

---

# An Overview of Measurement in the Social Sciences

---

George W. Bohrnstedt

---

American Institutes for Research

---

Prepared for the National Academies "Workshop on Advancing Social Science Theory: The Importance of Common Metrics." Washington, DC. February 25-26, 2010.

## **An Overview of Measurement in the Social Sciences**

My objective in this paper is to provide a short history and high level review of measurement in the social sciences. I begin with a short discussion of metrology, the science of measurement, and the lessons that we might learn from it. I then move on to a short discussion of the history of measurement in the physical sciences followed by an overview of measurement approaches in the social sciences. In the final section, I suggest through example that the use of standards, although arbitrary, can in some circumstances serve as useful common metrics.

### **Lessons Learned from Metrology**

Metrology is the scientific study of measurement. It seeks to understand how we have worked throughout history to convert latent scientific constructs into meaningful measurements through rigorous, objective procedures and practices (Sydenham, 1979). There are several interesting examples I can only touch on here but are spelled out in detail in Otis Dudley Duncan's very helpful book *Notes on Social Measurement*. (Duncan, 1984). One example comes from the history of currency and coinage. As far back as the time of the ancient Egyptians, grains were used as currency as payment for labor, that is as a type of currency. (Duncan, 1984, p.56). Silver was used commonly as a means of exchange as early as the 18<sup>th</sup> century B.C. It was carefully weighed in order to establish that fair trades occurred between it and the services it was used to purchase. In order to standardize the use of precious metals coinage was invented, probably around the 6<sup>th</sup> century B.C. (Pareti, Brezzi, and Petech (1965). The various geopolitical units such as city-states had their own coins and there was much variability in how they were valued from unit to unit. As a result, there was a push to develop systems of agreed-upon equivalent units—that is a pressure for standards of measurement equivalence developed. Fair trade required that there be agreement about the value of various coins as measures of value.

Another example is the measurement of length. Body parts were used-- the feet and hands in particular. Feet were used to measure distance, and, of course, still survives as a measure of length. The same is true for hand which is still used to measure the height of horses. Of course, because of the variability in the size of our hands and feet, there was a strong push for standardizing the meaning of feet and hands.

Measurement reform was part of the agenda of the French Revolution with mandate that there be uniformity in weights and measures (Duncan, 1984, pp.20-23). The French Assembly abolished the then current units of weight and length and turned to scientists by forming a commission to come up with a “perfect” system “based on a constant model, found in nature” (Langevin, 1961 p. 86). Based on the commission’s work, the French adopted the decimal system and defined the meter as one ten-millionth of a quarter of the earth’s meridian; the meter was to be the fundamental unit in the newly adopted measurement system (Duncan, 1984, p. 21). The work gained international significance when Talleyrand, then Minister of Foreign Affairs, brought together a delegation of scientists from nine countries to put the system into final form. This had the effect of giving great scientific stature to the work. But as Duncan states: “Unfortunately, no system of human contrivance is ‘perfect.’ Subsequent scientific work has vitiated the presumption that metric units enjoyed a privileged relationship to nature.” (p. 22). He continues: “A recurrent theme in the co-evolution of dimensions, units, and standards is the need for agreement on what are, after all, mere conventions of measurement and the need for enforcement of uniformity in practice.” (p. 16).

In the U.S., Thomas Jefferson, the first Secretary of State, recognized the need for the government to be involved in the standardization of coins, weights and other measures as early as 1790. His proposal for such standardization was endorsed by Washington, Hamilton, and Madison, but in spite of its merits, Franklin’s proposal was not adopted. Indeed, it took until 1901 for the U.S. Congress to establish a National Bureau of Standards although there were two precursor organizations established, the first of which was the Office of Standard Weights and Measures in the Department of Treasury in 1824.<sup>1</sup>

As a further and final example of the adoption of standards as a social process, we note the failure of the U.S. to adopt the metric system universally in spite of the obvious advantages it would have had operating in the world economy as we do today. This example underscores the importance of the societal context within which science operates.

What are we to take-away from this very brief set of historical examples? I think there are three things: (1) Measures are social constructs and the process of gaining standardization around

---

<sup>1</sup> <http://qanda.encyclopedia.com/question/national-bureau-standards-founded-752701.html>

measures is very much a social process involving negotiations among social actors; (2) Standardization is impelled along when there are strong commercial, political or scientific reasons for doing so; and (3) Science has a strong and central role to play in the development of standards for measurement.

### **Measurement in the Physical Sciences**

Measurement in the physical sciences differs from most measurement in the social sciences in that the physical sciences have developed standards based on theory and experimentation. For example, to measure the length of an object, one uses a meter stick. Another example is the use of specialized display readouts. There is typically some type of sensor that measures the phenomenon of interest and transmits the data to a display unit such as a meter or digital display unit using computer screens. Other examples include scales, ohmmeters, and thermometers. Importantly, all of these types of instruments are calibrated for accuracy against a standard. The maintenance and these standards is the mission of the National Institute of Standards and Technology (NIST) which is part of the U.S. Bureau of Commerce.

Originally founded as the National Bureau of Standards (NBR) in 1901, the NBR was the nation's first federal physical science research laboratory. There are seven base units maintained by NIST called the International System of Units, or SI for short. They are:

<i>Base Quantity</i>	<i>Name</i>	<i>Symbol</i>
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electrical Current	ampere	A
Thermodynamic Temperature	Kelvin	K
Amount of Substance	mole	mole
Luminous Intensity	candela	cd

The definitions of each are very precise. For example, a meter is defined as the length of the path travelled by light in vacuum during a time interval of  $1/299,792,458$  of a second.<sup>2</sup> A candela is defined as the luminous intensity, in a given direction, of a source that emits

---

<sup>2</sup> <http://physics.nist.gov/cuu/Units/current.html>

monochromatic radiation of frequency  $540 \times 10^{12}$  hertz and that has a radiant intensity in that direction of 1/683 watt per steradian<sup>3</sup>

Other quantities can also be derived from the seven base quantities. Some examples of the hundreds of such derived quantities include:

$$\begin{aligned} \text{Area} &= \text{m}^2 \\ \text{Velocity} &= \text{m/s} \\ \text{Acceleration} &= \text{m/s}^2 \\ \text{Luminance} &= \text{cd/m}^2 \end{aligned}$$

It is also possible to use the derived measures to construct other derived measures. Thus, for example,

$$\text{Force} = \text{mass} * \text{acceleration}.$$

While most of the equations in the physical sciences are based on the base units and the derived quantities, Zebrowski (1979) points out that there are also some cases where empirical equations are used, equations derived purely from the measurement of observations. The notion of empirical equations will become important when we discuss the psychophysical work of S.S. Stevens below.

It is important to note that the definitions in the physical sciences are updated from time to time as scientific information is itself updated. The definition of the meter the French Academy of Science adopted in 1791 was changed in 1889 because the prototype that was used was short by 0.2 millimeters as a result of researchers having miscalculated the flattening of the earth due to its rotation.<sup>4</sup> The definition was again changed in 1927, 1960 and 1983 when the current definition was put in place.

A critically important question is whether phenomena under measurement have a reality apart from the measures used to assess them. In the physical sciences the answer would seem to be an affirmative one, at least according to Einstein:

---

<sup>3</sup> Ibid. Also, a steradian is related to the surface area of a sphere in the same way a radian is related to the circumference of a circle

<sup>4</sup> <http://physics.nist.gov/cuu/Units/meter.html>.

“I think that a particle must have a separate reality independent of the measurements. That is an electron has spin, location and so forth even when it is not being measured.”  
(Einstein)<sup>5</sup>

Having stated this, one surely cannot see or touch weight or length although we can sense both. Importantly, both measures of these constructs can be calibrated against tangible standards as stated above. Certainly some of the phenomena in which social scientists are interested also have a reality apart from their measurement. Examples include date of birth, age, marital status, number of children, date of death and so forth. Many of the variables demographers use are of this type and often are expressed in rates. But the picture becomes much murkier when one thinks of cognitive constructs such as intelligence, religiosity, or impulsivity. Any of what we call attitude, value, ability or personality constructs fall within this category of murkiness. So do organizational constructs such as school climate, organizational learning as well as societal level constructs such as anomie, social disorganization and so on. Yet these types of unobservable or latent variables are the workhorses of much of social science. We can only observe their “reality” by examining the covariation between observed indicators presumed to be causally related to the latent variables or by looking at their hypothesized effects on measures of other constructs. We return to this issue below, but want to ensure that the reader understand this fundamental difference between physical measurement and much of social measurement with respect to the nature of their latent constructs. In the social sciences, most of our constructs are similar to what Northrop (1947) called “concepts of intuition” – concepts we perceive as opposed to concepts which are derived from axiomatic, deductive theory in the physical sciences which he calls “concepts of postulation” (Northrop, 1947, pp.82-83).

As Duncan (1984, pp. 159-160) points out, one of the most remarkable characteristics of equations expressing physical quantities is their simplicity—they involve products of quantities that are raised to some integral power. We simply don’t have anything quite like this simplicity in the social sciences perhaps with the exception of economics and psychophysics in psychology. Duncan in describing this issue points to the econometrics textbook by Carl Christ (1966) where he lists over 20 economic concepts and their formulas based on a relative small number of dimensions. For example, the price of a good is defined as Money/Quantity. But as Duncan goes on to say, it is not clear whether or how such dimensional equations have helped improve

---

<sup>5</sup> [http://thinkexist.com/quotes/albert\\_einstein](http://thinkexist.com/quotes/albert_einstein)

theory-building in economics. I will leave this for the economists to decide. I will discuss the case of psychophysics, using S.S. Steven's work as the prime example, in the next section where I address measurement in the social sciences more generally.

Before turning to that discussion, however, there is one other thing that is worth noting about measurement in the physical sciences and that is the interplay between theory and measurement.

We have perhaps underemphasized the role of observation and measurement in the development and refinement of theories in the physical sciences in our discussion above. The observation and recording of data are hugely important in the history of science. One only need read how Galileo came to describe acceleration by rolling balls down incline to understand the important interplay between data and theory in the physical sciences. And while measurement in some areas of the physical sciences is very accurate, in other areas, there are huge errors in measurement. Duncan (1984) quotes Hunter (1980) who states: "...the quality of many scientific measurements is suspect. (p. 874). Duncan goes on to note that William Kruskal in an informal communication notes that the discrepancies in some astronomical measurements are larger than would be tolerated even in some kinds of social measurement. (p.165). One only has to look at the history of astronomy to see estimates of the age of the earth that within the past 50 years have ranged from 10 to 19 billion years, that is, they vary by a factor of nearly two. By the way, the estimate as of 2006 was 13.7 billion years (Primack and Abrams, 2006). The point is this: There is measurement error in the physical sciences as well as in the social sciences.

