3). The MIMIC model was fit using LISREL, as described above, specifying the standardized solution to get the beta-weights. If one examines the statistical values in the LISREL output, the standardized b-weights are 0.117, and −0.165, for gender and age, respectively. Likewise, the correlations of gender and age with the latent variable are 0.141 and −0.182, respectively, with an R-squared of 0.046. As expected, the MIMIC R-squared is larger than the observed variable R-squared reported above (Lu et al., 2005). Note that this discrepancy in R-squared is event even though our test reliability is 0.91 for our sample, as described above.

[5] Note that the geometry we introduce necessitates (in the finite sample case) a least-squares estimation method, of which weighted least-squares is of course a case. Because, of the limitations described of using weighted least-squares with in our MIMIC model, I do not focus on the standard errors and instead limit the analyses to the consistent parameter estimates and constrain my conclusions to those descriptive in nature, rather formal statistical inference.

It is precisely due to the negative bias of the R-squared for the observed score regression variable model that I recommend using the latent variable MIMIC model in validation studies. Herein, I introduce the Pratt index for the MIMIC model so variable ordering can also be investigated. The corresponding Pratt indices for the MIMIC model are 0.355 and 0.645 for gender and age; these results are similar to those of the observed variable regression above, with age being the more important predictor.

Finally, given the coding for the gender variable, as a group the female respondents scored higher on the latent variable of depression. Likewise, the older respondents tended to have a lower level of depression compared to the younger respondents in this sample, as reflected in the negative regression coefficient. When the predictive relationship of age was modeled in a separate analysis for males and females via this generalized MIMIC model, age was an important (negative) predictor for the female respondents but age was not an important predictor for male respondents. Age is unrelated to depression level for men whereas older women in this sample are less depressed than younger women – there may a nonlinear relationship here. These findings contribute to the empirical underpinnings of CES-D score interpretation. Therefore, this sort of known-groups information is useful to researchers using the CES-D and hence supports, as described at the beginning of this chapter, the inferences made from CES-D test scores and motivates the discussions of value implications and social consequences of CES-D score use.

4. Closing remarks

This chapter had two aims, one fairly general and the other fairly specific. The general aim was to provide some observations and comments on the variety of psychometric models, methods, and theories currently at our disposal, whether for validity and more generally in the wide-sense application of measurement. With this general aim, I intended to go beyond a simple cataloging and description of the variety of strategies, analyses, and statistical applications in validity and draw on my experiences and knowledge in psychometrics, pure and applied mathematics, statistical science, and philosophies of science to shed some light on validity theory and the practice of validation.

The more specific aim of this chapter was to first discuss several foundational issues including presenting a framework to consider the strengths of our measurement inferences depending on the data at hand. Next, I considered statistical methods that are particularly relevant to validation practice. My reading of the vast literature on validity theory and practice dating back to the early parts of the 20th century leaves me with the impression that the history of psychometrics and measurement validity exhibits both a pattern of growing understanding and utility and a series of unending debates on topics of enduring interest. If one spends just a few moments reflecting on the history of many different sciences one will see that this is characteristic of a maturing science.

Integrating and summarizing such a vast domain as validity invites, often rather facile, criticism. Nevertheless, if someone does not attempt to identify similarities among apparently different psychometric, methodological, and philosophic views and to synthesize the results of various theoretical and statistical frameworks, we would

probably find ourselves overwhelmed by a mass of independent models and investigations with little hope of communicating with anyone who does not happen to be specializing on "our" problem, techniques, or framework. Hence, in the interest of avoiding the monotony of the latter state of affairs, even thoroughly committed measurement specialists must welcome occasional attempts to compare, contrast, and wrest the kernels of truth from disparate validity positions. However, while we are welcoming such attempts, we must also guard against oversimplifications and confusions, and it is in the interest of the latter responsibility that I write to the more general aim.

