

How to Build a Domain Theory

On the Validity Centered Design of Construct-Linked Scales of Learning and Growth

C. Victor Bunderson

Brigham Young University & The EduMetrics Institute

INTRODUCTION

Measurement experts, and especially those sympathetic to the traditions of these volumes, emphasize the importance of putting forth substantial efforts to *build useful scientific variables*, instrumented so as to yield measures, not mere scores. Building a domain theory is more than its instruments, but the development of the instrument and the development of the theory are synergistic. A great instrument is inseparable from a great theory. A domain theory gives an account of both sides of the person/item map – the substantive processes employed by the persons, and the construct-relevant sources of task difficulty. This article explains the concept of a domain theory, places it within other types of theory needed in educational research, and shows how it and its validity argument are built by giving attention to the design processes of validity-centered design.

The general problem of measuring learning and growth in educational or training domains is the background for the examples and discussion herein. In order to describe how to build a domain theory, with its necessary construct-linked measurement scales, this article is organized under the following four headings:

1. **Domain Theories are one of several interrelated theories, and good measurement instruments and good theory are inseparable.** A domain theory is one of several types of descriptive theory important in educational measurement and instruction. This paper deals with a domain-specific learning theory of progressive attainments, or domain theory for short. In learning domains, there are prescriptive theories as well– instructional-design theories, implementation design theories, and methods, models, and theories to guide measurement instrument development. The development of construct-linked scales makes it possible to link theory to data, and thus to test the predictions of domain theories. These scales in turn can be used to test the predictions of the other types of theory, which themselves may have been used in the development of the measurement instruments. Thus instruments and theories evolve together.
2. **What are domain theories, and in what way are they essential in measuring human differences in learning and growth?** This section defines domain theory and considers types of predictions that can be made and confirmed / disconfirmed based on them. It also shows the central role domain theories play in verifying causal inferences experimentally.
3. **A design discipline is needed: Validity-centered design of construct-linked scales.** This section traces the evolution of the unified validity model promulgated by Samuel Messick through his last paper in 1998. In this paper he used the term “constructing construct validity”. This term implies a design process. Validity-centered design not only considers the six aspects of construct validity discussed by Messick, but also adds three other aspects to round out the process.

4. **Other Issues of Importance in Developing Domain Theories.** Theory-based calibrations are needed to verify the predictions of domain theories. Rigorous design experiments are vital to obtaining the data to build a convincing validity argument for theories and their associated construct-linked scales.

DOMAIN THEORIES ARE ONE OF SEVERAL INTERRELATED THEORIES; GOOD MEASUREMENT INSTRUMENTS AND GOOD THEORY ARE INSEPARABLE

Two classes of theory are necessary, prescriptive and descriptive theories. In the area of measuring learning and growth, where educational measurement is usually focussed, there must initially be two kinds of prescriptive or design theories, one for the measurement instruments and one for the learning materials and environment. A design theory for the implementation process is also necessary. A formalized design process for measurement instruments would provide guidance for designing, developing, and improving the measurement instruments. Instructional-design theory for the learning system guides the process of designing, developing, and improving the instructional materials that can produce measured growth on the instrument's scales. Because of the history and philosophy of the field of measurement, the idea of design has been avoided, except in one area -- experimental design. Ironically, in experimental design, the need for two inter-related prescriptive theories is implicit. Consider Figure 1 below, a true experimental design from Campbell and Stanley (1966) that defines a simple transfer experiment. Random selection is used in an attempt to assure that the experimental group, *X*, and the control group, *C*, are equivalent in their starting positions on the outcome variable of interest, designated as *O*.

Simple Experimental Design – Transfer Experiment			
X group	R Random Selection	X eXperimental treatment	O Observation {Post-treatment measure}
C group	R Random Selection	{Control Group}	O Observation {Post-treatment measure}

Figure 1. A True Experimental Design to test a substantive process theory of transfer from *X* to *O*.

Experimental design attempts to assure the validity of causal inference – in this case that the treatment *X*, and not pre-existing variations in the groups *X* and *C*, is the cause of improvements in *O*. The field of educational measurement has not emphasized the notion that there is or can be a design discipline with prescriptive propositions – even theories -- for good

design of scientifically sound “O’s”. Experimental scientists have usually not emphasized the similar notion that there is or can be a design discipline for designing and developing the *X*’s – the treatments used in experiments to affect progress on the *O* measure. The nature of such a design process is considered further herein, largely in connection with the discussion of validity-centered design.

Prescriptive or design theories (and their predecessors), models or simply formal methods, transform an initial set of objects or ideas into useful products. Figure 2 uses a (transformative) arrow to depict a prescriptive theory for the *X*’s – instructional-design theory. Reigeluth (1983, 1987, 1999) has compiled many examples, and has defined the nature of such theories. Wiley (2000) used domain theory, discussed below, as a part of a theory-development effort (design theory) to specify learning objects and their sequencing. Good measurement practitioners teach disciplined processes of design and development for quality instruments, but these methods have not been formalized yet as a set of widely accepted prescriptive principles – a design theory. Some may feel it premature to call such a set of prescriptions a theory. At the least, there is a disciplined process of identifying and documenting a set of partly empirical, partly formalizable, teachable principles about the process of design. Such a disciplined process applies to the design of measurement instruments, instructional systems, implementation plans, and other artifacts of human construction.