But there is one important difference that needs to be noted; it is the availability of strong theories against which to test one's observations. In the physical sciences where we have good theory, measurements can often be used to confirm, reject or refine theories. Thomas Kuhn (1961) argues that in good scientific practice, one uses measurements to compare two theories with each other (Duncan, 1984, p. 169). Needless to say, examples of where measurements are used to compare competing theories in the social sciences are quite rare.

Kuhn also felt strongly that measurement needed to follow theory and not the other way around. He argued: "...the route from theory or law to measurement can almost never be travelled backward. Numbers gathered without some knowledge of the regularity to be expected almost

never speak for themselves” (Kuhn, 1961, p.45). Einstein would seem to agree when he said: “It is the theory that decides what can be observed.”<sup>6</sup> In economics, Koopmans (1947) argues that measurement without theory is not nearly as productive in the generation of useful knowledge as measurement guided by theory. Not everyone agrees with this position, however. Borgatta (1961), for example, argues that one should be more inclusive rather than less inclusive by measuring many variables that may be related to the phenomenon of interest rather than just those identified as plausible because of a single theory; that is, use a “shotgun” instead of a “saltshaker.” Bradburn and Cartwright (2010) address the relationship between research and theory in their paper in this workshop as well.

What do we take away from this history? (1) We have not discovered or figured out how to define the kind of fundamental quantities in the social sciences that exist in the physical sciences; (2) Our concepts are large in number and for the most part do not bear the kind of simple relationships to one another as is true in the physical sciences, (3) We lack strong mathematical theories against which to evaluate our measurements and vice-versa. Whether these are *sine quo non* for the development of better measures within the social sciences is difficult to say. But all three of these points would seem to be important to consider when seeking improving measurement in the social sciences.

## **Measurement in the Social Sciences**

We now briefly discuss the historical development of measurement in sociology and psychology and then delve more deeply into some of the areas that are of particular importance for measurement in the social sciences today<sup>7</sup>

---

<sup>6</sup> [http://thinkexist.com/quotes/albert\\_einstein/11.html](http://thinkexist.com/quotes/albert_einstein/11.html).

<sup>7</sup> In economics, since the main variable of primary interest is capital, there does not seem to be a lot of concern about measurement per se. But there have been concerns about the effects of measurement error (called “errors in variables” by economists). The earliest reference I could find in the literature on this topic is an article Durbin (1954). The topic is also covered in detail by Carl Christ in his econometric textbook which was first published in, where the matter is handled through an instrumental variable approach.<sup>7</sup> (Christ, 1966). It was Arthur Goldberger, however, who in a series of articles showed how unbiased estimates could be obtained in econometric models in the presence of errors in variables using structural equation modeling with unobservable variables (Goldberger, 1971; 1972a; 1972b).

## Sociology

In sociology, Le Play is often considered the father of the modern social survey. He had workers live with families in order to gather data on attitudes and beliefs, family budgets and family expenditures as ways to determine families' standards of living. From these data, sets of quantified cases studies were developed which became the data for LePlay's writings. Bowley is credited with the introducing the notion of probability sampling into survey research methodology in the 20<sup>th</sup> century (Duncan, 1984, 107). Needless to say, the survey has and continues to be the framework within which much of the measurement within the discipline of sociology (as well as related social sciences) occurs right up to the present time.

Most of the measurement done by sociologists and other survey researchers has been what Torgerson (1958) calls *subject-centered measurement* where one places the respondent on a continuum that runs from low to high. This is typically done through the use of Likert scales with response categories such as "Definitely disagree," "disagree," "neutral," "agree," and "definitely agree" (Likert, 1932). This is in contrast this with *stimulus centered measurement* where one orders stimuli from high to low on a scale which characterizes measurement in what is called psychophysics which is discussed below.

There are other rating scales that order subjects as well. One member of this class is called "Behaviorally anchored rating scales (BARS)" (Smith and Kendall, 1963). These typically have five to nine categories where each category has a verbal description of the behavior to be rated. BARS are very commonly used in the ratings of performance, especially in personnel psychology for the rating of job performance.

Another type of rating scale is anchored only at the extremes. Perhaps the best example of the use of these types of scales is the semantic differential developed by Osgood, Suci and Tannenbaum (1957) on their innovative work on the measurement of meaning. In this work, subjects are asked to rate objects, persons or phenomenon on a set of scales using adjectives such as "hot—cold" "strong—weak" where an adjective and its antonym are the anchors. Using

factor analyses to analyze the data, Osgood and colleagues extracted three dimensions that could be used to measure meaning – evaluation, potency and activity.<sup>8</sup>

Yet another form of subject-centered measurement that enjoyed some prominence in the mid-20<sup>th</sup> century was the Guttman scale (Guttman, 1950). To fit a Guttman scale the following criterion must be met: a positive response to the  $j$ th item implies a positive response to items  $j-1$ ,  $j-2$ .... $1$ . That is, if one agrees with the fourth item in a group of  $k$  items, one must also agree with the third, second and first items. This also implies that agreement with the third item means that one also agrees with the second and first items, and so on. A good example of a Guttman scale is the Bogardus social distance scale where respondents were asked to respond with a “yes” or “no” answer to each of the following questions:

1. Are you willing to permit immigrants to live in your country?
2. Are you willing to permit immigrants to live in your community?
3. Are you willing to permit immigrants to live in your neighborhood?
4. Are you willing to permit immigrants to live next door to you?
5. Would you permit your child to marry an immigrant?

If one answers “yes” to question 4, the implication is that he or she will have answered “yes” to the preceding three questions as well.

One of the problems with Guttman scaling is that it is a deterministic rather than probabilistic scaling model. Guttman suggested a coefficient of reproducibility as a way to judge the fit of data to the model. An entire cottage industry developed around how to do “error analyses” for item patterns that did not fit a Guttman scale. But we shall spare the reader reviewing this literature since it is now really only of historical interest. This is not to diminish the importance of Guttman scaling; indeed, it is the precursor of what are the strongest methods we have for measurement today – Item Response Theory (IRT) methods (Andrich, 1985). IRT models, like Guttman scaling *orders* items and persons on a scale from low to high. By contrast, many if not most of the scales used in current work based on factor analytic approaches are composites of

---

<sup>8</sup> As a relevant aside, the paradigm for measuring meaning developed by Osgood, Suci, and Tannenbaum may be a good example of where a common metric has been quite well established.

items that may correlate well together, but for which we are not at all certain where they operate on the underlying continuum of interest.

## **Psychology**

There have been two separate strands of work in psychology that have had important implications for the development of social and psychological measurement. The first is psychophysics and the second is the measurement of intelligence. We now turn to a discussion of each.

***Psychophysics.*** In psychology, the earliest work on measurement is generally associated with Fechner's research on psychophysics (Boring, 1961) which focused on human sensation and perception. Although Fechner did some work on esthetic judgments as well, he did not seem interested in applying his methods to social measurements more generally. That is, he focused mostly on the measurement of sensation and perception. It was L. L. Thurstone, however, who first took Fechner's work and applied it to social phenomena, including which nationalities students would prefer to associate with, which criminal offenses were perceived as the most serious, attitudes towards the church, and so on (Thurstone, 1959). The other important development for our purposes that grew out of work in psychophysics is the work of S.S. Stevens about whom I will have more to say below given his influence on what are meant by levels of measurement.

*L.L. Thurstone.* While psychophysics was aimed primarily at trying to scale sensations and perceptions, Thurstone used its methods to scale attitudes and values and is credited as the first psychologist with the measurement of each. (Thurstone, 1927a; 1927b; 1928) His approach to the scaling of values had subjects consider pairs of stimuli simultaneously in making judgments—known the *method of paired comparisons*. For example, one of his papers focused on measuring the seriousness of various crimes. He approached this by developing a list of 19 crimes. Each crime was then paired with every other crime resulting in  $19 \times 18 / 2 = 171$  pairs. Subjects were then asked for each pair, which of the crimes was the more serious. In some pairs one of the crimes is clearly seen as more serious than the other; in other cases both members of the pair are seen as very close to each other in seriousness. That is, just as in psychophysics

there are some cases of where sensations are discriminated as barely different from each other, and other cases of where they are sensed as very different from each other. The sizes of these differences are then used as the basis for scaling the stimuli.

Thurstone's approach to the measurement of attitudes was somewhat different than the measurement of values but he demonstrated that this approach was also derivable from the laws of psychophysics (Thurstone, 1928). He defined an attitude as a latent, unobservable variable that lay on a continuum from very favorable to very unfavorable. Observables that could be used to measure attitudes included behaviors and opinions. He focused on the use of opinions. Unlike most contemporary attitude measurement work, Thurstone separated the scaling of items from the scaling of persons. He began by having one group of subjects generate a large number of statements with instructions to cover a diversity of opinions about some object, law or phenomenon. Examples included attitudes towards prohibition, militarism and the church. After paring the list to something under 100 statements, he asked 200-300 subjects to sort the statements into one of 11 piles where the middle category reflects a neutral attitude and the two end points measure "strongly affirmative" and "strongly negative," respectively. He suggested that the 11 categories be thought of by the judges as roughly equally-spaced on the continuum. He then examined each pair of statements in terms of the percent that placed it in a given category. It was by examining the scale separation between pairs of items that allowed him to use his *comparative law of judgment*. In Thurstone's words:

"If 90 per cent of the judges or readers say that statement  $a$  is more militaristic than statement  $b$  ( $p_{a>b}=.90$ ) and if only 60 per cent of the readers say that statement  $a$  is more militaristic than statement  $c$  ( $p_{a>c}=.60$ ) then clearly the scale separation ( $a-c$ ) is shorter than the scale separation ( $a-b$ ). The psychological scale separation between any two stimuli can be measured in terms of a law of comparative judgment..." (Thurstone, 1928, p.541).

The degree of separation between pairs of statements was then used create scale scores for each statement. To place persons on the scale, he would choose say 25-30 of the opinion items that were relatively evenly spread over the attitude continuum and then these would be administered to separate samples. One's score on the attitude scale is simply the number of items agreed to from the list of 25 or 30 opinion items.

There are a couple of other things to note about Thurstone's work on attitude measurement. First, he understood that an important assumption about his approach was that the opinions one had needed to be independent of where they placed items among the 11 categories. That is, it was assumed that one's own religiosity, for example, was unrelated to where items measuring attitudes towards the church were placed on the attitude continuum. This notion of *invariance* is an important one for measurement, that is, that the items have the same meaning for the various subpopulations of respondents to whom they are administered. It also means that if additional items are added to the scale, that the ordering and the distance between any two respondents,  $i$  and  $j$  on the latent dimension is also invariant. Second, Thurstone plotted cumulative probability distributions for each item as way to see where they were operating along the attitude continuum. In doing so, he noted that the slope of these curves gave useful information about how well each of the items discriminated in the measurement of the attitude. These curves look very much like the item characteristic curves associated with current IRT approaches to measurement which we address later in this paper. That is, Thurstone's work on attitude measurement was an important precursor to modern measurement approaches.