4.1. An overview: The melodic line, with some trills and slides

In addition to reviewing the validity literature from a psychometric perspective, this chapter includes three novel methodological contributions: (i) the DLD framework, (ii) the extension of the Bollen and Lennox thought-experiment to content validity studies with subject matter experts, and (iii) a latent variable regression method that allows for variable ordering.

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

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. There are many metaphors discussed in the literature for the process of validation: (a) the stamp collection, (b) chains of inference, (c) validation as evaluation, and (d) progressive matrices to name just a few. In my early work on validity I 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. I am sure I am not alone in the use of the courtroom metaphor. Today I think in less adversarial terms and rather think of it as jazz – as in the musical style. With validation as jazz I want to principally borrow the tenets of sound coming together, but that the coming together is not necessarily scripted. All sorts of notes, chords, melodies, and styles come together (including, of course, improvisation that is particular to that one song or performance) in a creative way to make music.

I was once told by a music teacher that he could not teach me to play jazz. Instead, he could teach me music and some jazz sounds, songs, and styles. It would then be my task to bring it all together, sometimes in an unscripted way, to make jazz music. Perhaps the same applies to the process of validation, there is no one methodology or script that can be applied in all measurement contexts.

There is, however, one framework that is general enough to have broad appeal. Because it was not developed with validation, per se, in mind, this framework has a limitation that needs to be addressed. That is, if one were to insert a component of developing and stating a theory of the phenomenon of interest, the portrayal of the measurement process by Hattie et al. (1999) is a rich and useful framework. A modified version of the Hattie et al. framework would involve a conceptual model (theory) of the phenomenon, leading to conceptual models of the measurement, leading to test and item (task) development, leading to test administration, test use and test evaluation, and back to the conceptual model of the phenomenon; with all the constituent elements and statistical psychometric methods described by Hattie and his colleagues in Figure 1 and the remainder of their chapter. It is noteworthy that there is a feedback back to the conceptual model of the phenomenon. Having said this, however, it is important to note the distinction I make between validity, per se, and the process of validation. I consider validity to be the establishment of an explanation for responses on tasks or items – the emphasis being inference to the best explanation as the governing aspect, and the process of validation informs that explanatory judgment. The modified Hattie et al. framework, therefore, is a useful general description of the process of validation, but is not validity, per se.

Another broad theme in this chapter involves establishing the bounds and studying the limits of our inferences. In this light there are many interconnected ideas I discuss above that deal with the purpose and use of models, as well as a new framework, the DLD framework, to incorporate both the inferences to a domain (or construct) and the inferences to a population of examinees or respondents to your measure.

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

4.2. In terms of the psychometrics of validity, when psychometricians speak, what are they really saying?

With the psychometrician's responsibility in mind, and the general aim of this chapter at hand, let me close by answering a question that I often hear: When psychometricians speak, what are they really saying? Or, put another way, when psychometricians do what they do, what are they doing?

Psychometricians use models to help us go from the data we have, to the data we wish we had, by augmenting our data with assumptions. These assumptions need to be both empirically and conceptually (i.e., from theory) validated. Furthermore, psychometricians use models that are descriptive in nature so that we usually do not know the "processes" involved and hence have, at best, a heuristic purpose. In this light, item response theory is hardly an elaborated process model. This point is important to keep in

mind when thinking of alternatives to psychometric models. This lack of explanatory focus has been the root of a long-standing angst among some measurement specialists. The most recent attempt at relieving this angst has been to prevail on cognitive theory to lend an explanatory hand. My only observation on this front is that not all cognitive theories are explanatory so that we need to be careful that, in our quest for explanatory power, we do not inadvertently supplant one heuristic model with another while deluding ourselves that our new model is explanatory. Furthermore, these cognitive theories need to be empirically tested – see, for example, Zumbo et al. (1997) wherein we tested a social-cognition theory of item discrimination.