Figure 2 symbolizes such a theory for learning and instruction. In Figure 2, information about needs, goals, objectives, and contents are transformed by the processes prescribed in the design theory into useful instructional products such as learning maps, learning objects, and lesson materials. Developing measurement instruments is a prescriptive design process as well, even though historically this process has been grossly oversimplified by calling it a process of “operationally defining the variable”. Another prescriptive design process is implementation design, in which the managing, monitoring, and adapting to local situations is planned and executed. This design process is discussed in more detail in Bunderson & Newby (2002).

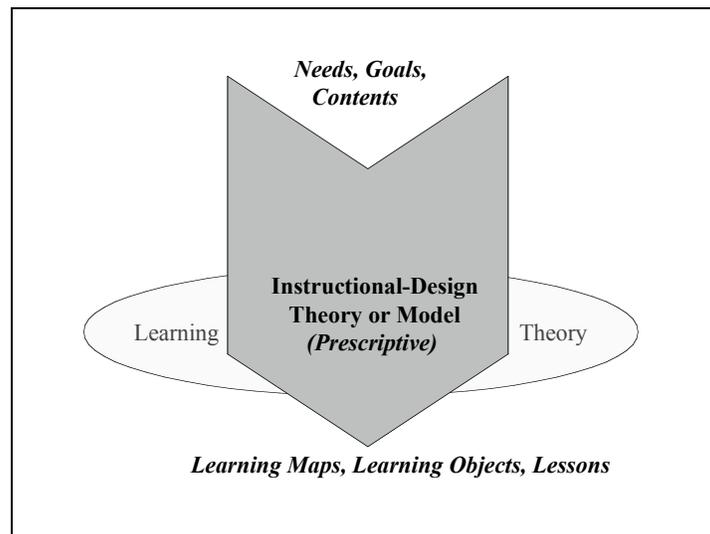


Figure 2. A prescriptive design theory and its background descriptive theory.

Descriptive theories are more common than prescriptive in psychology and the human sciences, and in Figures 2 and 3 are depicted not by transforming arrows, but by circles or

ovals. The oval behind the arrow in Figure 2 reminds us of the role of learning theory; that is, that principles consistent with learning theories, verified by learning research, are transformed into the prescriptive statements of instructional-design theories. Both descriptive principles and prescriptions require empirical confirmation within their scope of application. This scope of application may be delimited by means of two other types of descriptive theory, as depicted in Figure 3.

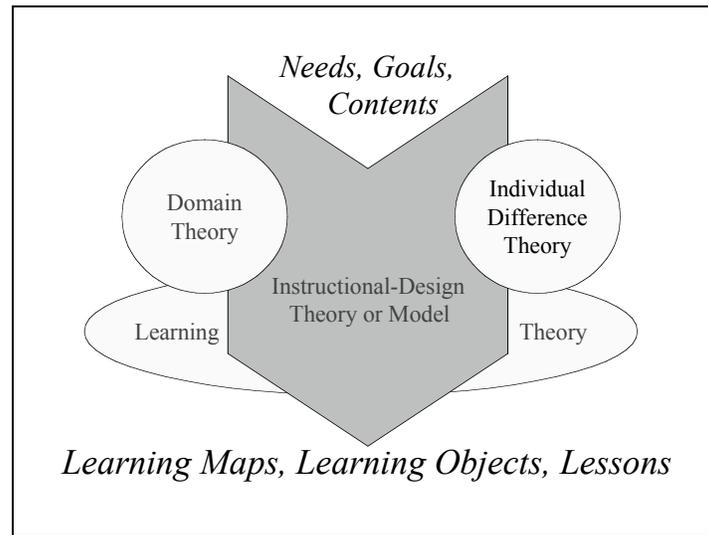


Figure 3. Two descriptive theories, domain theory and individual difference theory, are needed to guide – and to validate – prescriptions

Figure 3 adds two more descriptive theories to the mix of theories needed in the study of human learning and growth – domain theories and individual difference theories. Domain theories are specific to a particular content domain. Measurement scales and instruments designed to be linked to the constructs in the domain theories enable us to assess progress before, during, and after learning sequences. Individual difference theories describe individual variations in motivation, affect, personality, temperament, style, or cognitive or psychomotor abilities that cut across learning domains. This paper will deal extensively with only the first of these two descriptive theories, subject matter domain theories, even though the validity-centered design process and standards for a good validity argument apply to both.

Actually, these two types of descriptive theory are quite similar. Both can be conceived of as increasing levels of propensity to respond in a correct or particular way. In a subject-matter or job domain, increasing expertise enables one to do better in knowledge or performance of that subject matter. In the domain of individual differences in learning and thinking, we can also describe ordered levels of performance (perhaps just propensity to respond in a particular way). In either case, we need a domain theory to explain what is easy, what is somewhere between easy and hard, and what is hard. When we understand these things, we can measure, give meaningful feedback, and can teach people who desire to do so how to rapidly progress from lower to higher levels in either type of domain. These lead to the construction of learning maps that reflect progress up the domain scales, or reveal standing on a map of individual difference variables.