### S.S. Stevens

S., S., Stevens was far more interested with phenomena associated with classical psychophysics than was Thurstone, notably the scaling of sensations. His approach required subjects to make judgments using ratios of stimuli. For example, he would present Tone A and tell the subject that its loudness is a "10." They were then told to rate the loudness of each succeeding tone by comparing it to the first tone in terms of ratios, e.g., Tone B is three times as loud as Tone A, or Tone C is half as loud as Tone A. By plotting the perceptual data against the actual stimulus magnitudes across different stimuli, he noted that they formed a similar pattern. This led him to postulate a general psychophysical law which was stated as:

$$\psi = \alpha \phi^\beta$$

where  $\psi$  is the perceived magnitude,  $\phi$  is the actual magnitude of the stimulus intensity,  $\alpha$  is a constant that varies depending upon the units of measurement used, and  $\beta$  is the parameter to be

estimated depending upon the stimulus (e.g., loudness of a tone, brightness of a light, etc.). When put in logarithmic form, the law becomes linear in form:

$$\log \psi = \log \alpha + \beta \log \phi$$

where the intercept is  $\log \alpha$  and the slope to be estimated for each stimulus is  $\beta$ . Duncan (1984, p. 177-178) notes that while Stevens calls this a law, it was not one derived from axiomatic theory; instead it is an empirically derived law and therefore does not share the characteristics of the base and derived measures in the physical sciences discussed above. This does not mean that Stevens' work is flawed in any fundamental sense, only that it does not enjoy the same status as fundamental measurement in the physical sciences.

While Stevens did little in the area of application of his psychophysical methods to social phenomena, some sociologists and political scientists have. Examples include the work of Shinn (1969), Hamblin (1971, 1974), Rainwater (1972) and Kaplan, Bush, and Kerry (1979).

***The Measurement of Intelligence.*** Another major strand in psychology relates to the measurement of intelligence where Galton's work is generally credited. His work provided the basic ideas for correlation that led to the work of Spearman in the early 1900s that provided the groundwork for factor analysis with his emphasis on the *g-factor* or general factor in intelligence. Another heavy contributor to work in relating intelligence to factor analysis was L.L. Thurstone. Thurstone challenged Spearman's work by positing a set of general factors called primary mental abilities (Thurstone, 1938) which were analyzed by multiple factor analytic methods developed by Thurstone (1947). This work was important in its own right, but also is important because factor analysis became one of the major workhorses in the development of social measurements. The work of Thurstone and others who developed factor analytic methods prior to the mid-to-late 1960s are called *exploratory factor analysis* (EFA) methods. They are referred to as exploratory methods because they can be used to explore the dimensionality of a set of items designed to measure a construct or set of constructs. But, EFA cannot be used to confirm the dimensionality or structure of a set of items in a rigorous way. However, in the early 1960s Karl Jöreskog (1969) developed *confirmatory factor analysis* (CFA) which does allow one to test rigorously hypotheses about the number of dimensions underlying a

set of items – a technique which enjoys much popularity today as a way to build social and psychological measures. We shall have more to say about both EFA and CFA below.

In the sections that follow immediately below, I spell out the roles of both psychophysics and factor analysis for social measurement in more detail since both have had crucially important influences.

**S.S. Stevens on Scales of Measurement**

Although his work in the field of psychophysics is very well known and has been influential, S.S. Stevens is perhaps best known because of his paper “On the theory of scales of measurement” (1946). In it, he defines four hierarchical scales of measurement based on the kinds of mathematical transformations that are allowable under each. The four types are: nominal (N), ordinal (O), interval (I) and ratio (R). The accompanying table shows the scales, permissible transformations and examples of each (Stevens, 1975).<sup>9</sup>

<i>Scale</i>	<i>Permissible Transformations</i>	<i>Examples</i>
Nominal	Substitution of any number with any other number	Football Jersey numbers
Ordinal	Any change that preserves order	Hardness of minerals
Interval	Multiplication by or addition of a constant	Fahrenheit, centigrade scales
Ratio	Multiplication by a constant	Length, Weight

That there are in fact different levels or types of measurement is a point about which there is no quarrel. The problem is the prescriptive language used by Stevens (1951) about what statistics could be used to describe or analyze the data collected as a function of the level of measurement of the tool used to collect it. For example, when one has what he called nominal level measurement, analyses according to Stevens must be limited to statistics such as the mode, and contingency coefficient. Statistics that are permissible for ordinal scales include the median and rank order correlations--statistics whose meanings are preserved when monotonic

<sup>9</sup> For interesting look at the examples used by Stevens over time see Duncan (1984), p. 124, Table 4-2.

transformations are applied to the data. Interval level scales allow one to compute means, standard deviations and Pearson product moment correlations—statistics whose meanings are unchanged when linear transformations are applied to the data. The availability of ratio level data also allows the computation of geometric means and coefficients of variation, both of which are unchanged by rescaling one's data. Importantly, there is cumulateness to the operations that can be done as one moves up the levels of measurement hierarchy. That is, one can do operations on ordinal data that are permissible for nominal data. In turn any operations permissible for ordinal level data are also permissible for use with interval data, and so on.

There have been several scathing critiques of the Stevens' dicta about what statistics can and cannot be computed because of the level of measurement of one's scales. I especially recommend Paul Velleman and Leland Wilkinson's review article "Nominal, Ordinal, Interval, and Ratio Typologies are Misleading" in *The American Statistician* (Velleman and Wilkinson, 1993).<sup>10</sup> The earliest attack appears to have been made by Fredrick Lord (1953) in which he argues that the choice of an appropriate statistic depends upon the meaningfulness of the statistical analysis taken and not the level of measurement of one's scales. John Tukey (1961) agreed with Lord and further argued about the importance of the meaning of the data in making a decision about what which statistical analyses are appropriate. Furthermore, he argues that just because Stevens' scale types are absolute doesn't mean that the statistical methods must be absolute as well. In a similar vein Baker, Hardyck, and Petrinovich (1966) and Borgatta and Bohrnstedt (1980) argue that much of the statistical power of parametric statistics is lost when one uses rank order statistics on types of measures typically used in the social sciences.

Borgatta and Bohrnstedt (1980) further argue that in most cases the underlying variable of interest is a continuous one where equal intervals are assumed and the fact that one does not measure it at the interval level only means that one is measuring with error. Abelson and Tukey (1959) make a similar argument. Guttman (1977) makes the more general point that the use of statistical techniques hinges on the kinds of questions asked of the data at hand and on the kind of statistical evidence one would accept to inform us about those questions. He argues that one should try to minimize a loss function.

---

<sup>10</sup> Also see <http://www.cs.uic.edu/~wilkinson/Publications/stevens.pdf>.

Duncan (1984) has many criticisms of Stevens' approach to scales of measurement, but perhaps the most devastating one is Stevens' use of the term "nominal scale." He sees that much of what Stevens refers to at the nominal level is not measurement at all but mere labeling in some cases and classification in others. He goes on to make the point that many dichotomous variables (which would be viewed as nominal by Stevens) have an important place in measurement and conditions such as "present versus absent" or "on versus off" are important examples of this. Other examples include the dummy coding of variables such as religious identification (e.g., Catholic or not) and political party identification (e.g., Republican or not). Duncan (1984, p.122-126) also takes strong issue with Stevens' definition of measurement which is "...the assignment of numbers to objects or event according to rules" (Stevens, 1946). Drawing on Cohen and Nagel (1934, p.294) Duncan states: "Measurement is not only the assignment of numerals, etc. It is also the assignment of numerals in such a way as to correspond to *different degrees of a quality* or property of some object or event." Duncan goes on to state: "...the purpose of measurement is to quantify" and notes that for the philosopher Bunge (1973, p.108) "To quantify...is to introduce a functional correspondence between the degrees of a property and a number." That is, the assignment of numbers should represent a functional relationship between the numbers and the *degree* to which some object or phenomenon possesses some quality or characteristic. That is, measurement involves *magnitude*.

Unfortunately, there was a generation of social scientists trained in 1950s who were taught because of Stevens' dictum they could not use parametric statistics, except where one had interval or ratio levels of measurement, which tended to be rare except for economists and demographers. Virtually every statistics text in the social sciences used Stevens' levels of measurement as the guidepost for what statistics could or could not be used for a particular type of measure. Luckily, by the 1970s and 1980s this began to turn around but one still sees reference to levels of measurement in statistics texts in the social sciences. Stevens cast a long, long shadow.

### **Factor Analysis and the use Linear Composites in the Social Sciences**

One of the most common approaches to the development of composite measures is to write a set of Likert or other rating type items that are presumed to measure the concept of interest, and then

to factor analyze them to determine their dimensionality. Typically the question is whether one factor is sufficient to explain the covariation among the items. Or if the concept is multi-dimensional, the investigator asks how many factors it takes to explain the covariation among the items. The investigator then builds one or more linear composites of the items and examines their internal consistency reliability typically using Cronbach's  $\alpha$  (Cronbach, 1951). One of the problems with this approach is that it was largely atheoretical. Scales are often built without regard for other measures of the same or similar content that already exist. Many of them borrow items from each other. For example, Schuessler (1982) did an analysis of the domain of what he characterized as "social life feelings." He examined concepts such as anomie, alienation, external-internal control, life satisfaction, cynicism, anomie, normlessness, etc. He found roughly 1000 items contained in the various composite scales and reduced them by half when eliminating duplicates or near duplicates. He then pilot tested the remaining items and based on a pilot test, developed a questionnaire that eventually had 237 items. They were then administered to a national sample of 1500 adults and factor analyzed them. He ended up creating a dozen new scales, all with acceptable reliabilities, and created norms for each based on the fact that it was a national sample.

While appreciating Schuessler's work was a useful exercise in showing the overlap and duplication as well as the separation of a number of similar sounding concepts, Duncan (1984) characterizes this approach to measurement as no more than "...a "correlational" science of "inexact constructs" (p.207). In his view this approach to the development of measures is a dead end and will never result in the kinds of fundamental measures that characterize the physical sciences. We end up with a plethora of poorly defined, fuzzy, and poorly measured concepts. These are concepts that Bradburn and Cartwright in their paper refer to as 'ballungen' which translate roughly in English to "agglomerations" or "congestions," concepts that are fuzzy on the edges.

