

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

George W. Bohrnstedt  
American Institutes for Research

# Overview of Presentation

- Measurement in the physical sciences
- History of measurement in the social and behavioral sciences
- S.S. Stevens on levels of measurement
- The role of classical test score theory in current measurement
- Item response theory (IRT) and Rasch models
- Index construction
- Examples using standards to build common metrics
- Conclude with some “takeaways”

# Learning from Measurement in the Physical Sciences

- Metrology and the scientific study of measurement
- The role of governments in common metrics
- The role of science in common metrics
- Conclusions, standardization:
  - Very much a social process
  - Driven by strong commercial and political interests
  - Depends heavily on a strong role from science

# Measurement in the Physical Sciences

- Characterized by standards based on strong theory and experimentation are key to measurement
  - Meter as a standard for length
  - Seconds as a standard for time
- Typically read-outs are used as displays (e.g., thermometer, meters) that are closely calibrated to standards
- The role of National Bureau of Standards (NBS), now National Institute of Standards and Technology (NIST), in maintaining and updating measures

# Seven Base Units Maintained by NIST

<i>Base Quantity</i>	<i>Unit</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

# Derived Measures

Examples of derived measures:

$$\text{Area} = \text{m}^2$$

$$\text{Velocity} = \text{m/s}$$

$$\text{Acceleration} = \text{m/s}^2$$

$$\text{Luminance} = \text{cd/m}^2$$

An example of derived measures derived from other derived measures:

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

# Do Phenomena Have a Reality beyond Their Measurement?

“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)

- Cannot see or touch weight or length, for example (although we can sense both), but both are calibrated against tangible standards
- We do have some clear tangible measures in the social sciences, e.g., birth, age, marital status, number of children, date of death
- The picture becomes much murkier when one thinks of cognitive concepts, including attitudes, values, and beliefs
- Additional examples include organizational concepts such as school climate and organizational learning, as well as societal level concepts such as anomie, social disorganization

# Latent Constructs in the Physical and Social Sciences

- In the social sciences, most of our constructs are similar to what Northrop called “concepts of intuition” – concepts we perceive
- By contrast the physical sciences have what Northrop calls “concepts of postulation” – concepts derived from axiomatic, deductive theory
- In the social sciences we often can only observe “reality” for many concepts by examining the covariation between observed indicators

# The Interplay between Theory and Measurement

- In the physical sciences, where we have good theory, measurements can often be used to confirm, reject, or refine theories
- In the social sciences, often not clear whether it is the theory that is faulty, or the measures, or both
- In the physical sciences, theory is often viewed as a necessary precursor for measurement
- Lack of strong theory in the social sciences likely plays into the lack of well-accepted common metrics

# What Are the Takeaways From This Review?

- 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
- Our concepts are large in number, “fuzzy,” and do not have a simple relationship to one another as is true in the physical sciences (“Ballungen” concepts)
- Lack strong axiomatic theories against which to evaluate and inform our measures
- Unclear if these criteria drawn from the physical sciences are *sine qua non* for the development of better measures within the social sciences or not

# Measurement in the Social Sciences – Sociology

- LePlay in the mid-1850s is credited with establishing what has become modern day survey – still the most popular method for collecting data in sociology
- Most measurement in sociology is what Torgerson called *subject-centered* – where persons are placed on a continuum
- Likert scales and the semantic differential are two examples of subject-centered measures

# Measurement in Sociology (cont.)

- Can contrast subject-centered with *stimulus-centered* measurement, where one orders stimuli from high to low on a scale

# Sociology – Guttman Scales

- Guttman scales had some popularity in the 1950s and 60s. It was both subject- and stimulus-centered measurement in that it ordered both items and persons on the scale
- An example is the Bogardus social distance scale, with a series of questions such as:
  - Are you willing to permit immigrants to live in your community?
  - Are you willing to have immigrants live next door?