The quality of the set of measurement instruments and related maps that depict learner progress within a domain are entirely related to the quality of the domain theory. Moreover, the predictions of the domain theory can be confirmed or disconfirmed based on the measures from the instruments. The prescriptions from the design theory can also be confirmed or disconfirmed from experiments in which the instruments provide a construct-valid assessment of progress on the “*O*” variables. Thus, instruments for the “*O*’s”, instructional systems for the “*X*’s”, and their corresponding theories evolve together and are inseparable.

WHAT ARE DOMAIN THEORIES, AND IN WHAT WAY ARE THEY ESSENTIAL IN MEASURING HUMAN DIFFERENCES IN LEARNING AND GROWTH?

Domain Theory Defined

A domain theory is a descriptive theory of the contents, substantive processes, and boundaries of a domain of human learning and growth that gives an account of construct-relevant sources of task difficulty; and conjointly, an account of the substantive processes operative at different levels of growth along the scale(s) that span the domain.

Based on the constructs that account conjointly for difficulty and level of processing, and using measurement instruments linked to the constructs in the domain theory, testable predictions can be made about the relationships between tasks, processes, and locations along the scale(s).

Content and performance descriptions of domains. Instruments are developed to assess learning progress and performance in domains of human knowledge, proficiency, and accomplishment. Domains are usually signified by familiar content or topical names, like calculus, American history, accounting, network engineering, and nursing. Topics signify categories of what people may know, but they do not describe what people can do, so other uses of language, such as objectives or work-models are needed. These specify what the learner must do with contents, problems, questions, etc. Objectives, tasks, and models of what a proficient person does when performing the work (see Gibbons, et al. (1995) on work models), provide a performance-oriented transition language beyond verbal names. In this way, the general boundaries of the domain are initially established, but the process is not complete until a set of essentially unidimensional measurement scales is developed, which together define the scope and boundaries of the domain of learning and growth. Because perfect unidimensionality is impossible to achieve, we must settle for essential, or near, unidimensionality. Delineating a domain to a point where essential unidimensionality in each of a small set of domain-spanning scales is achieved is not always accomplished in practice, as it is often a time-consuming and lengthy process.

Scales that span the domain. Constructing a set of domain-spanning measurement scales elevates a good qualitative model of a domain to a theory. Extensive research is often needed to understand the nature of the substantive process constructs that explain the ordering from lesser to increasingly able learners. This research may start with qualitative probes to get an initial idea of the domain constructs, followed by the discovery of qualitative order relationships, eventually leading to a qualitative model of the domain. The construction of many tasks – items, testlets, performance rubrics – is necessary, as it provides confirming evidence of theoretical

propositions about how many essentially unidimensional scales are needed to span the domain. An increasingly mature domain theory identifies a set of unidimensional scales that approximate equal interval scales, each giving a precise parameterization of the expertise space of a specific knowledge domain. It has a pool of real and possible tasks calibrated on the scales that can be used to measure accurately learner progress within the domain. When a domain theory is achieved, measurement instruments and theory have gone hand in hand the whole way.

Domain Theory Gives a Conjoint Account of Tasks and Substantive Processes

The term “domain theory” in educational measurement was used by Messick (1995) in an important article on validity. According to Messick (1995):

“A major goal of domain theory *is to understand the construct-relevant sources of task difficulty*, which then serves as a guide to the rational development and scoring of performance tasks and other assessment formats. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed” (*italics added*).

In this statement, Messick stresses the task side of a scaling of persons and tasks. In other parts of this and other writings, he stresses the substantive processes (occurring “in the heads” of the people). A view that measurement in the human sciences can only aspire to be fundamental if it meets the conditions of additive conjoint measurement implies this truth: theorists must understand that the constructs of difficulty and thinking process underlying their scales are conjointly explained. It is not enough to theorize only about the tasks – columns in a data matrix, or about the people – rows in the data matrix. Measurements can emerge only through the interaction of both, so therefore any domain theory that gives the resulting scale any explanatory power must also give an account of both. The “construct-relevant” sources of task difficulty lie in the substantive processes, whether cognitive, affective, or psychomotor. Except for psychomotor, these processes are usually invisible, and their ordering along each scale is initially unknown. All processes and their order are a part of the domain-theoretic account. The substantive processes must be anchored to task features at landmark levels of task difficulty in order to give them empirical substance and interpretive utility.

Figure 4 opens up the usual “Wright map” of persons and items that conjointly gives the empirical display of task difficulty compared to person proficiency. In this figure person proficiency is given its own set of descriptors, ranging from beginner to advanced, while task difficulty increases along the horizontal axis.