### **Classical True Score Theory**

Although rarely made explicit by those building composite measures, the underlying justification for the factor analytic approach in building measures typically flows from classical

test score theory (or “true score theory” as it is sometimes referred to).<sup>11</sup> An observed score,  $x$ , is assumed to be related to or to reflect an underlying true score,  $\tau$ , as follows:

$$x = \tau + \varepsilon$$

where  $\varepsilon$  is random measurement error. For a set of  $x$ s the underlying  $\tau$  can be thought of as an underlying factor assuming that all of the variance is common variance, that is, assuming that the item has no other reliable variance specific to itself or held in common with other factors.

Unfortunately, in the 1950s and 1960s when many of the types of constructs Schuessler factor analyzed were constructed, there was no way to really test these types of assumptions given that only exploratory factor analytic methods existed at that time. Among other things, exploratory factoring methods have an infinite number of solutions leading to the so-called “rotation problem.” Another issue was that most factor methods at the time failed to make the distinction between sample and population statistics as it applied to factor analysis. Hence, rigorous tests of model fits were not possible.

### **Confirmatory Factor Analytic Methods**

As a result of these limitations, it was with open arms that the *confirmatory* factor analytic methods developed by Jöreskog (1969) as a subset of his more general linear structural equation models with latent variables was welcomed. Jöreskog further enhanced the importance of this work when he wed confirmatory factor analytic methods to congeneric measurement theory. A set of measures  $x_1, x_2, \dots, x_n$  is defined as congeneric if their true scores,  $\tau_1, \tau_2, \dots, \tau_n$ , are correlated at unity with each other, implying that all the  $\tau_i$  are linearly related to a random variable  $\tau$ , that is, to an underlying true score. As a result, congeneric measurement implies that

$$\tau_i = \mu_i + \beta_i \tau$$

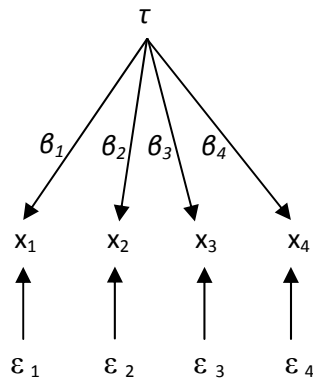
and since  $x_i = \tau_i + \varepsilon_i$ , it follows that  $x_i = \mu_i + \beta_i \tau + \varepsilon_i$ . That is, an item is a linear function of a weighted true score plus error plus a location parameter captured by  $\mu_i$ . Assuming that the  $x_i$  follow a multivariate normal distribution, the parameters of the model can be efficiently

---

<sup>11</sup> We will a type of index construction below that does not depend upon either score theory or factor analysis.

estimated by Jöreskog's (1969) maximum likelihood method. The method yields large-sample standard errors for all parameter estimates as well as the overall chi-square goodness-of-fit statistic, which allows one to *test* the assumption that measures are congeneric and whether one's items could be accounted for by a single factor. The method was fully generalizable to the multiple factor case as well. And most important, the structural equation modeling procedures developed by Jöreskog allowed one to posit causal relations between latent variables as well as between a set of observed indicator variables and the latent variables.<sup>12</sup> A path-analytic representation of a four-variable confirmatory factor model where the items are presumed to be explained by a single underlying true score,  $\tau$ , is displayed in Figure 1.

**Figure 1. A Schematic of a Confirmatory Factor Analysis with Four Congeneric Items**



While Jöreskog's work was of monumental importance in moving the classical test score theory method forward, most of the measurement models being used must still be regarded as primitive. One of the major problems with the typical composite measure is that unlike Guttman scaling, there is no explicit concern about where the items are measuring on the latent variable of interest. Add the atheoretical way in which many subject-centered measures are conceptualized and developed, one can see why Duncan called the approach “correlational science with inexact constructs.”

A different type of criticism of this is approach to putting together composite measures is that it privileges reliability over validity. Heise and Bohrnstedt (1970) show, for example, through a

<sup>12</sup> See Bohrnstedt (2010) for a review of the use of classical test score methods including confirmatory factor analytic methods to measurement in survey research.

path analytic approach applied to factor analysis that such measures may contain both valid and invalid sources of reliable variance but that most researchers blithely assume that because they have a “respectable”  $\alpha$ , they have a good measure. In fact, it is very difficult to construct measures with more than 8-10 items that in fact are “pure” in factorial content. As a result, many measures are reliable by current conventions but contain some amount of invalid (albeit reliable) variance.

### **IRT Models and Scale Construction in the Social Sciences<sup>13</sup>**

Item Response Theory (IRT) was devised to measure ability and achievement but has been used less often to model responses for the measurement of social phenomena. This has begun to change however, because of the work of Reiser (1980), Thissen and Steinberg (e.g., Thissen and Steinberg, 1984; Thissen, Steinberg, Pyszczynski, and Greenberg, 1988; Steinberg and Thissen, 1995; Thissen and Steinberg, 2009), Embretson and Reise (2000) and others. Classical test score theory takes a person’s score on an underlying latent variable of interest (e.g., an attitude or belief) to be the sum of responses to a set of items hypothesized to measure the latent variable. IRT theory takes an approach which is almost the opposite. It asks, given the overall distribution of responses to a set of items and a given person’s responses to them, what is the best estimate of the person’s underlying true score? <sup>14</sup>

#### **The One Parameter Logistic Model (1PL)**

The simplest IRT model is used for items with dichotomous responses scored 1 or 0. The basic model assumes that a person  $p$ ’s probability of scoring 1 on the  $i^{\text{th}}$  item is

$$\Pr[x_i(p) = 1 | \theta_p, \beta_i] = \frac{e^{(\theta_p - \beta_i)}}{1 + e^{(\theta_p - \beta_i)}}$$

where  $\theta_p$  represents person  $p$ ’s true score (or “ability”) and  $\beta_i$  is the “difficulty” of item  $i$ . The greater an item’s difficulty, the smaller the percentage of respondents who agree with it, that is,

---

<sup>13</sup> The sections on IRT theory draw heavily on Bohrnstedt (2010).

<sup>14</sup> For an excellent brief introduction to IRT theory see Thissen and Steinberg (2009).

the smaller the number of responses coded 1 rather than 0.<sup>15</sup> The equation above describes the *one-parameter logistic (1PL)* model since it estimates only one parameter ( $\beta_i$ ) for each item. The model assumes that a single underlying continuous latent variable accounts for the covariation among the items and (similar to a single factor in confirmatory factor analysis). Furthermore, the model assumes that items are *locally independent*. Local independence means that at any given value of the latent variable, the *xs* are statistically independent (Hambleton, Swaminathan and Rogers (1991, p. 10). Another way to view it is that if there is local independence, a latent variable,  $\theta$ , will account for the covariation among a set of items.

The single parameter in the 1PL model is the difficulty of the items, that is, the probability of correctly answering (or in the case of social science applications, agreeing with) each item. As a result of having a single parameter, the result is that the *item characteristic curves (ICCs)* for all the items are parallel.<sup>16</sup> An ICC depicts the modeled probability of correctly answering (or agreeing with) an item as a function of the respondent's true score  $\theta_p$ . The 1PL model typically assumes that this takes a logistic form, as opposed to the linear relationships between true scores and observed variables assumed in the classical true score and the congeneric model described above. (Hambleton et al. 1991, p.13). The implication for a 1PL model is that all of the items discriminate along the latent variable equally well.

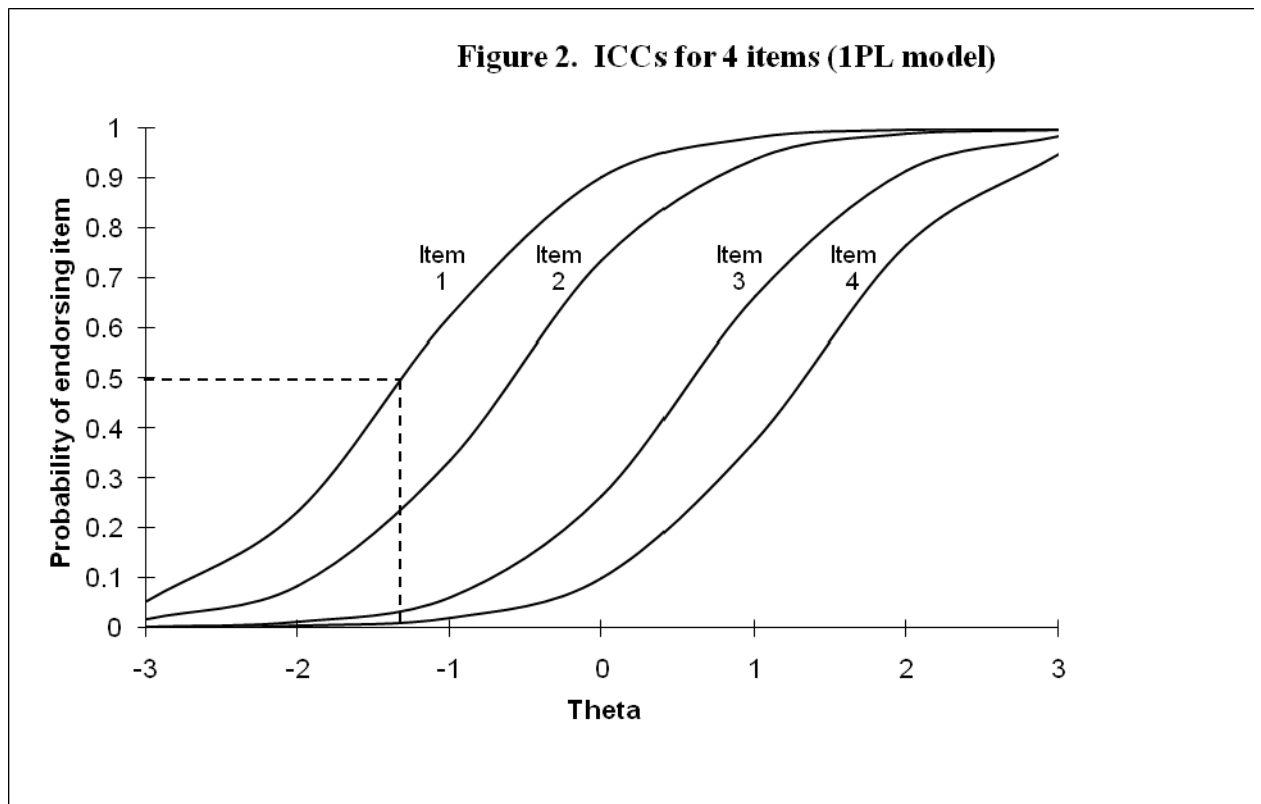
ICCs for a hypothetical 4-item 1PL model are shown in Figure 2; they assume that the latent variable has mean 0 and standard deviation 1. The values for the 4  $\beta_i$  are -1.3, -0.6, 0.6, and 1.3 for items 1, 2, 3, and 4 respectively. Note first that each ICC traces the probability of endorsing a given item as a function of the true score. As the latent variable  $\theta$  increases, the probability of endorsing each item goes up. Second, note that the ICCs have the same *shape* for each item—they only differ in their location along the underlying true score scale. At a true theta of 0, for example, the probability of agreeing with Item 1 is higher than that of agreeing with Item 2; the probability that Item 2 equals 1 is in turn higher than that for Item 3, and so on. This implies that Item 2 is more difficult than Item 1, that Item 3 is more difficult than both Items 1 and 2, and so on. Now consider someone with a true score ( $\theta$ ) of 2. Again the

---

<sup>15</sup> The terms “ability” and “item difficulty” derive from educational measurement, where items are scored as either right or wrong.