# Sociology – Guttman Scales (cont.)

- Guttman scaling is a deterministic model that has fallen out of use but is an important precursor to Item Response Theory (IRT) scaling

# Psychology – Psychophysics – Thurstone

- L.L. Thurstone took Fechner's psychophysical work on sensation and perception and applied it to attitude and value measurement using the *method of paired comparisons*
- Separated the scaling of items from the scaling of persons
- The scaling of attitude items lead him to develop the *comparative law of judgment*

# Psychology – Psychophysics – Thurstone (cont.)

- Thurstone's work in attitudes was also important because he
  - Developed the notion of *invariance in measurement*:
    - Items have the same meanings for all respondents and therefore subgroups of respondents
    - With the addition of items to the scale, the ordering and distance between any two respondents remains unchanged
  - Scaled both items and persons
  - Plotted cumulative probability distributions for each item as a way to show how well each discriminated in the measurement of the attitudes

# Psychology – Psychophysics – S.S. Stevens

- S.S. Stevens' work is very closely associated with classical psychophysics – scaling sensations
- Subjects would make judgments (e.g., loudness) using ratios
- He plotted perceptual values against actual stimulus values across various stimuli and noticed a regularity in the resulting pattern

# Psychology – Psychophysics – S.S. Stevens (cont.)

- Led him to posit a general psychophysical law stated as:

$$\psi = \alpha\varphi^\beta$$

where

$\psi$  is the perceived magnitude,

$\varphi$  is the actual magnitude of the stimulus intensity,

$\alpha$  is a constant that varies depending upon the units of measurement, and

$\beta$  is the parameter to be estimated

# Psychology – Psychophysics – S.S. Stevens (cont.)

When put in logarithmic form the equation becomes linear in the logs:

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

- Steven's "law" is an empirically, not theoretically, derived law
- He did little to apply his work to social phenomena, but Shinn, Hamblin, and Rainwater have in political science and sociology

# Psychology – the Measurement of Intelligence

- Galton's early work led to Spearman positing a *g*- or general factor to explain intelligence. Laid the groundwork for the earliest work in *factor analysis*.
- That led to work in bi-factor analysis and then to Thurstone's multiple factor analysis in which he posited a set of *primary mental abilities*
- Multiple factor analysis has since become known as *exploratory factor analysis*, because it is used to explore the dimensionality of a set of items

# Psychology – the Measurement of Intelligence (cont.)

- In early 1960s, Karl Jöreskog developed *confirmatory factor analysis* (CFA), which allows one to test rigorously dimensionality of items as well as hypotheses about which factors the various items measure
- EFA and CFA have played important roles in measurement in the social sciences as we shall see in a later section of the talk

# S.S. Stevens and Levels of Measurement

- Stevens is better known in the social sciences for his work on levels of measurement than his work in psychophysics, although the former grew out of the latter
- He defines four levels or types of measures – nominal (N), ordinal (O), interval (I), and ratio (R)
- Each has permissible as well as non-permissible transformations

# Levels of Measurement (cont.)

<i>Scale</i>	<i>Permissible Transformations</i>	<i>Examples</i>
N	Substitute any number with another	Football jersey numbers
O	Any change that preserves rank order	Hardness of minerals
I	Multiplication by or addition of a constant	F and C temp. scales
R	Multiplication by a constant	Length, weight

For Stevens, “true” measurement in the sense of what is found in the base units in the physical sciences requires ratio-level measurement.

# Levels of Measurement (cont.)

- Nominal level measurement – mode, contingency coefficient
- Ordinal – median, rank order correlation
- Interval – mean, SD, correlation, regression
- Ratio – geometric mean, coefficient of variation

Note the cumulativeness to the permissible operations as one moves up the measurement hierarchy.

# Critiques of Steven's Dicta

Lord (1953) – Choice of statistics depends upon the meaningfulness of the analysis undertaken, not the level of measurement

Tukey (1961) – Argues that just because scale types are absolute doesn't mean that statistical methods employed must be absolute as well

Others – the power of statistical analysis is lost when one uses rank-order as opposed to parametric analyses. Also, when the underlying variable is continuous in nature, parametric statistics are appropriate – one is just measuring with error.

# Critiques of Steven's Dicta (cont.)

Duncan (1984) is especially harsh on Steven's dicta about the use of nominal variables

- Argues that he confuses classification with measurement
- Dichotomous variables play an especially important role in the social sciences (i.e., presence versus absence, on versus off)

# Critiques of Steven's Dicta (cont.)

- Also takes issue with Steven's very definition of measurement, which is the assignment of numbers to outcomes
- Argues that instead, measurement is the assignment of numbers that correspond to *different degrees of a quality or property* of some object or event. That is, measurement involves the *magnitude* of the concept being measured.
- Much damage done by Steven's dicta in that it was routinely picked up in social science statistics texts.