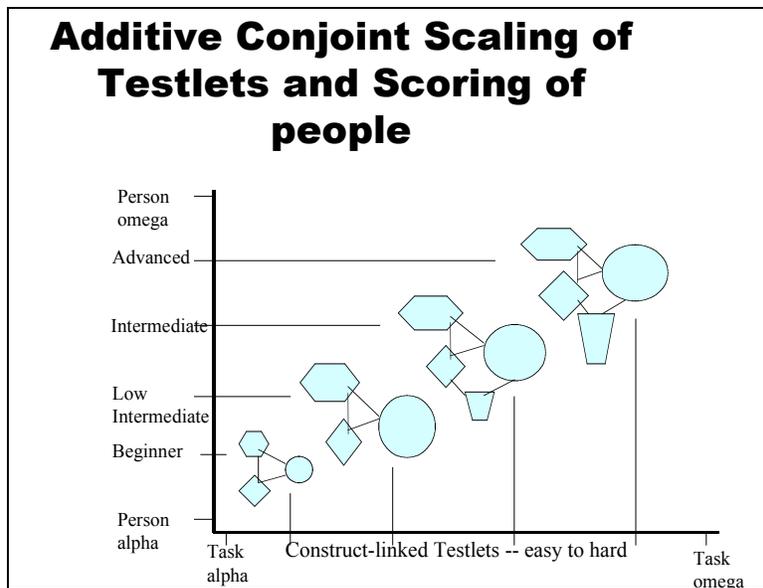


Figure 4. Substantive processes made visible by diagrams, and related both to levels of person proficiency and levels of task difficulty.

Along the horizontal axis are the tasks, ranked from easiest – task alpha – to the most difficult – task omega. The vertical axis ranks people from the most novice – person alpha – to the most advanced – person omega. Along the horizontal axis we have used the term “testlets” to refer to the need for greater attention to the design of a set of tasks with a close link to a theory of ordered substantive processes. A testlet will typically deliver more than the two states of right or wrong in conventional items. It thus gives more opportunity for identifying ordered levels of maturity of the processes. The shape and arrow diagrams in Figure 4 are so constructed as to imply a Guttman-like ordering of substantive processes. Each shape subsumes the lower level of the similar process. For example, the process represented by the diamond grows larger as proficiency and difficulty increase. A new process is added at a higher level (trapezoid) and grows at an even higher level. Nested sequences like this aid in designing for unidimensionality. We have found the use of the concepts of task and person alpha and omega very useful in delineating the boundaries of a domain. What is the lowest level of proficiency of a person who might barely be able to be admitted to an educational program in this domain? What is the simplest task that this person could barely pass? How may a person with proficiency just beyond the reach of an educational program be described, and what “task omega” could be constructed that this person could pass, but no one below could do?

Provided essential unidimensionality has been achieved, we can accurately link scores on the testlets to ability of the person (the vertical axis). There must be a theory of the distribution of expertise that permits the construction of stable 'meter-sticks' for measuring progress on each scale. The measurement resulting from the adaptive and asynchronous administration of these testlets permits a progress interpretation related to constructs of growing expertise for each individual, and for groups. As in adaptive testing, no two students need take exactly the same set of questions, nor follow the same sequence, as they progress through the learning domain.

Predictions Possible from Domain Theories

Once the construct-linked scale is developed, its predictive powers can be used. Given an unscaled task, we can predict its location among the previously scaled tasks. There may be situations in which a certain location along a scale is not being thoroughly tested. We can design specifications for a task that will calibrate at that location. More broadly speaking, given a description of a substantive process, we can predict its location. Given a location, we can design specifications for a task or testlet that will tap the processes at that location.

Figure 5 is useful in identifying the most basic predictions possible from a domain theory. Starting on the person side of the vertically arranged Wright map, we see a distribution of persons ordered by their calibrated theta measures. At certain points along the scale, our domain theory identifies a linkage to substantive processes operative at that level. Thus, given a description of one such process (designated by one of the right-pointing arrows), we can predict the location of persons for whom we have other evidence that they are just barely possessed of that substantive process. Similarly, because of our theory-based knowledge of the ordering of the substantive processes and their connection to difficulty features of item or testlets, we can design specifications for a task or testlet that will calibrate to the desired location. Moving to the task side, if we are given a task or testlet, we can determine by examination whether it is appropriate to the domain scale at all, and if so, approximately where it will calibrate when we are able to obtain data on it.

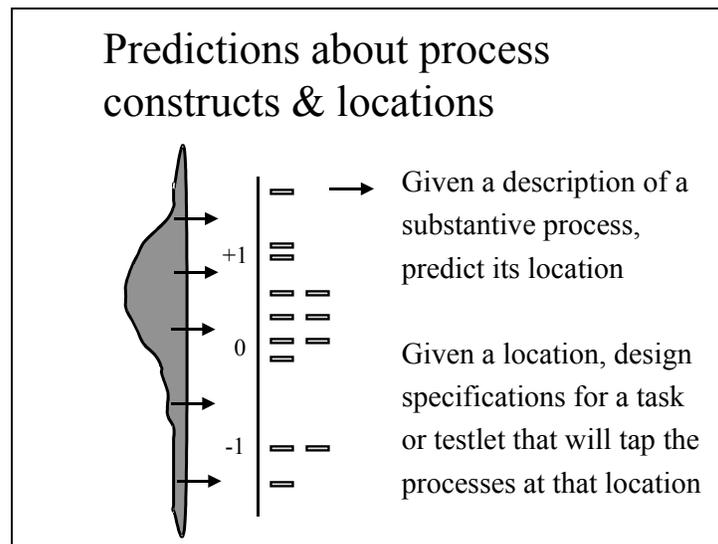


Figure 5. Predictions about the location of tasks and substantive processes along a conjointly calibrated construct-linked scale.