<sup>16</sup> Another way to think about this is that the discrimination parameters to be discussed when describing the 2PL model below are constrained to be equal for all the items.

probability of agreeing goes down as one moves from Item 1 to Item 4, just as for the person with a true score of 0; but a respondent with a true score of 2 is much more apt to endorse any given item than is a person with true score 0. That is, item 4 is more difficult to agree with than item 1 because item 4 represents a more extreme position on the construct being measured. Another way to look at it is that respondents with higher true scores are more likely to endorse



any of the items. Likewise, those with lower true scores are *less* likely to endorse any item. Finally, note that if one draws a line parallel to the  $x$  axis where the  $y$ -axis is 0.5 until it intersects the ICC for Item 1, and then drops a line from there through the  $x$ -axis that it equals the difficulty of the item. In Figure 1, this is shown for Item 1 to be -1.3. Another way of thinking of the difficulty of an item is that it is the point where the probability of agreeing with, or getting the item right, is .50.

### The Rasch Model

A particular version of a one-parameter IRT model is known as the *Rasch model*, after the Danish psychometrician who developed it (Rasch, 1960). The model assumes that the

probability of getting an item correct (or agreeing with it) can be explained completely by the item and person parameters and that this cumulative probability distribution for a give item is described by a logistic function. This is exactly what the equation for the 1PL looks like except that as stated, for most 1PL analyzes the data are centered at zero instead of being in the native metric of the responses. If the item fits a Rasch model, the total score for a person is a sufficient statistic to describe a person's position on the latent variable (as well as for describing where the item is operating on the latent variable).

Rasch models have some very appealing measurement characteristics, especially the ability to create interval level measures. Perline, Wright and Wainer (1979, p. 253) note that when the Rasch model holds, measures are on an interval scale with the implications that the distance between, e.g., a 2 and a 3 on a Rasch scale is the same as that between a 3 and a 4. This *equal interval property* does *not* hold for scales that sum either dichotomous or Likert-type ordinal items using either a classical true score or an IRT model.<sup>17</sup> Whereas one is often counseled when building items base on true score theory to use items that have difficulties close to .5, by contrast, with IRT scaling one wants items with difficulty levels that span the full range of the true score scale. Wright (1977, p. 102) notes that the Rasch model is the only latent trait model for dichotomous items for which adding the responses to form a total score is justifiable.

The difference between 1PL and Rasch models has to do with the invariance criterion described in the context of Thurstone scaling. Here is how Rasch himself described what he meant by invariance:

- (1) The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
- (2) Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion (Rasch, 1961, p. 332).

---

<sup>17</sup> Likert scales use ordinal response categories running from, say, "Definitely disagree," to "Probably disagree," to "Probably agree," to "Definitely agree." Distances between these categories are typically unknown, but as we see below, it is possible to estimate them.

The first point implies that for any pair of items, their relationship to one another does not depend upon which set of individuals responded to them. It also means that if other items are added to the scale, the comparison between any other pair of items will remain invariant. The second point means that the comparison between any two individuals on the scale should be invariant with respect to which items they are compared on, and in addition, independent of whom else might have or will take the test at some future point in time. Obviously, invariance is a strong assumption, *but* this is how measurement in the physical sciences works. Any two objects that are measured by two different scales designed to measure weight should remain in the same relationship to each other regardless of what the weighing instrument is. Furthermore, the scales should work just as well in Russia or Brazil as they work in the U.S. Or if we use different scales in different places or different times, they should result in the same differences between the two objects being measured. When we have invariance, we are able to say that we have truly measured something and done so with confidence. It allows for comparisons we can believe in.

When one has items that fit a Rasch model, the resulting measure has the invariance quality just described. One can substitute items and if they fit the Rasch model, the distance between any two individuals on the scale is unaffected. Similarly, the items all measure the same regardless of who is taking them and where they are taking them. This characteristic is obviously very attractive, but it does require some very restrictive assumptions. It is difficult to fit more than a relatively small number of items to a Rasch model. This fact notwithstanding, Duncan (1984) challenges social scientists to take seriously the attempt to build scales using Rasch scaling because of its attractive features. He states: “In my view, what we need are not so much a repertoire of more flexible models for describing extant tests and scales (as interesting as such models might be) but scales built to have the measurement properties we must demand if we take ‘measurement’ seriously” (p. 217).

### **The Two-Parameter IRT Model (2PL)**

Two-parameter IRT models are attractive because of the limitation of equal slopes which applies to all one-parameter IRT models, not just Rasch models. The problem is that with the two-parameter models one loses all the virtues of true measurement associated with the Rasch scale.

Nevertheless, the two-parameter model has become a workhorse in psychometric work because it is nearly impossible to fit more than a few items to a one-parameter model.

The two-parameter logistic (2PL) model assumes that the probability of answering item  $i$  “correctly” (1) is:

$$\Pr[x_i(p) = 1 | \theta_p, \beta_i, \alpha_i] = \frac{e^{\alpha_i(\theta_p - \beta_i)}}{1 + e^{\alpha_i(\theta_p - \beta_i)}} \quad (11.47)$$

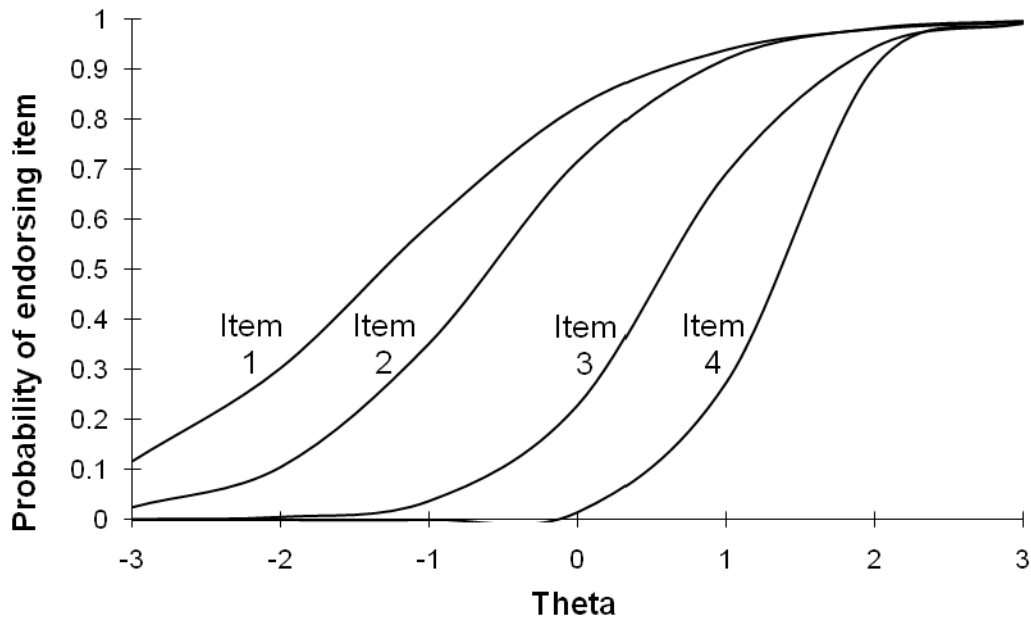
Like the 1PL model, it assumes that item responses at a given true score are locally independent. For each item, the 2PL IRT model estimates an item difficulty parameter  $\beta_i$  (as in the 1PL model) and an additional “item discrimination” parameter,  $\alpha_i$ . The item discrimination parameter permits item responses to be differentially related to the underlying true score  $\theta$ ; items with higher  $\alpha_i$  values make sharper distinctions between respondents whose true scores lie above and below the item’s difficulty level  $\beta_i$ . The difference between the 2PL and 1PL IRT models resembles that between a single-factor model for congeneric items such as, in which factor loadings can vary freely, and a model where the loadings are constrained to be equal.<sup>18</sup> The discrimination parameters add flexibility that allows the 2PL model to fit data substantially better than the 1PL does.

Figure 3 displays ICCs showing the relationship between the true score  $\theta$  and the probability of responding positively to four dichotomous items for a 2PL IRT model. The ICCs in Figure 3 look very different than those in Figure 3, however, because the items have different discrimination ( $\alpha_i$ ) and difficulty ( $\beta_i$ ) parameters. The larger the  $\alpha_i$ , the steeper the slope of the corresponding ICC. In Figure 3, the items were purposely constructed with characteristics such that Item 1 is the “easiest” item and has the worst discrimination whereas Item 4 is the most difficult and has the highest discrimination power. The other two items fall in between on the values of their difficulties and discrimination parameters.

---

<sup>18</sup> Indeed, software packages for structural equation modeling offer options that perform confirmatory factor analyses using dichotomous items. When such models properly specify a logistic relation between the items and the factor(s), one is in fact doing an IRT analysis.

Figure 3: Hypothetical Example of a Four-Item 2PL Model



### The Graded Response Model

The IRT models considered thus far accommodate only dichotomous items. IRT models exist for both nominal and ordinal items with multiple responses, but the items of most interest in the social and behavioral sciences are assumed to have continuous underlying true scores. Hence, here we focus on models for items with ordered response categories such as “Definitely disagree,” “Disagree,” “Agree,” and “Definitely agree.” These responses typically are scored from 0 to 3, or 1 to 4. The technical details of these models are substantially more complex than those of models assuming dichotomous responses.<sup>19</sup> A Graded Response Model (GRM) (Samejima, 1996a; 1996b) estimates complex IRT models for scales including ordinal items. It allows different items in a scale to have different numbers of response categories. As important as the graded response model could become for measurement in the social sciences, the details are simply too complex to get into in what is intended to be a “high level overview” of measurement

<sup>19</sup> See Chapter 5, “Polytomous IRT Models,” pp. 95-124 in Embretson and Reise (2000).

models in the social sciences. The interested reader is directed to Yen and Fitzpatrick (2006) or Thissen and Steinberg (2009).