Psychometricians often have their favorite model. As I noted above, like Pygmalion we sometimes fall in love with our models. It is confusing to hear that "all the models are the same" and "my model is different and you should use it everywhere and all the time". For example, it is confusing for practitioners (and psychometricians) to hear that all models are realizations of factor analysis or generalized linear latent variable models from one quarter, and then hear from another quarter that the Rasch model has special properties that no other model allows, or that generalizability theory is limited because it does not allow invariance of examinees or items. The reason for this confusion is that one can make convincing arguments for the unification of measurement models (see McDonald, 1999; Zumbo and Rupp, 2004; Zimmerman and Zumbo, 2001) and, at the same time, convincing arguments about, for example, the advantages of *generalizability theory* over classical test theory, item response theory over *generalizability theory*, item response theory over classical test theory, and structural equation modeling over *generalizability theory*, classical test theory, and item response theory, and so on (Zumbo and Rupp, 2004). This is not uncommon in mature mathematical disciplines and may well be due, in part, to the distinction between writing a model as a mathematical statement on paper versus the additional conditions that need to be invoked to estimate the parameters in these parameter-driven models. In this light, it is important to note that, in practice, models are not just defined by how they are written on paper but also by the parameterization and estimation with data.

It may come as a surprise to some who advocate the exclusive use of one measurement model in test validation research (e.g., the claim that the Rasch model should be used exclusively to validate all health outcome measures) that it can be shown that IRT (and some forms of factor analysis) is a first-order approximation to generic (classical) test theory model. This is somewhat expected once one recognizes that the generic classical test theory model statement $X = T + E$ is axiomatic to all measurement models (Lord, 1980). The different measurement models arise depending on how one defines T and E; if T is a latent variable, you have IRT or factor analysis, but if T is a complex model statement of the measurement study with facets and domains, one may have a mixed or random effects ANOVA and hence generalizability theory. Depending on the strength of the T and E assumptions one may have weak or strong forms of classical test theory. In fact, one can consider the generic measurement model statement, $X = T + E$, on par with the generic regression model statement in Eq. (3). If you apply the geometry in Zimmerman and Zumbo (2001) one can show that classical test reliability is, in essence, a Pratt index – a partitioning of the explained variation in the test score attributable to the model, just like an R-squared value in regression.

Advocating one measurement model for all occasions is an overstatement. At best, one can say that one particular model might be useful in characterizing and solving several measurement problems or that, depending on the test validation needs, one instantiation of a model with its corresponding parameterization and estimation methods may be more appropriate.

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

Therefore, as formalized in the DLD framework I described in Section 2.3, two distinct directions of generalizability are typically of interest, which require an understanding of the reliability and validity properties of scores and inferences. As Zumbo and Rupp (2004) note, first, it is of interest to make statements about the functioning of a particular assessment instrument for groups of examinees who share characteristics with those examinees who have already been scored with it. Second, it is of interest to make statements about the functioning of item sets that share characteristics with those items that are already included on a particular test form. For example, it is often of interest to show that the scores and resulting inferences for different examinee groups are comparably reliable and valid if either the same instrument is administered to the different groups, a parallel version of the instrument is administered to the different groups, or selected subsets of items are administered to the different groups. This also specifically implies that researchers should report estimates of reliability coefficients and other parameters for their own data rather than relying on published reports from other data and that validity needs to be continually assessed rather than being taken for granted based on prior assessment calibration(s).

In summary, then, the item (or task) responses created by the interaction of examinees with items (or tasks) on a measure are considered to be *indicators* or markers of unobservable or *latent* variables. I use the term *latent variable* to refer to a random variable that is deliberately constructed or derived from the responses to a set of items and that constitutes the building block of a statistical model (e.g., θ scores in IRT or factor scores in factor analysis). The statistical problem of measuring a latent variable can be characterized as involving two key tasks: (a) to find a set of indicators (items, scales, tasks, performances, or more generally referred to as measurement opportunities) that we believe that the latent variable will imply, and (b) to find a methodology for constructing a summary measure or scalar measure of the latent variable from these indicators. Denoting the set of indicators by $x = (x_1, x_2, \ldots, x_q)$ the second part of the problem is to find a function $\varphi(x)$ so that the numerical value of φ can be regarded as an appropriate scalar measure of the unobserved or latent variable. In this light, it is important to keep in mind that the main goal of modeling test data should always be to

make valid inferences about the examinees but inducing latent variables into the data structure cannot mechanically increase the validity of these inferences. No matter how sophisticated the psychometric model, the statement of $\varphi(x)$, and estimation routines have become, a test with poor validity will always remain so.