These predictions from a domain theory are all testable. The shape of the person distribution is testable, the location of persons possessed of known substantive processes is testable, the predicted location of items is testable. Predictions of transfer effects to other domain scales, or the effects of instructional treatments on moving persons up a domain scale are all testable.

Examples of Domain Theories

Attention to the issues of linking calibrated scale values to constructs in good theories is practiced by an increasing number of investigators. Domain theory by this name is being used primarily by researchers connected to the Brigham Young University Department of Instructional Psychology and Technology, and more broadly in the future, with the EduMetrics Institute. Among the subject-specific domain theories developed to date, the one with the most extensive investigation is the creation of Diane Strong-Krause on the topic of English as a Second Language (ESL) (Strong-Krause (2000, 2001, 2002)). She described the four stages of construct-linked scale development (CLSD), which are construct delineation, construct-linked ordering, invariant scale development, and construct-linked scaling. In her research, she is developing a domain theory of ESL communicative competence in speaking by using CLSD methodology. In Strong-Krause (2001) she used tree-based regression in an effort to understand the construct-relevant sources of task difficulty in a set of ESL speaking tasks.

Among the cross-domain theories of individual difference domains, the work of Margaret Martinez is the most related to domain theory concepts. She developed the learning orientation construct and theory, and an instrument to measure both overall learning orientation and provide a profile on three subscales (Martinez (1999a, 1999b), and Martinez and Bunderson (2000)).

Strong-Krause's work benefits from extensive efforts in the field of language teaching to define the domain of speaking competence. She is able to build on the American Council of Teachers of Foreign Language (ACTFL) proficiency model and interview protocol, and on excellent theoretical work by several linguistic scientists. By contrast, Cindy Xin (2002) is developing a domain theory, and associated construct-linked scales in a very new domain, wherein there are few if any experts, and little theory. This is the domain of *engaged collaborative discourse in asynchronous computer conferences*. It falls somewhere between a subject matter domain and an individual differences domain. Despite a paucity of either theory or experience, based on her own extensive qualitative investigations, and on a model of increasingly more advanced *moderating functions* developed initially by Andrew Feenberg (Feenberg & Xin, 2002), Xin has made good progress in a very new domain by using the methods of validity-centered design to create a promising theory, a construct-linked measurement scale, and a tool to facilitate both on-line collaborative discourse and data collection. Over time it will be possible to provide additional aspects of validity evidence for the domain theory of engaged collaborative discourse.

As can be seen by the examples given, domain theory can be used to bring order to both well-plowed and new fields.

A DESIGN DISCIPLINE IS NEEDED IN THE DEVELOPMENT OF A DOMAIN THEORY: VALIDITY CENTERED DESIGN OF CONSTRUCT LINKED SCALES

The present state-of-the-art in developing excellent, theory-based measurement instruments is improving rapidly. Embretson & Reise (2000) review several examples of IRT applications dealing with substantive cognitive processes and developmental paths. Bond & Fox (2001) explain how the Rasch model may be used to create developmental scales and to use Wright maps with great interpretive power. Mark Wilson has been teaching a constructive approach to measurement for many years to Berkeley students, using construct modeling and mapping, and a textbook on the subject is expected in the near future. The approach taken here is

compatible with these developments, but starts from two somewhat different perspectives: (1) The idea of design disciplines, even design science [Simon (1969); Reigeluth, Bunderson & Merrill (1978); Brown (1992)]; and (2) the unified validity model promulgated by Messick (1989, 1992, 1995, 1998). The idea of design disciplines, even design theories, was discussed above in section 1. In this section the relation of validity-centered design to unified validity will be emphasized. This four-fold model is depicted in Figure 6.

As background for Messick’s choice of two columns, Test Interpretation and Test Use, consider the following cycle of activity that characterizes the generally sequential process of developing a measurement instrument, then using it through interpretation and use.:

Observe/Compare → Measure → Interpret → Take Action (Use)

There is also a path from observe/compare to Interpret that skips measurement. This path is taken by qualitative researchers. Where is validity to be found in this total process? Messick finds it mainly in Interpretation and Use, which form the two columns of the validity matrix. This accords with his definition of validity (Messick (1995)):

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick (1989,1995)).

He expands upon this definition as follows:

Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores...what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails.

Unified Validity Concept -- 1995		
	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Construct Validity Value Implications	Construct Validity Relevance/Utility Value Implications Social Consequences

Four Facets of Validity (After Messick, 1995)

Figure 6. Messick’s unified validity concept

This concept of validity explains the two columns in Figure 6, and follows the last two stages of the observe→measure→interpretation→action cycle discussed above. One could also

argue that construct validity is inherent in demonstrating a morphism between the real world, considered initially in the observe/compare process, and the metric used in the measure step. According to the representational theory of measurement (Krantz, et al. (1971), Roberts (1979)), measurement – (the use of the real number system or some part of it as a representation for sets and relations in the real world) is possible if there is a homomorphism between the “real” (empirically observable) sets and relations in the world and the mathematical model representing them. We can point out that Messick assigned to construct validity the central role in all four quadrants of his validity model, and that construct validity surely has the meaning that the measures are strongly connected to and represent the sets and relations in the real world. Thus despite the emphasis on interpretation and action, the link between observations and measures has not been lost.