Much more could and should be said about IRT models than this brief introduction allows. For example, building parallel forms of a test to measure a given construct is very difficult. Item construction is never easy, but IRT theory provides better methods for determining whether the items in Version B are similar to those in Version A than are available using true-score models. Item replacement is easier with IRT theory since one knows in some detail what characteristics a replacement item should have. In spite of its brevity, this introduction has hopefully illustrated the value of IRT models for item development. It offers far more information about how one's items and their response categories operate along the latent true score scale than what one obtains using more traditional classical true score methods. But the advantages of IRT should not obscure the value of classical test score theory approaches to measurement. Indeed, the two methods can complement each other in the development of measures as we shall argue below.

### **How “Good” Measures that Reflect Latent Constructs Are Constructed**

Most of us have been trained to think of measurement as an “item fitting” exercise. What do I mean by this? Most of us who build measures do not approach the task as developing a set of items that will meet the invariance criterion required of a Rasch scale. Instead, we borrow or develop items that we think represent some concept or concepts of interest, define a population of persons to whom we think it is relevant and then administer the items to a sample from that population. We then we engage in a type curve fitting. We factor analyze to eliminate the items that “don’t work.” Similarly, if we are using a 1PL IRT model we drop items that appear not to fit or we move from a one-parameter to a two-parameter model. If a single factor model doesn’t work, we move to a multiple factor model. If a single trait doesn’t work with an IRT approach, we move to a multi-dimensional approach. When we engage in this “item trimming” process we almost never go back and ask how well what we have left represents what we originally intended to measure. And we rarely replicate our work to see if the items do show invariance. Indeed, we rarely look at how the items operate for subpopulations within our samples. As a result, what we end up with is something far from invariance, something highly dependent upon the items and samples we have. This makes generalizeability from one study to another chancy

at best. Finally, this kind of approach will never lead to the kind of standardization we seek. Instead, I think Duncan (1984) is right (although I didn't at the time he wrote the book); we ought to begin with the end in mind, which is to attempt to construct measures that meet the invariance criterion. We will make many mistakes along the way, but at least we will be aiming at the right target.

Does taking this position mean that we should completely give up on what we have been doing? I don't think so, because as noted, it is very difficult to construct measures that meet strict measurement criteria. Unlike Duncan, I think there is a role for factor analysis, but it has to be used thoughtfully. Until we have good, solid axiomatic theories that embody fundamental social constructs in the same sense that the physical sciences have base and derived units against which to measure, to some degree we have to muddle along. But we can probably do it better than we have been. As we will show below, by specifying our domains carefully and examining measures that already exist to measure them, the use of exploratory factor analysis, confirmatory factor analysis and IRT modeling can move us towards better measurement in the social sciences. As a nice example of this combined approach, the reader is directed to the U.S. Health and Human Service's ongoing effort to measure health-related quality of life known as the Patient-Reported Outcomes Measurement Information System (PROMIS). See Reeve et al., 2007 for the details as to how the development of the quality of life measures is being undertaken.

It is important to have evidence that our measures capture the theoretical constructs as posited. Here are some steps that would seem to make sense when constructing subject-centered measures:

1. *Define the concept as carefully as possible. This will help in specifying the domain of the concept clearly and as completely as possible.* Any particular construct is intended to measure a *domain of meaning*. Most domains have various facets (Guttman, 1959) and the same principles of stratification used to sample persons can be used to improve the content validity (sometimes also called "coverage") of a measure. Social scientists often construct a few items on an ad hoc, one-shot basis, and apparently believe that the items measure the intended construct. In fact, constructing good measures is tedious, arduous, and time-consuming. The idea of sampling the facets of a construct's domain of meaning

is intuitively appealing, but most domains cannot be enumerated in the same way as a population of persons or objects is, so in practice the task is performed less rigorously than one would like. Two guidelines can be provided. First, researchers should search the literature carefully to determine how the concept to be measured has been used. This is done through literature searches. Others may have already developed measures of the concept that would need to be examined and evaluated. Second, researchers should ask whether their own observations and insights about the construct under consideration point to additional facets—especially if they have hunches about how the concept relates to a (set of) of other variable(s), especially outcome variables that will be the criterion-related validity of the measure constructed to be tested. Using these two approaches, one either uses or develops *sets* of items that capture the various facets or strata within the domain of meaning (Tryon, 1959). No simple criterion tells whether a particular domain of meaning has been properly sampled. Two precautions, however, can be taken to help ensure that the various facets within the domain are represented.

First, stratify the domain into its major facets. Note the most central meanings of the construct, making certain that all major meaning facets are represented. If a particular facet appears to involve a complex of meanings, subdivide it further into substrata. *The more refined the strata and substrata, the easier it is to construct items later, and the more complete the coverage of meanings associated with the construct.* Second, the items to be used should be pretested on a sample of persons similar to those in the studies that will use them. Pretest samples should be large enough to permit use of powerful multivariate tools of the sort described below.

*2. Use exploratory factor Analysis to get a sense of the dimensionality of the construct and whether the items load on factors as hypothesized.* This step has to be taken thoughtfully. Items can load on factors either because they are measuring a common construct or because they share methods variance. Examples of this include items that have a common stem (e.g., Most of my....) or a common stimulus (e.g., a focus on “teachers”). Furthermore, one needs to look carefully at the items that don’t load on any factors as well as those that do. It may be that some of these “singlets” reflect important facets or strata of the construct. Instead of eliminating them because they don’t share

variance with other items, one should instead build or find other items that appear to reflect this facet or stratum.

*3. Once one has determined what appears to be the dimensionality of the construct, use confirmatory factor analysis to examine the goodness of fit.* One of the virtues of confirmatory factor analysis is that one can posit factors that not only capture construct or facet variance, but methods variance as well and statistically test the fit. And assuming that the methods and construct variance are independent of each other, one can estimate the variance due to each. Ideally, one would work to eliminate any methods variance associated with common stems or common stimuli as a way to “purify” the items. One cannot do this with EFA.

*4. Estimate the reliability of one’s measures using internal consistency measures.* One should ensure that one’s measures have values at least as large as .7 using internal consistency measures such as Cronbach’s  $\alpha$  (Cronbach, 1951) or Heise and Bohrnstedt’s  $\Omega$ . (Heise and Bohrnstedt, 1971). I would not eliminate items based on these measures, however, since items that fit a Rasch model may not have a high internal consistency coefficient as scales constructed from a CTST model.

*5. Choose sets of items that correlate well together and represent separate dimensions or sub-dimensions and attempt to fit them to a Rasch model.* More generally, Rasch and IRT models are to be preferred over measures developed from CTST since they allow for item replacement without affecting the meaning of the measures. And as pointed out above, one of the objectives is to represent the full range of the latent variable which is better done with Rasch models. Or at least, one can test to see whether one is covering the full range.

There is no need to belabor this point, since it was well-discussed above. Suffice it to state that the goal should be to construct measures that meet the criteria of fundamental measurement.

*6. If measures do not meet the Rasch assumptions and fit statistics, use IRT models with more relaxed assumptions, such as a 2PL model.*

7. Whichever model is used, make certain that the parameter estimates are invariant for important subpopulations formed on the basis core socio-demographic measures (e.g., gender, race/ethnicity). Successful replication with maximally diverse samples adds credibility that one's measures are robust against the invariance criterion.

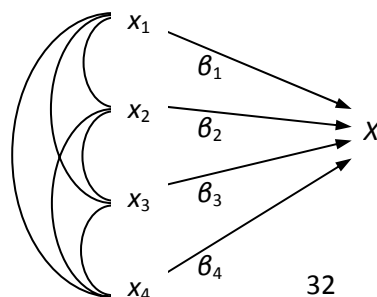
8. Examine the scales and subscales for sparseness and weakness of items and develop other items to test as additions or replacements before going into the field. Where one has the luxury of working within a longitudinal or trend design, always test new and revised items as possible replacements for weaker items.

Following these steps does not guarantee that a measure will meet the criteria for good measurement, but it certainly will move one closer to the model than the haphazard way in which many measures are constructed.

### **Index Construction Where the Indicators Determine Rather than Reflect a Construct**

To this point, our discussion has focused on measurement that assumes a set of observed indicators reflect or are caused by an unobserved or latent variable. But there are many cases in sociology, economics and social indicator research where the assumption is that the indicators comprise the construct under study. This type of measurement model is sometimes called “formative” compared to the “reflective” model discussed above. (Diamantopoulos and Winklhofer, 2001). Bollen (1984) refers to these as “causal” as opposed to “effect” indicators. Probably the most important construct of this type in sociology is socio-economic status which is typically seen as a function of income, education and occupation. An example from economics is the Consumer Price Index which is based on the cost of a marketbasket of goods and services. Unlike Figure 1, the representation of indicators as components or determinants of an index has the arrows going in the other direction as shown in Figure 4.

**Figure 4. A Schematic Representation of an Index Composed of Four Variables**



In the typical index, the  $x$ s are simply added together into a composite,  $X$ . Furthermore, if the  $x$ s are in the same metric (e.g., have the same response categories), the  $\beta$ s typically will equal each other and equal 1.0. In other cases, one might weight the  $x$ s inversely to the size of their variances or weight them based on theory, differential utilities, or preferences associated with them. Weighting by preferences might occur if a community were to decide to build an index on the quality of services provided by the community. That is, a decision about how to weigh the various services that comprise the index might be based on a community survey where respondents are asked about the relative importance of each service. Importantly, there is no statistical way to determine how the weights should be assigned.

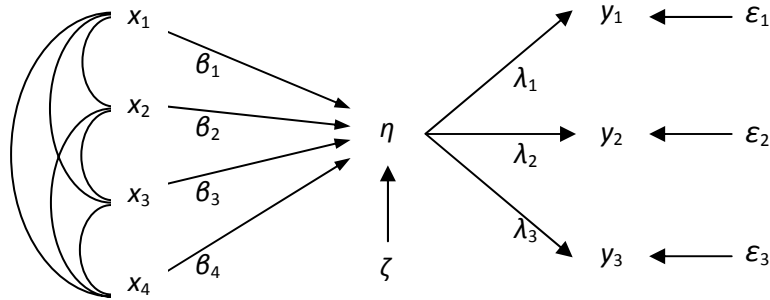
Notice that the composite does not have an error term associated with it. The index is determined solely by the  $x$ s that compose it and the weights associated with them. And finally, for this type of index, there is no requirement that the  $x$ s are internally consistent, say as measured by Cronbach's  $\alpha$  as is the case in an index where the  $x$ s reflect a latent construct (Bollen and Lennox, 1991; Coltman, Devinney, Midgley, and Venaik, 2008).