References

Angoff, W.H. (1988). Validity: An evolving concept. In: Wainer, H., Braun, H.I. (Eds.), *Test Validity*. Lawrence Erlbaum, Hillsdale, NJ.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling* **12**, 411–434.

Asparouhov, T., Muthen, B.O. (2005). Multivariate statistical modeling with survey data. Presented at the Conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA (available via the web at http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf).

Bollen, K.A., Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* **10**, 305–314.

Bollen, K.A., Ting, K. (2000). A tetrad test for causal indicators. *Psychological Methods* **5**, 3–22.

Bond, T.G., Fox, C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Social Sciences*. Lawrence Erlbaum, Hillsdale, NJ.

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

Cook, T.D., Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston.

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

Cyr, A., Davies, A. (2005). Item response theory and latent variable modeling for surveys with complex sample designs: The case of the national longitudinal survey of children and youth in Canada. Presented at the Conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA (available via the web at http://www.fcsm.gov/05papers/Cyr_Davies_IIIC.pdf).

de Finetti, B. (1974–1975). *Theory of Probability (vols. 1–2)*. Wiley, New York.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics* **20**, 115–147.

Embretson, S.E. (1994). Applications of cognitive design systems to test development. In: Reynolds, C.R. (Ed.), *Cognitive Assessment: A Multidisciplinary Perspective*. Plenum Press, New York, pp. 107–135.

Freedman, D.A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics* **12**, 101–223 (with discussion).

Green, P.E., Carroll, J.D., DeSarbo, W.S. (1978). A new measure of predictor variable importance in multiple regression. *Journal of Marketing Research* **15**, 356–360.

Gullicksen, H. (1950). *Theory of Mental Tests*. Wiley, NY.

Hambleton, R.K., Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. In: Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences*. Kluwer, The Netherlands, pp. 153–171.

Hattie, J., Jaeger, R.M., Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education* **24**, 393–446.

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.

Jonson, J.L., Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement* **58**, 736–753.

Jöreskog, K.G. (2002). Structural equation modeling with ordinal variables using LISREL. Retrieved December 2002 from http://www.ssicentral.com/lisrel/ordinal.htm.

Jöreskog, K.G., Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* **10**, 631–639.

Jöreskog, K.G., Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research* **36**, 341–387.

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

Kaplan, D., Ferguson, A.J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling* **6**, 305–321.

Kristjansson, E.A., Desrochers, A., Zumbo, B.D. (2003). Translating and adapting measurement instruments for cross-cultural research: A guide for practitioners. *Canadian Journal of Nursing Research* **35**, 127–142.

Lindley, D.V. (1972). *Bayesian Statistics, a Review*. Society for Industrial and Applied Mathematics, Philadelphia.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale, NJ.

Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Lu, I.R.R., Thomas, D.R., Zumbo, B.D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling* **12**, 263–277.

Maller, S.J., French, B.F., Zumbo, B.D. (in press). Item and test bias. In: Salkind, N.J. (Ed.), *Encyclopedia of Measurement and Statistics*. Sage Press, Thousand Oaks, CA.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Erlbaum, Mahwah, NJ.

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: Wainer, H., Braun, H.I. (Eds.), *Test Validity*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 33–45.

Messick, S. (1989). Validity. In: Linn, R.L. (Ed.), *Educational Measurement*, 3rd ed. Macmillan, New York, pp. 13–103.

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: Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences*. Kluwer Academic Press, The Netherlands, pp. 35–44.

Mislevy, R.J. (1991). Randomization-based inferences about latent traits from complex samples. *Psychometrika* **56**, 177–196.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement* **33**, 379–416.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior* **15** (3–4), 335–374.