Adding the second row *Consequential Basis* (for test interpretation and use) has proved to be controversial. Critics have argued that negative consequences are solely the fault of misuse, not of sources of invalidity such as construct under-representation or construct-irrelevant variance. Messick himself did not in 1988 see construct validity as falling in the two cells on the second row, but by 1992 construct validity appeared in all 4 cells, and was clearly articulated as the essence of validity. In his last paper on the subject (Messick, 1998) argued that evidence of negative social consequences was a likely indicator of construct-irrelevant variance:

Unanticipated consequences signal that developers may have been incomplete or off-target in test development, and hence, in test interpretation and test use. By 1995 the model had become progressive, all flowing to the bottom right corner, as shown in Figure 6. By 1998 Messick wrote about the constructed nature of validity, using the term “constructing construct validity”. In commenting on the process of constructing construct validity, Messick described construct validity as having six key aspects for which evidence and theoretical rationales should be sought. These six aspects are 1) content, 2) substantive process, 3) structure (including the number of dimensions, and relationship of score structure to each), 4) generalizability (including both reliability and generalizability to different subgroups of people and different task types), 5) external (including both predictive and concurrent evidence), and 6) consequential aspects. These six aspects for constructing (designing) construct validity are keys to validity-centered design. But it is a validity argument that must be constructed as well as a set of scales.

Validity Argument

Cronbach (1988) introduced the term “Validity Argument” after analogy to House’s (1977) notion of “the logic of evaluation argument”. Cronbach was in agreement with the complex but unified nature of validity, and its inseparable connection to values and consequences. Instruments cannot be validated themselves, and interpretations and uses change constantly. As a result, an instrument (and, we add, its associated theories), are never validated once and for all. There is an interplay between evidence and instrument features, and the instrument and theory evolve and are improved to reflect evidence and the correction of identified inadequacies. All we can do is continue to improve the argument for the theory, the construct-linked scales, the instrument, and its delivery system. We do this as Messick stated, through evidence and theoretical rationales.

By 1998 the idea of “constructing construct validity” went hand-in hand with the concept of validity argument. Construct validity is an argument that the empirical evidence and the theoretical rationales favor the interpretations and actions recommended. While the strong

design-as-method and even as-theory position taken in this paper may go beyond where Messick was willing to go, we like to think that he would approve of constructing not only construct validity, but all other aspects of validity as well, and doing so by a principled, well-documented, rigorous design process. Quantitative psychologists have been quite happy with the idea of experimental design. Perhaps more of us can become comfortable with the principled design of both instruments and theory. Why not forthrightly seek to design and revise and experiment until we have evidence for all six aspects of construct validity, as well as evidence for desirable values, positive consequences, utility as ease of use, and all aspects of the unified validity model? Together, this evidence and theoretical rationale provides an increasingly strong validity argument.

Validity-Centered Design

Validity centered design as its current practitioners understand and use it is the beginning of a principled design process for designing and developing improved domain theories, the construct-linked measurement scales associated with them, and documenting the evidence for a validity argument. The validity argument is not accomplished all at one (and indeed, never ends), but is improved step by step as we complete work on each aspect of validity. It also includes planning for future activities to improve other aspects of the validity argument in an ongoing process.

In validity-centered design, we consider nine aspects of validity from the very beginning. These nine aspects may be grouped into three major categories. The last two major categories, (II and III, incorporating aspects 4-9) encompass Messick's six aspects of construct validity. Category I goes beyond these six aspects, but is compatible with the overall unified validity model.

- I. **Design for Usability, Appeal, Positive Expectations.** This aspect of validity has the first and highest priority in the view of those who might finance an instrument designed to be used widely. This category of design is not enough, as the instrument and its theory will fail unless the basic core of inherent construct validity is also considered from the first. However, usability and appeal is generally required before organizations will invest in some new measurement instrument. Activities that lead to this aspect of validity are often found in treatments of *user-centered design*. Common characteristics are the following:

1. **Overall appeal.**
2. **Usability.** The instrument will be easy to use, understandable, quick and efficient.
3. **Perceived value** to the target users, **perceived positive consequences.**

Design for appeal and usability can establish superficial face validity, but in order for users to continue to perceive true value and positive consequences, there must be a strong foundation in inherent construct validity. Without a good blueprint, and continuing improvement (established in category II, below), the instrument might be so off-target that the perceived value cannot be achieved, nor will positive consequences occur. Users quickly notice when real value is not forthcoming.