# Factor Analysis and the Use of Linear Composites in the Social Sciences

- Typical to use Likert items surveys to measure concepts of interest
- One then uses EFA to check on dimensionality of the items
- The models are based on classical true score theory (CTST) where  $x_i = \tau + \varepsilon_i$
- The results are used to build linear composites, which are checked for internal consistency reliability

# Criticisms of the EFA Approach

- Ad hoc – what Duncan called “ a correlational science of inexact constructs”
- Rarely is the replicability of the dimensionality checked for relevant subgroups or in new or holdout samples
- This approach will never lead to the kinds of fundamental measures found in the physical sciences
- Privileges reliability over validity

# The Use of Confirmatory Factor Analytic Methods to Build Composites

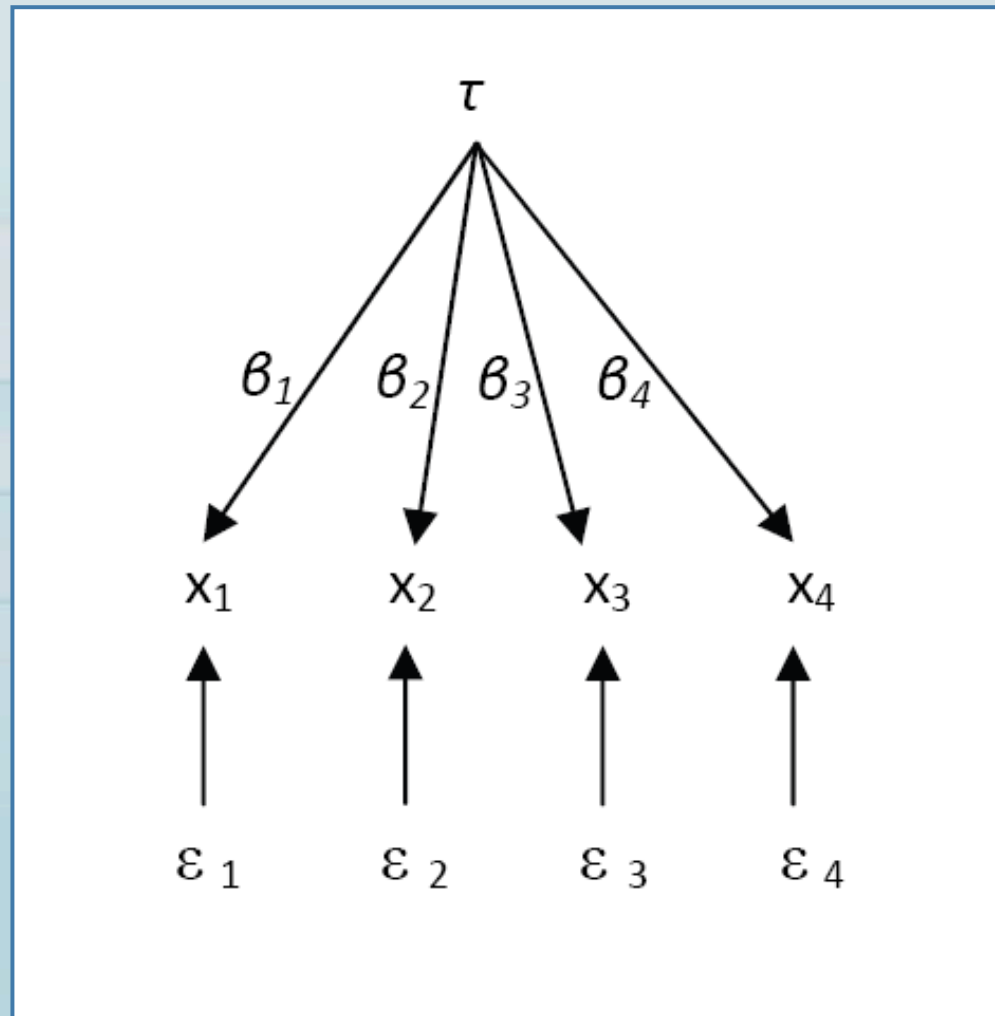
Jöreskog (1969) define *congeneric measurement* as:

$$x_i = \mu_i + \beta_i \tau + \varepsilon_i$$

That is, an item is a linear function of a weighted true score, a constant to take into account the unit of a measure and random measurement error.