There is a way to determine statistically how to weight the components of a formative index *if* one has both multiple causes and multiple indicators. In this situation, one can use structural equation modeling (Bollen, 1989) to estimate all of the parameters of interest. Logically, these models are called multiple indicator-multiple cause (MIMIC) models. Early examples of MIMIC models can be found in Hauser and Goldberger (1971), Bohrnstedt (1977) and Hauser and Wong (1989). A schematic of a MIMIC model is shown in Figure 6. Models of this sort allow one to determine the weights on the component measures going into the latent variable of interest, labeled  $\eta$  in Figure 5. It is of paramount importance that one has specified this type of model correctly to get accurate estimates of the causal path between the components and the latent variable of interest. In this regard, it can be instructive to use slightly different specifications on the indicator side of the model (the  $y$ s) to check the stability of estimates of the  $\beta$ s associated with the  $x$ s.

While I have chosen to include the discussion of MIMIC models in a section on formatively determined indices, note that the variable of interest, say SES, is in fact unobserved, and no index is created. The tradeoff, and it is an important one, is that these models do allow one to obtain useful estimates of impact on the “causes” side of the model. As Hauser shows in his

paper which follows, careful modeling can reveal the relative impact of such causes on outcomes such as intergenerational mobility.

**Figure 5. Multiple Indicators, Multiple Causes (MIMIC) Model**



**The Use of Standards Might be a Useful Step Forward in Certain Situations**

In this final section of the paper, I introduce one other thought on the role that standards might play in measurement in the social sciences. Certainly this approach will not work in all cases, but perhaps it will work in at least some. Consider the area of educational assessment. Since the early 1990s there has been a movement as part of the education reform movement to establish both content and performance standards by subject area. As a leader in this movement, the National Assessment Governing Board (NAGB) set performance standards, called Achievement Levels, first in mathematics and then in reading, and eventually in almost all of the tested subjects. The Governing Board began by broadly defining what is meant by Advanced, Proficient, and Basic Performance. Then, using methods that grew out of standard setting used for certification by professions (e.g., board certification in medicine) sets of expert judges examined the NAEP items for a given subject area and examined where they performed on the NAEP scale (that is, their relative “difficulty”). They then, first individually, and then as a group placed cut points along the NAEP scale that in their judgment represented Advanced, Proficient, and Basic performance. This process was done for Grade 4, 8 and 12 NAEP within each subject area. There has been controversy about both the way the standards were set as well as the where the cutpoints were made on the NAEP scale, especially where the cutpoints were placed

for “Advanced” performance.<sup>20</sup> In addition to the standards that NAGB set on NAEP, the *No Child Left Behind* legislation required states to set their own performance standards on their state assessments and are required to show annual yearly progress against them with the goal of 100% of their students being at least proficient by 2014.

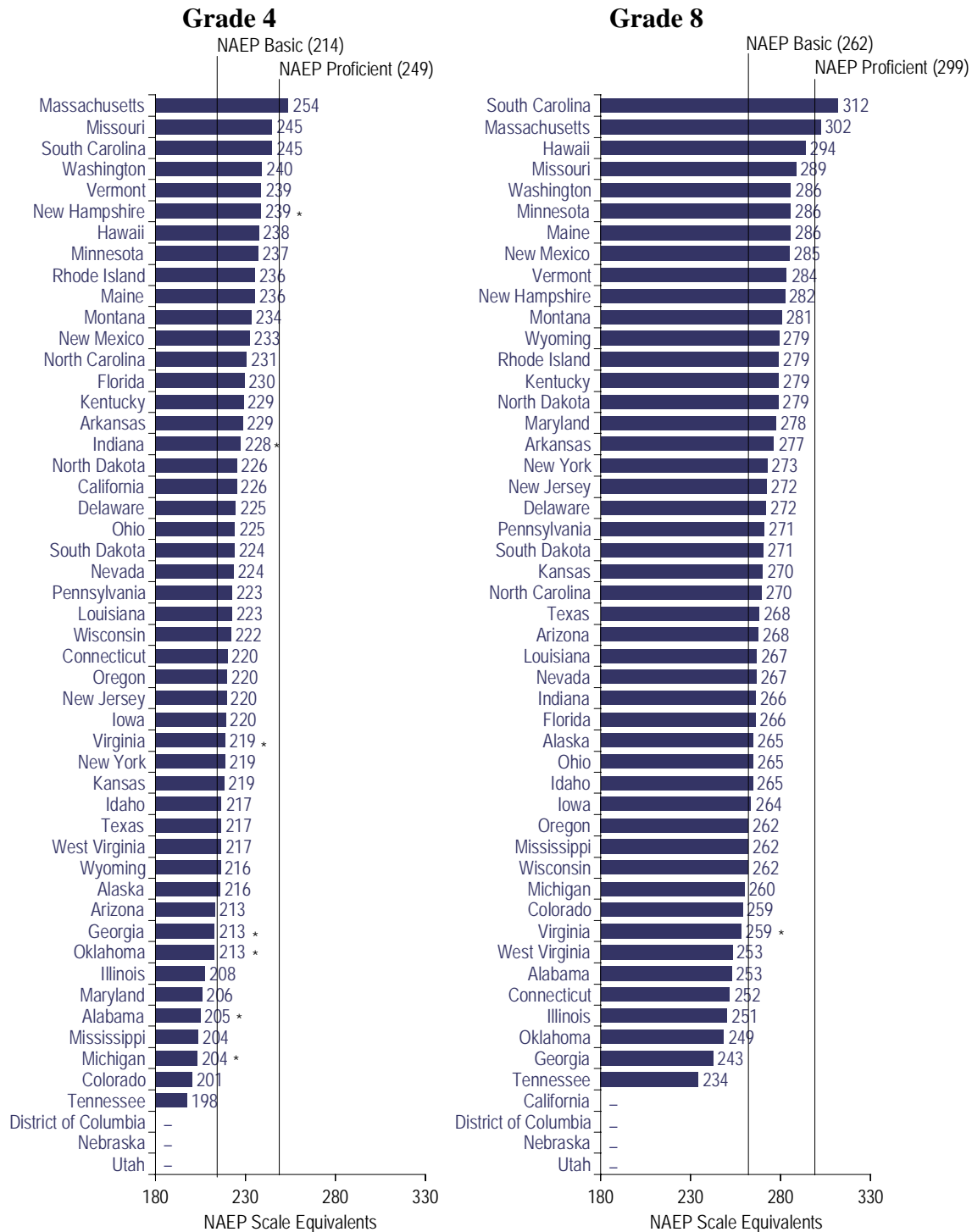
Because the states use different assessments and have set their own standards on them, a natural question that arises is whether they can be compared, and if so, how? The answer is they can be compared. But doing so requires finding a *common* standard against which they all can be compared. The National Center of Education Statistics has issued three reports since 2000 in which they project all of the state standards onto the NAEP scale. This can be done because all of the state’s must participate in NAEP if they are to receive Title I funds. Figure 6 shows how the various state standards in Grade 4 and Grade 8 mathematics stack up when measured against the 2007 NAEP Achievement Levels on mathematics. Briefly what is seen is that 1) only two of the states’ standards for proficiency reaches the NAEP standard for proficiency (South Carolina, Grade 8 and Massachusetts in Grades 4 and 8) , and 2) there is a tremendous amount of variation in where states have set their standards (the range is over two standard deviations on the NAEP scale).

A project building on this work has been done by my colleague Gary Phillips. He is interested in international benchmarking. There are five levels of performance set by the International Association for the Evaluation of Educational Achievement (IEA) on Trends in International Mathematics and Science Study (TIMSS), an international assessment carried out over various years, the most recent of which is in 2007. Phillips (2009) argued that the 5 levels could reasonably be labeled “A” for the highest, “B” for the next highest, etc. It turns out that a “B” on the TIMSS scale is nearly equal to Proficient on the NAEP mathematics scale, a criterion that several of the top performing nations in the world meet.

---

<sup>20</sup> See Glaser, Linn and Bohrnstedt (1993)

**Figure 6. NAEP scale equivalent scores for the state grades 4 and 8 mathematics standards for proficient performance, by state: 2007**



— State assessment data not available.

\* Relative error greater than .5.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 Mathematics Assessments. U.S. Department of Education, Office of Planning, Evaluation and Policy Development, *EDFacts SY 2006-07*, Washington, DC, 2008. The National Longitudinal School-Level State

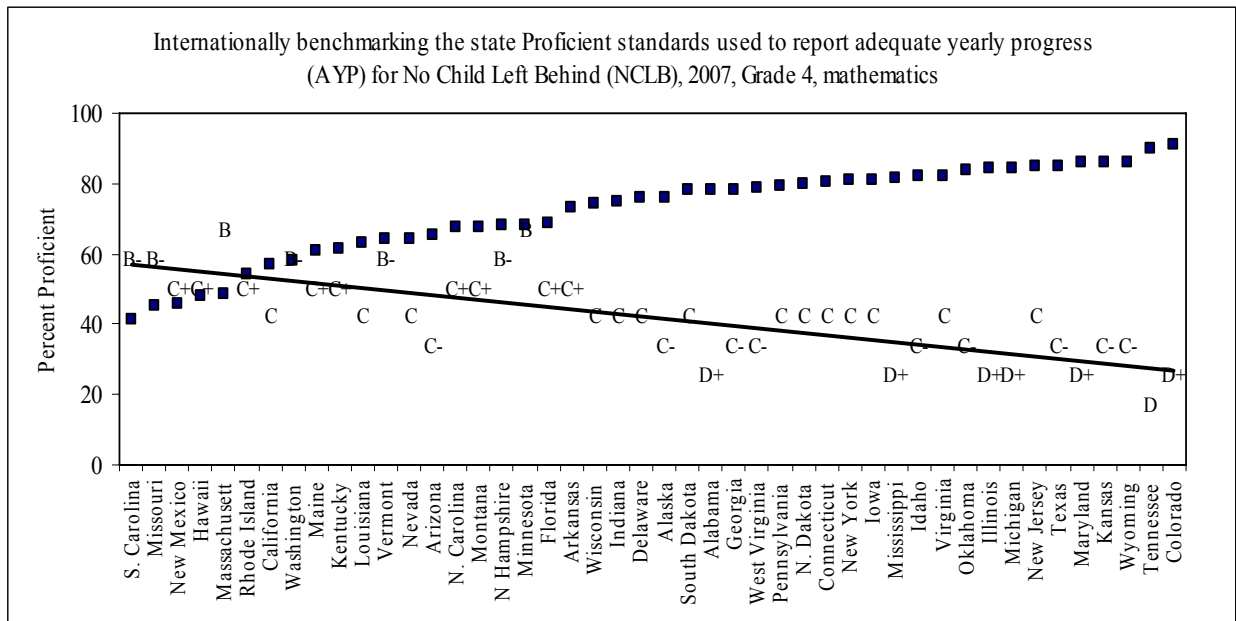
Phillips (2010) next took the results from the NAEP mapping analysis reported in Figure 6 and projected them onto the TIMSS performance measures. The results for Grade 4 mathematics are shown in Figure 7. The states' standards are arrayed such that the states with the highest standards are on the left side of the figure and those with lower standards on the right. Phillips fit a straight line to the figure. The TIMSS "grade" is shown for each state as well. South Carolina, Missouri, Massachusetts, Vermont, New Hampshire, and Minnesota all set standards at B or B<sup>-</sup> levels – reasonably high standards. By contrast, seven states set their standards at D<sup>+</sup> or D level.