II. Design for Inherent Construct Validity. Construct validity is the link between reality and the scores or measures produced by an instrument. This aspect of validity starts with the blueprint. Are we measuring important invisible mental processes related to the valued human practices hypothesized to exist in the users? Do the scales we construct through scoring the questions connect with important aspects of reality? There are three aspects to the blueprint:

4. **Content** coverage and appropriateness.
5. **Substantive processes** -- The important but typically invisible mental processes used by those whom we would wish to score as more successful on an instrument, or affective attributes of persons such as their beliefs, attitudes, and values. It is only through theories of the cognitive, linguistic, affective or perhaps psychomotor processes that we can design appropriate questions or performance tasks to get at different degrees of these usually invisible processes.
6. **Structure** of the constructs. The starting number of questions or tasks is expected to collapse into a smaller number of separate unidimensional measurement scales. The scales we design should correspond with an hypothesized, then increasingly validated structure.

III. Design for Reliability and for Evidence of Criterion-Related Validity. This aspect of validity is attained through analyzing the data from using the instrument – along with other measures. Except for reliability and generalizability to different groups of people, scores from the instrument must be correlated with other measures -- other instruments and outcome criteria.

7. **Generalizability.** Evidence that the scoring methods and scores are reliable, and generalize to different genders, racial groups, national groups, etc.
8. **External.** Evidence that the scores predict other valid criteria of what is being measured; also, evidence that other instruments correlate or do not correlate as would be expected by the nature of their constructs.
9. **Consequential.** Evidence that positive results (consequences) do occur over time, and that unexpected but negative consequences do not occur over time. This is an extension of the *perceived* positive consequences listed under category I, above. In this aspect of validity, we obtain *evidence* of the actual occurrence of positive or negative consequences.

Validity-centered design is a work in process. The only other quantitative discipline with design in its title, experimental design, has far more completed work and far wider acceptance. It is broadly interdisciplinary. Validity-centered design is interdisciplinary as well, and reaches out not only to tough-minded experimental logic, but also encompasses ways to design and develop at least two sets of artifacts: learning and instructional materials in domains of interest and measurement systems for measuring progress in the same domains. It is important that good work be accomplished in defining disciplined design approaches to domain theories and construct-linked scales. As Herbert Simon (1969, 1981) has stated about design science:

“If I have made my case...the proper study of mankind is the science of design, not only as a technical discipline but as a core discipline for every liberally educated person.”

However, to him, a science of design is not intellectually soft, intuitive, informal, nor “cookbooky”.

“The professional schools will reassume their professional responsibilities just to the degree that they can discover a science of design, a body of intellectually tough, analytic, partly formalizable, partly empirical, teachable doctrine about the design process.”

OTHER ISSUES OF IMPORTANCE IN DEVELOPING DOMAIN THEORIES

Theory-Based Calibrations are needed to verify the predictions of domain theories

The current dominant philosophy of science apparently favored by most practitioners in educational measurement is logical empiricism. Trout (1998) gives an extensive treatment of logical empiricism and contrasts it to a newer philosophy, *Measured Realism*. Logical empiricism, a descendent of logical positivism, does not privilege theory equally with data. Data rules, and theory to the logical empiricist brings with it shaky, speculative, and metaphysical content.

An argument can be made that to construct a domain theory of learning and growth across all the learners, from novice to the most advanced, requires theory-based calibrations (Bunderson & Newby (2002)). If predictions can be made about the location of tasks and of thinking processes along scales, then a close approximation to actual calibrations can be generated from theory, or even from the experience of teachers who know a domain and its students well. If a theory is good, these theory-based calibrations may correlate highly with later empirical calibrations. These correlations may rival good reliabilities. Theory-based calibrations are also of significant help in testing the predictions of domain theories.

Rigorous Design Experiments are vital to obtaining the data to build a convincing validity argument for theories and their associated construct-linked scales.

In the field of research on human learning and growth, one-time snapshots of learning effects are not enough. They do not test each of the several theories involved. They may be too artificial and contrived actually to allow the key constructs to exert their effects. Design experiments over repeated cycles in live settings offer a powerful alternative solution to these problems.

Other papers given at this conference (Bunderson & Newby (2002) and Strong-Krause (2002)) provide details on both of these points.

CONCLUSIONS

This paper presents a framework for a set of inter-related theories, some prescriptive, some descriptive. When a set of principled design processes contain prescriptive principles for achieving desired ends, and these can be verified empirically, we may call these design theories. Instructional design, measurement instrument design, and implementation design are three examples. Learning theory is a descriptive theory used in guiding the initial creation of prescriptive principles of instructional design. Learning theory as it has been developed and is now taught is not specific to particular subject-matter domains. Thus there is a need for a

domain-specific learning theory. Such a theory could be called a learning theory of progressive attainments in a specific domain of learning. This article uses the term domain theory to refer to a descriptive theory of the contents, substantive processes, and boundaries of a domain of human learning and growth that gives an account of construct-relevant sources of task difficulty; and conjointly, an account of the substantive processes operative at different levels of growth along the scale(s) that span the domain.

Domain theories, and how to build them and their associated construct-linked scales, is the central concern of this paper. The predictions of a domain theory, and how they might be verified, were discussed. Examples of domain theory work were given.