Diagrammatically this is what congeneric measurement looks like for 4 items:

# A Plot of a Confirmatory Factor Analytic Model



# Confirmatory Factor Analytic Methods to Build Composites (cont.)

- CFA can be estimated by ML methods, which provide standard errors and chi-square tests for goodness of fit (absent from most EFAs)
- Allows one to *test* the assumption that measures are congeneric
- Generalizes to the multiple factor model
- Allows one to test for methods factors, correlated errors, equal loadings, etc.

# Limitations of CTST for Scale Construction in the Social Sciences

- Even when using CFA, one of the limitations is that there is no explicit concern with how well the items measure the entire range of the latent variable
- One of the virtues of Thurstone and Guttman was an explicit recognition of where on the latent dimension one's items are measuring as well as how well the items discriminate
- But both of their models suffered from a lack of strong statistical estimation methods, as do most exploratory factor analysis methods

# IRT Methods for Scale Construction in the Social Sciences

- Item Response Theory was developed in the in the early 1960s primarily to measure latent ability and achievement
- It is worth noting that Paul Lazarsfeld, a sociologist interested in survey research, had worked out much of the mathematics for it in what he called latent trait theory. However, he did not have the computational machinery to estimate the models efficiently.

# IRT Methods for Scale Construction in the Social Sciences (cont.)

- While there were a few early applications of IRT to social measurement (e.g., Reiser, 1980), it has been David Thissen and his student Lyn Steinberg who have done more than any others to make this transition
- But there is increasing interest in IRT applications for the measurement of social and psychological latent concepts (e.g., the measurement of health-related quality with the Patient-Reported Outcomes Related Information System, or PROMIS)

# The One Parameter IRT Model (1PL)

- Simplest IRT model is for dichotomous items where the probability of person  $p$  getting item  $i$  right or agreeing with the item is given by

$$\Pr (x_i = 1 | \vartheta_p, \beta_i) = e^{(\vartheta_p - \beta_i)} / (1 + e^{(\vartheta_p - \beta_i)})$$

where  $\vartheta_p$  is a person  $p$ 's true score (e.g., "true" attitude) and

$\beta_i$  is the "difficulty" of item  $i$ .

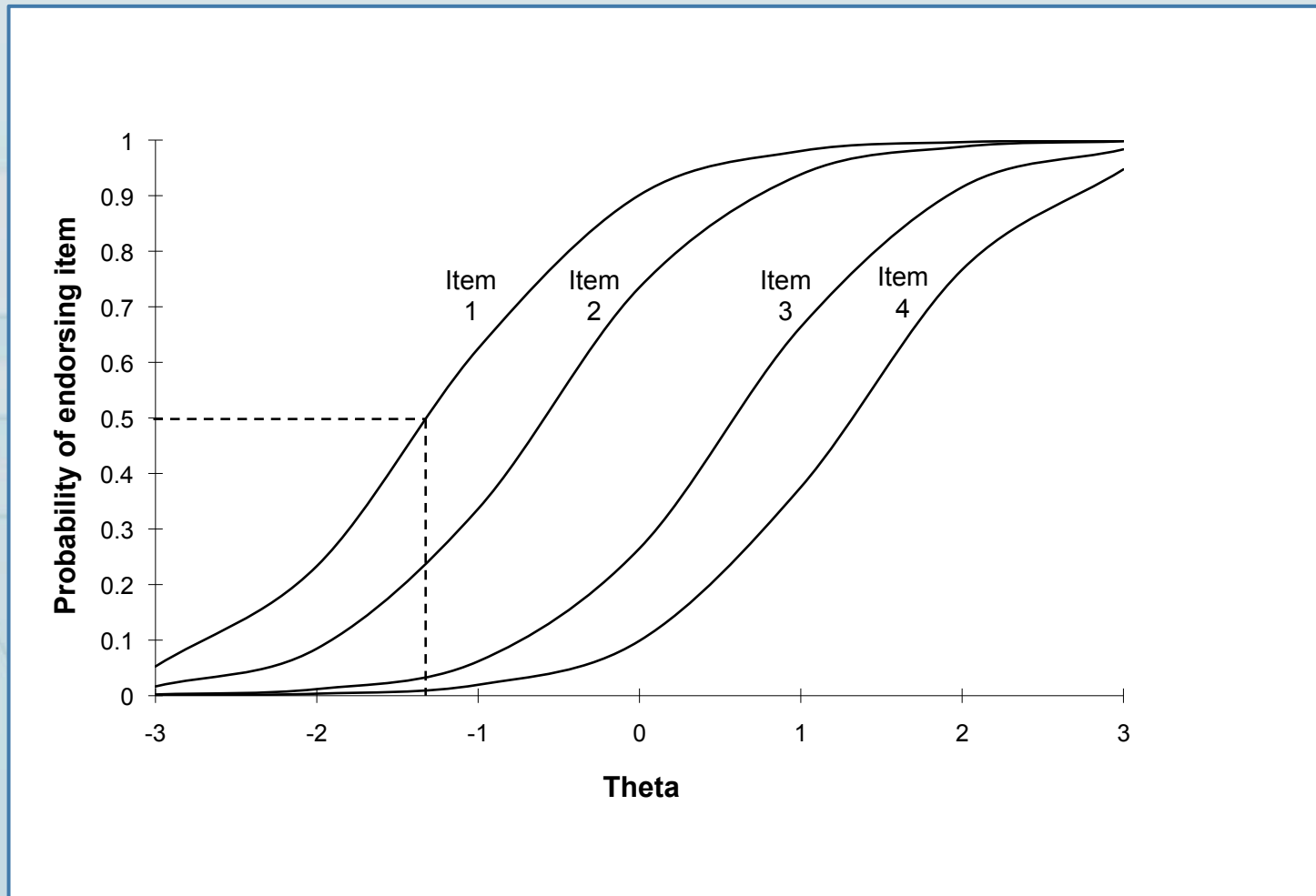
# The One Parameter IRT Model (1PL) (cont.)