Phillips also plotted the percent proficient on states' assessments in Grade 4 mathematics by state (top line in Figure 5) to demonstrate the inverse relationship between where states set their standards and the percentage of their students labeled "proficient." Note that those states that set high standards have relatively lower percentages of their students meeting proficiency on their state assessments than those states that set lower standards.

Again, this is all made possible by putting results on common scales – NAEP and TIMSS.

To summarize this section, I have used two examples that show that is possible to measure where the various states have set their proficiency standards by using common metrics – in one case the comparisons are with a U.S. standard and in the second, an international standard. Neither of the metrics has true zero points, nor are the distances between the points on the two scales likely equal. Furthermore, the standards set are at some level arbitrary since they depended upon professional judgment. Nonetheless, both are examples of *common metrics* that I believe have clear utility. How many other examples of this sort that can be found, I am not certain, but I thought it worth mentioning.

**Figure 7: An Example of the Value of a Common Metric for International Benchmarking**



**Summary**

Let me briefly summarize what I think are the main take-ways from my paper:

- (1) Measures are social constructs and the process of gaining standardization around measures is very much a social process involving negotiations among social actors;
- (2) Standardization is impelled along when there are strong commercial, political or scientific reasons for doing so; and
- (3) Science has a strong and central role to play in the development of standards for measurement.
- (4) We have not discovered or figured out how to define the kind of fundamental quantities in the social sciences that exist in the physical sciences;

- (5) Our concepts are large in number and for the most part do not bear the kind of simple relationships to one another as is true in the physical sciences,
- (6) We lack strong mathematical theories against which to evaluate our measurements and vice-versa.
- (7) There are models such as the Rasch model which if we could construct indicators that fit the model would allow us to develop measures that meet the invariance criterion that true measurement requires.
- (8) In the absence of being able to create Rasch models, there are nonetheless a set of criteria that can be employed that will increase the utility of our measures.
- (9) It is possible to construct common metrics that don't meet the criterion of strong measurement, but have utility nonetheless.

## References

- Abelson, R.P. and J.W. Tukey (1963) "Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order." *Annals of Mathematical Statistics* 34, 1347-1369.
- Andrich, D. (1985). "An elaboration of Guttman scaling with Rasch models for measurement". In N. Brandon-Tuma (Ed.), *Sociological Methodology*, San Francisco, Jossey-Bass. (pp. 33-80.).
- Baker, B.O., C.D. Hardyck, and L.F. Petrinovich, L.F. (1966). "Weak measurements vs. strong statistics: An empirical critique of S.S. Stevens' proscriptions on statistics." *Educational and Psychological Measurement*, 26, 291-309.
- Bohrnstedt, G. W. (1977) "Use of the Multiple Indicators-Multiple Causes (MIMIC) Model." *American Sociological Review* 42: 656-665.
- Bohrnstedt, G.W. (2010) "Measurement." In P. Marsden and J. Wright (Eds.), *Handbook of Survey Research* (Second Edition) New York: Academic Press (forthcoming).
- Bogardus, E. S. (1926) Social Distance in the City. *Proceedings and Publications of the American Sociological Society*. 20, 40-46.
- Bollen, K. (1984) "Multiple indicators: Internal consistency or no necessary relationship" *Quality and Quantity* 18:377-385.
- Bollen, K. (1989) *Structural equations with latent variables*. New York: Wiley.
- Bollen, K and Lennox, R. (1991) "Conventional wisdom on measurement: A structural equation perspective" *Psychological Bulletin* 110: 305-314.
- E. F. Borgatta (1961) "Toward a methodological codification: The shotgun and the saltshaker." *Sociometry*, 24, 432-435.
- Borgatta, E. F. and G. W. Bohrnstedt (1980) "Level of measurement - Once over again." *Sociological Methods and Research*, 9, 147-160.
- Christ, C. (1966) *Econometric Models and Methods*. New York: Wiley.
- Coltman, T. Devinney, T.M. Midgley, D.F. and Venaik, S. (2008) "Formative versus reflective measurement models: Two applications of formative measurement." *Journal of Business Research* 61:1250-1262.
- Cronbach, L.J. (1951) "Coefficient alpha and the internal structure of tests", *Psychometrika* 16, pp. 297-334.

- Diamantopoulos, A. and Winklhofer, H.M. (2001) "Index construction with formative indicators: An alternative to scale development." *Journal of Marketing Research* 38:269-277.
- Duncan, O.D. (1984) *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.
- Durbin, J. (1954) "Errors in variables," *Review of the International Statistics Institute*, 22. 23-32.
- Embretson E. and S. P. Reise (2000) *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Mahwah, NJ.
- Glaser, R., R. Linn, R., and G. W. Bohrnstedt, (Eds.) (1993). *Setting performance standards for student achievement*. Stanford University: The National Academy of Education.
- Goldberger, A. S. (1971) "Econometrics and psychometrics: A survey of communalities." *Psychometrika*, 36, 83-107.
- Goldberger, A. S. (1972a) "Maximum likelihood estimation of regressions containing unobservable independent variables." *International Economic Review*, 13, 1-15
- Goldberger, A. S. (1972b) "Structural equation models in the social sciences." *Econometrica* 40, 979-1001.
- Griliches, Z. and V. Ringstad (1970). Error-in-the-variables bias in nonlinear contexts. *Econometrica* 38, 368-370.
- Guttman, L. (1950)." The basis for scalogram analysis." In Stouffer *et al. Measurement and Prediction*. The American Soldier Vol. IV. New York: Wiley
- Guttman, L. (1977). "What is not what in statistics." *The Statistician*, 26, 81-107.
- Hambleton, R.K., H. Swaminathan and H.J. Rogers (1991), *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park, CA.
- Hamblin, R. K. (1971) "Ratio measurement for the social sciences." *Social Forces* 50: 191-206
- Hamblin, R. K. (1974) "Magnitude measurement and theory." In H. M. Blalock, Jr. (ed.) *Measurement in the Social Sciences: Theories and Strategies*. Chicago: Aldine, pp.61-120.
- Hauser, R. M. and A.S Goldberger (1971) "The treatment of unobservable variables in path analysis." In H. L. Costner (ed.) *Sociological Methodology: 1971*. San Francisco: Jossey-Bass, pp 81-117.
- Hauser, R. M. and Wong, R.S. (1989) "Sibling resemblance and inter-sibling effects in educational attainment." *Sociology of Education* 62: 149-171.

- Jöreskog, K.G. (1969), "A general approach to confirmatory maximum likelihood factor analysis", *Psychometrika*, Vol. 34, pp. 183-202.
- K.G. Jöreskog, K.G. (1970), "A general method for analysis for covariance structures", *Biometrika*, Vol. 57, pp. 239-251.
- Jöreskog, K. G. (1971), "Statistical analysis of sets of congeneric structures", *Psychometrika*, Vol. 36, pp. 109-134.
- Langevin, L. (1961) "The introduction of the metric system." *Impact of Science on Society* 11: pp. 77-95.
- Northrop, F.S.C. (1947) *The Logic of the Sciences and the Humanities*. New York: Meridian Books.
- Osgood, C.E., G.J. Suci, and P.H. Tannenbaum (1957) *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.
- Phillips, G. W. (2009) *The Second Derivative: International Benchmarks in Mathematics for U.S. States and Districts*. Washington D.C.: American Institutes for Research.
- Phillips, G.W. and T. Jiang (2010) "Internationally Benching State Performance Standards." Washington D.C.: American Institutes for Research.
- Primack, J.R. and N. E. Abrams (2006) *The View from the Center of the Universe: Discovering our Extraordinary Place in the Cosmos*. New York: Riverhead Books.
- Rainwater, L. (1972) *What Money Can Buy: The Social Meaning of Poverty*. New York: Basic Books.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology, pp. 321-334 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley: University of Chicago Press, 1980.
- Reeve, B.C., et al. (2007), "Psychometric evaluation and calibration of health-related quality of life item banks", *Medical Care*, Vol. 45 Supplement 1, pp. S22-S31.
- Samejima, F. (1969) "Estimation of latent ability using a response pattern of graded scores." *Psychometrika Monograph*, No. 17
- Shinn Jr., M. (1969) "An application of Psychophysical scaling techniques to the measurement of national power." *Journal of Politics* (31) 932-951.
- Smith, P.C. and Kendall, L.M. (1963) "Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales." *Journal of Applied Psychology* 47:149-155.

- Stevens, S.S. (1946) "On the theory of scales of measurement." *Science*, 103, 677-680.
- Stevens, S. S. (1951) "Mathematics, measurement, and psychophysics." In S.S. Stevens (ed.) *Handbook of Experimental Psychology* New York: Wiley.
- Stevens, S. S. (1975) *Psychophysics*. New York: Wiley.
- Sydenham, P.H. (1979) *Measuring Instruments: Tools of Knowledge and Control*. Stevenage, U.K.: Peter Peregrinus.
- Thissen, D. and Steinberg, L. (2009), "Item response theory", in Millsap, R. and Maydeu-Olivares, A. (Eds.), *The Sage Handbook of Quantitative Methods in Psychology*, Sage Publications, London, pp. 148-177.
- Thurstone, L.L. (1928) "Attitudes can be measured." *American Journal of Sociology*, 33, 529-554.
- Thurstone, L.L. (1927a). "A law of Comparative Judgment," *Psychological Review*, 34, 278-286.
- Thurstone, L.L. (1927b). "The method of paired comparisons for social values," *Journal of Abnormal and Social Psychology*, 21, 384-400.
- L. L. Thurstone (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- L. L. Thurstone (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.
- Torgerson, W. S. (1958) *Theory and Methods of Scaling*. New York: Wiley.
- Tryon, R.D. (1959) "Domain sampling formulation of cluster and factor analysis." *Psychometrika*, 24, pp. 113-135.
- Velleman, P. and L. Wilkinson (1993) "Nominal, Ordinal, Interval, and Ratio Typologies are Misleading." *The American Statistician* 47, 65-72.
- Yen, W. and A.R. Fitzpatrick (2006), "Item response theory", in Brennan, R.L. (Ed.), *Educational Measurement*, fourth edition, Praeger Publishers, Westport CN, pp. 11-153.
- Zabrowski, E. (1979) *Fundamentals of Physical Measurement*. North Scituate, MA: Duxbury Press.

February 24, 2010