Moustaki, I., Jöreskog, K.G., Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling* **11**, 487–513.

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: Wainer, H., Braun, H. (Eds.), *Test Validity*. Lawrence Erlbaum, Hillsdale, NJ, pp. 213–238.

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

Muthen, B., Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology* **25**, 267–316.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10**, 1–51, in Polish. English translation by D. Dabrowska and T. Speed (1990). Statistical Science 5, 463–80 (with discussion).

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research* **64**, 575–603.

Pratt, J.W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In: Pukkila, T., Puntanen, S. (Eds.), *Proceedings of the Second International Tampere Conference in Statistics*. University of Tampere, Tampere, Finland, pp. 245–260.

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.

Rupp, A.A., Dey, D.K., Zumbo, B.D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling* **11**, 424–451.

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. *Educational and Psychological Measurement* **64** (2004) 991, Errata.

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

Shadish, W.R., Cook, T.D., Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.

Sireci, S.G. (1998). The construct of content validity. In: Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences*. Kluwer Academic Press, The Netherlands, pp. 83–117.

Thissen, D., Steinberg, L., Pyszczynski, T., Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement* **7**, 211–226.

Thomas, D.R. (1992). Interpreting discriminant functions: A data analytic approach. *Multivariate Behavioral Research* **27**, 335–362.

Thomas, D.R. (2001). Item response theory and its application to the National Longitudinal Survey of Children and Youth. Report prepared for the Survey Research Methods Division, Statistical Society of Canada.

Thomas, D.R., Cyr, A. (2002). Applying item response theory methods to complex survey data. In: *Proceedings of the Statistical Society of Canada, Section on Survey Research Methods*. Statistical Society of Canada, Hamilton, Ontario.

Thomas, D.R., Hughes, E., Zumbo, B.D. (1998). On variable importance in linear regression. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* **45**, 253–275.

Thomas, D.R., Zumbo, B.D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics* **21**, 110–130.

Thomas, D.R., Zumbo, B.D. (2002). Item response theory and related methods with application to complex sample surveys. Notes for the day long session presented at the Statistics Canada XIX International Symposium "Modeling survey data for social and economic research". Ottawa, Canada.

Thomas, D.R., Zhu, P.C., Zumbo, B.D., Dutta, S. (2006). Variable importance in logistic regression based on partitioning an R-squared measure. In: *Proceedings of the Administrative Sciences Association of Canada (ASAC) Conference*. Banff, AB.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice* **16**, 33–45.

Wu, A.D.L., Zumbo, B.D., Thomas D.R. (2006). Variable and factor ordering in factor analyses: using Pratt's importance measures to help interpret exploratory factor analysis solutions for oblique rotation. Paper presented at the American Educational Research Association Meeting. San Francisco, CA.

Zimmerman, D.W., Zumbo, B.D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing* **1**, 283–303.

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.

Zumbo, B.D. (2001). Methodology and measurement matters in establishing a bona fide occupational requirement for physically demanding occupations. In: Gledhill, N., Bonneau, J., Salmon, A. (Eds.), *Proceedings of the Consensus Forum on Establishing BONA FIDE Requirements for Physically Demanding Occupations*. Toronto, ON, pp. 37–52.

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., Gelin, M.N. (in press). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated test and item bias. *Educational Research and Policy Studies*.

Zumbo, B.D., Hubley, A.M. (2003). Item bias. In: Fernández-Ballesteros, R. (Ed.), *Encyclopedia of Psychological Assessment*. Sage Press, Thousand Oaks, CA, pp. 505–509.

Zumbo, B.D., MacMillan, P.D. (1999). An overview and some observations on the psychometric models used in computer-adaptive language testing. In: Chalhoub-Deville, M. (Ed.), *Issues in Computer-Adaptive Testing of Reading Proficiency*. Cambridge University Press, Cambridge, UK, pp. 216–228.

Zumbo, B.D., Pope, G.A., Watson, J.E., Hubley, A.M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement* **57**, 963–969.

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