- Difficulty parameter indicates where the item is operating along the latent scale
- The model is probabilistic in that the value for  $\beta_i$  is that point on the latent dimension where the probability of getting it right or agreeing with it is .50.
- If one holds a strong attitude or value, then the probability of agreeing will be higher than .5; if one holds a weak attitude, then the probability of agreement will be less than .5.

# The One Parameter IRT Model (1PL) (cont.)

- The plot of the probability of agreement (or getting an item correct) as one moves from the negative end of the latent continuum to the positive end is called an *item characteristic curve* (ICC)
- There is one ICC for each item
- By convention, the metric of most 1PLs is standardized around a mean of zero and SD of 1.0

# A Plot of the ICCs for 4 Items Fit to a 1PL Model



# The Rasch Model

- A Rasch Model is a particular type of 1-parameter latent trait model
- While the trait measured can be standardized as was true for the 1PL, it can also be used in the native “right-wrong” or agree-disagree metric, which is appealing
- The model has some very appealing measurement characteristics

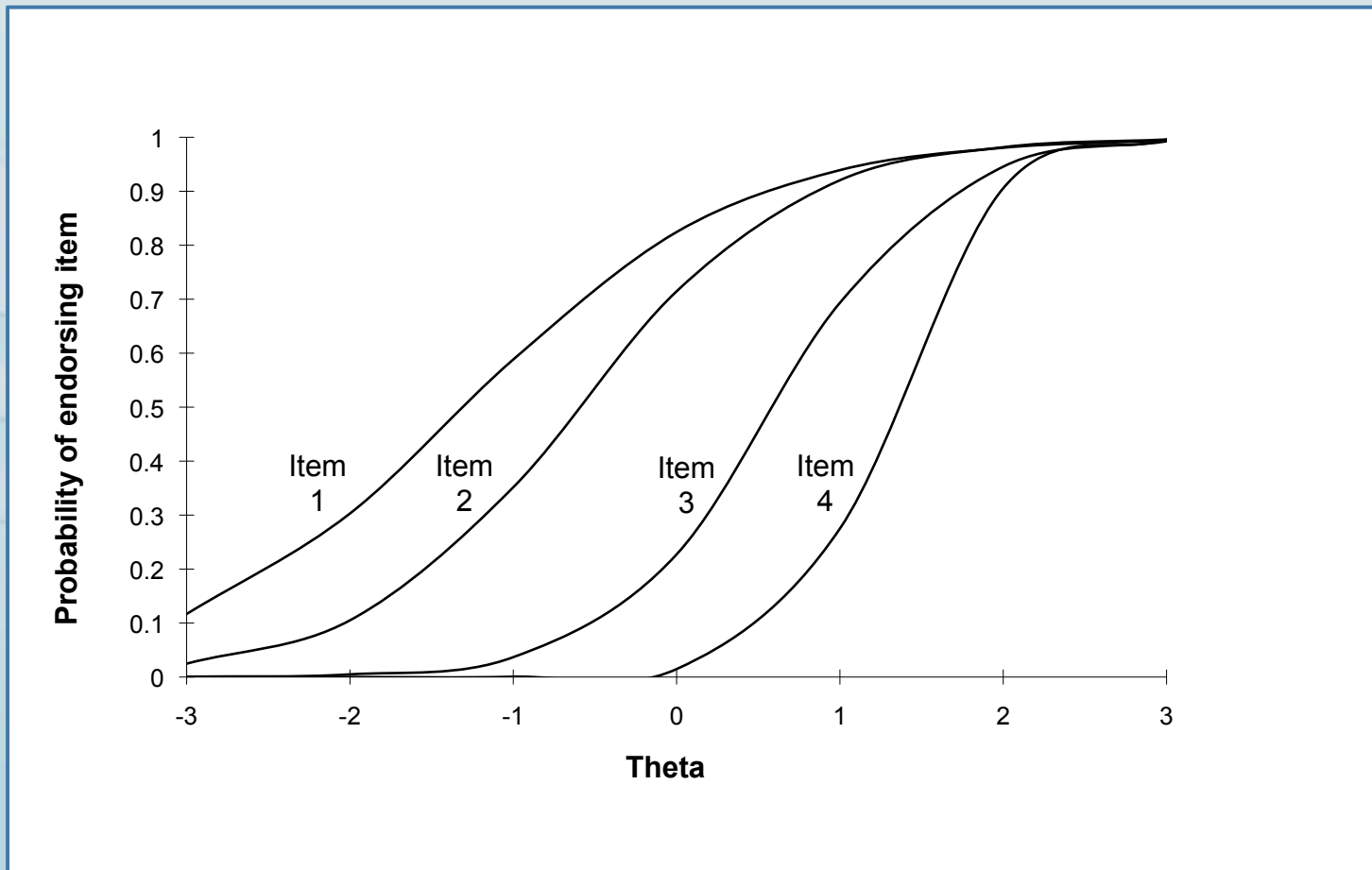
# The Rasch Model (cont.)

- Estimates both item and person parameters
- The distance between points are of equal intervals
- Unlike CTST-constructed scales, the items can range across the latent variable
- It meets the *invariance criterion* mentioned when referring to Thurstone's attitude measurement work above, namely:
  - The relationship between the items does not depend upon the persons who responded to them, and if other items are added the comparison between any two items remains invariant