Domain theories are built using a principled design process, and in synergy with the development of construct-linked scales that span the domain. One approach to building domain theories is validity-centered design. This approach relies on principles for developing disciplined design prescriptions from other fields, and applies it specifically to the task of constructing nine aspects of validity into a domain theory and its associated scales. The nine aspects of validity include the six aspects of construct validity proposed by Samuel Messick in his unified validity model. It adds to these other aspects of validity three aspects included in the concept of *user-centered design*. These additional aspects are consistent with the unified validity model, but have been developed more extensively in other fields that specialize in design. Perhaps ironically, putting user-centered design first in the design process rehabilitates *face validity* and gives it much greater importance, but without inherent construct validity, it is argued that the perception of value and positive consequences will be short-lived.

REFERENCES

- Bond, T.G. & Fox, C. M. (2001) *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J.: Erlbaum.
- Brown, A. L. (1992) Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings." *The Journal of the Learning Sciences*, 2, 2, 141-178.
- Bunderson, C. V., Martinez, M., & Wiley, D. (2000). Verification in a design experiment context: Validity argument as design process. *Symposium session at the annual meeting of the American Educational Research Association*, New Orleans, LA.
- Bunderson, C. V. & Newby, V. A. (2002). The role of design experiments and invariant measurement scales in the development of domain theories, in review at *Proceedings of the IOMW*.
- Campbell, D., & Stanley, J. (1966). Experimental and quasi-experimental designs for research. Reprinted from *Handbook of Research on Teaching (1963)*, Houghton Mifflin Company.
- Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun, Test Validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists* (multivariate applications book series), Lawrence Erlbaum Association.
- Feenberg, A., & Xin, M. C. (2002). *A teacher's guide to moderating online discussion forums: From theory to practice*. Available: <http://www.textweaver.org/modmanual4.htm>.

- Gibbons, A. S., Bunderson, C. V., Olsen, J. B., and Rogers, J. (1995). Work models: Still beyond instructional objectives. *Machine-Mediated Learning*, 5(3&4), 221-236.
- House, E.R. (1977) *The logic of evaluation argument*. Los Angeles: Center for the Study of Evaluation.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971) *Foundations of measurement, Volume 1, Additive and Polynomial Representations*, Academic Press, Inc. New York.
- Martinez, M. (1999a). *An investigation into successful learning--Measuring the impact of learning orientation, a primary learner-difference variable, on Learning*. Doctoral dissertation, Brigham Young University, Utah. (University Microfilms No. 992217)
- Martinez, M. (1999b). A mass customization approach to learning. *ASTD™s Technical Training Magazine*, 10(4), 24-26
- Martinez, M. & Bunderson, C. V. (2000), Building Interactive World Wide Web Learning Environments to Match and Support Individual Learning Differences, *J. Interactive Learning Research*, 11, No.3, 163-196
- Messick, S. (1989), Validity, in R. L. Linn, ed., *Educational Measurement* (pages 13-103), New York: Macmillan.
- Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *RR-92-39*, Princeton, NJ: Educational Testing Service,.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741-49.
- Messick, S. (1998). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment [*Research Report 98-4*]. Princeton, New Jersey: Educational Testing Service.
- Reigeluth, C. M., (1983), (ed) *Instructional-design theories and models: An overview of their current status*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Reigeluth, C. M. (1987) (ed) *Instructional theories in action*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Reigeluth, C. M. (1999) (ed) *Instructional-design theories and models: a new paradigm of instructional theory*, Mahwah, N.J. Lawrence Erlbaum Assoc.
- Reigeluth, C.M., Bunderson, C. V., & Merrill, M.D., (1978) Is there a design science of instruction? *J. Instructional Development*, 1(2), 11-16.
- Roberts, Fred S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences*, Addison-Wesley Publishing Company, MA.
- Simon, H. (1969) *The Sciences of the Artificial* (also (1981) 2nd Edition with additional chapters) MIT Press, Cambridge MA.
- Strong-Krause, D. (2000) Developing invariant, construct-valid measurement scales in spoken English as a second language, in Bunderson (2000) (chair), *Foundations of Design Experiments: Science Philosophy, Measurement, and Validation Methodology*; Symposium in Division D (Measurement) *American Educational Research Assoc.*, April 28, 2000.
- Strong-Krause, D. (2001). *English as a second language speaking ability: A study in domain theory development*. Unpublished doctoral dissertation, Brigham Young University, Provo, Utah. Available on the World Wide Web: <http://www.edumetrics.org/research/dissertations/strong-krause.pdf>

- Strong-Krause, D. (2002). Toward a domain theory in English as a second language, in review at *Proceedings of the IOMW*.
- Trout, J.D. (1998) *Measuring the Intentional World, Realism, Naturalism, and Quantitative Methods in the Behavioral Sciences*, Oxford University Press, New York.
- Wiley, D. A. (2000). *Learning object design and sequencing theory*. Unpublished doctoral dissertation. Retrieved January 17, 2001, from the World Wide Web:
<http://davidwiley.com/papers/dissertation/dissertation.pdf>
- Xin, M. C. (2002). *Validity centered design for the domain of engaged collaborative discourse in computer conferencing*, unpublished doctoral dissertation, Provo, Brigham Young University.