  - The distance between any two individuals is invariant regardless of which items they are compared on
- The difficulty is in finding more than a small number of items to fit a Rasch model

# The Two Parameter (2PL) IRT Model

- Because it is so difficult to find items to fit 1-parameter models, many psychometric applications use 2 or 3 parameter models
- In the 2PL, the first parameter estimates the difficulty of the item (as in the 1PL model) and the second parameter estimates how well it discriminates along the theta scale
- An item that discriminates well will have a steep slope centered around its difficulty. A poorly discriminating item will have a flatter slope.

# A Hypothetical Example of a Four Item 2PL IRT Model



# Graded Response Model

- The IRT models considered thus far are only for dichotomous items
- There are several IRT models that will accommodate multiple response categories where one does not have to assume more than ordinality, including the Graded Response Model (GRM)
- The problem with both the 2PL and the GRM is that one loses the nice measurement characteristics associated with the Rasch model

# Steps for Constructing “Good Measures” Where Items Reflect Constructs

1. Define the concept as carefully as possible, specifying the domain of meaning
2. Use factor analysis to explore the dimensionality of the concept
3. After determining what appears to be dimensionality, do a confirmatory factor analysis
4. Estimate the internal consistency reliability of the measures constructed based on the CFA
5. Fit the items for each dimension to a Rasch model

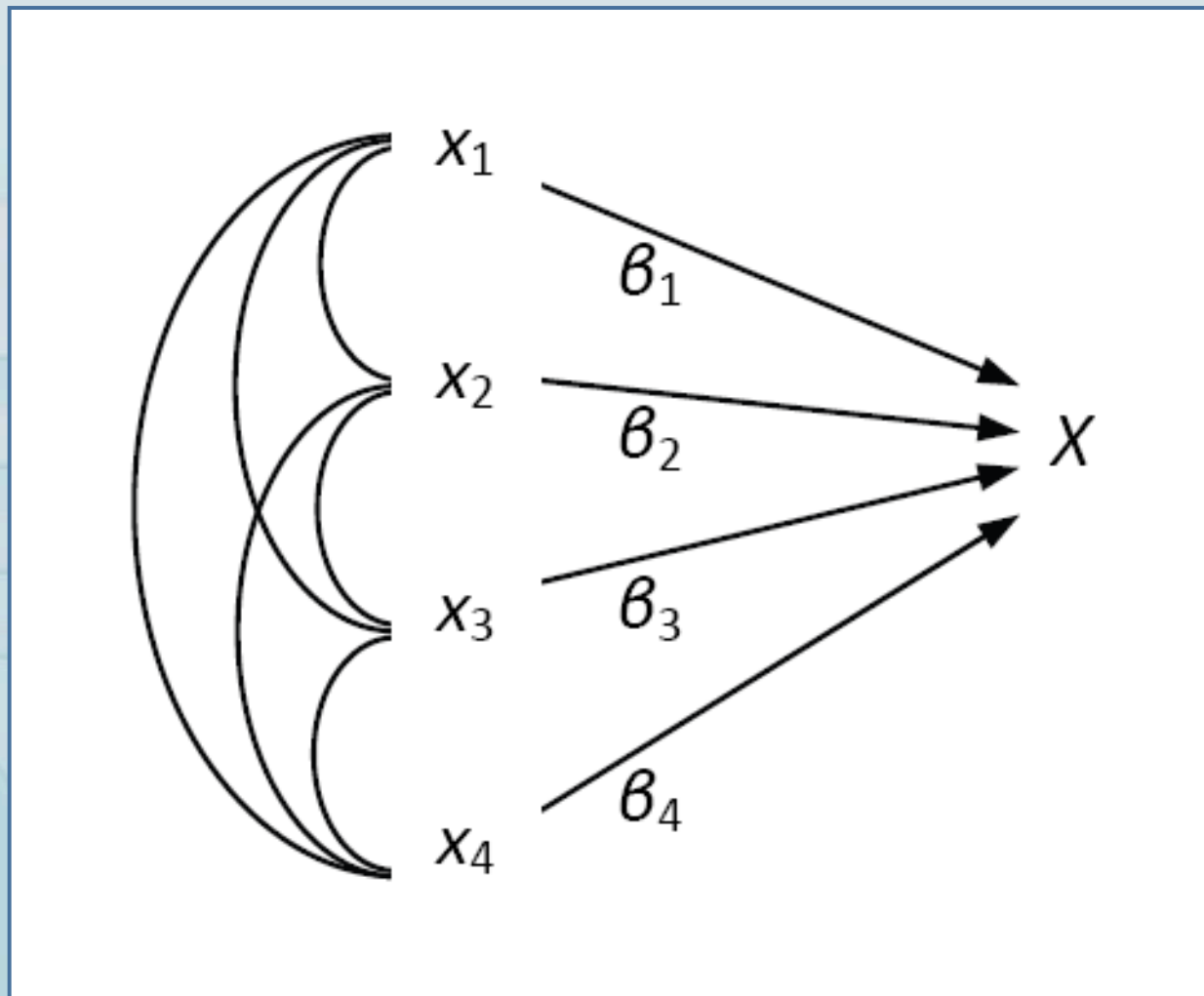
## Steps for Constructing “Good Measures” Where Items Reflect Constructs (cont.)

6. If the items will not fit a 1PL or Rasch model, fit them to a 2PL model
7. Ensure that parameter estimates are invariant for various subpopulations
8. Develop new items to bolster sparse areas on the latent dimensions

# Index Construction When Items Determine Rather than Reflect a Construct

- In sociology, economics, and policy research there are cases where the assumption is that the indicators define the construct rather than the other way around
- This is sometimes called a “formative” as opposed to a “reflective” model for index construction
- Examples include an SES index composed of education, income, and occupation, and the Consumer Price Index based on a market basket of goods and services

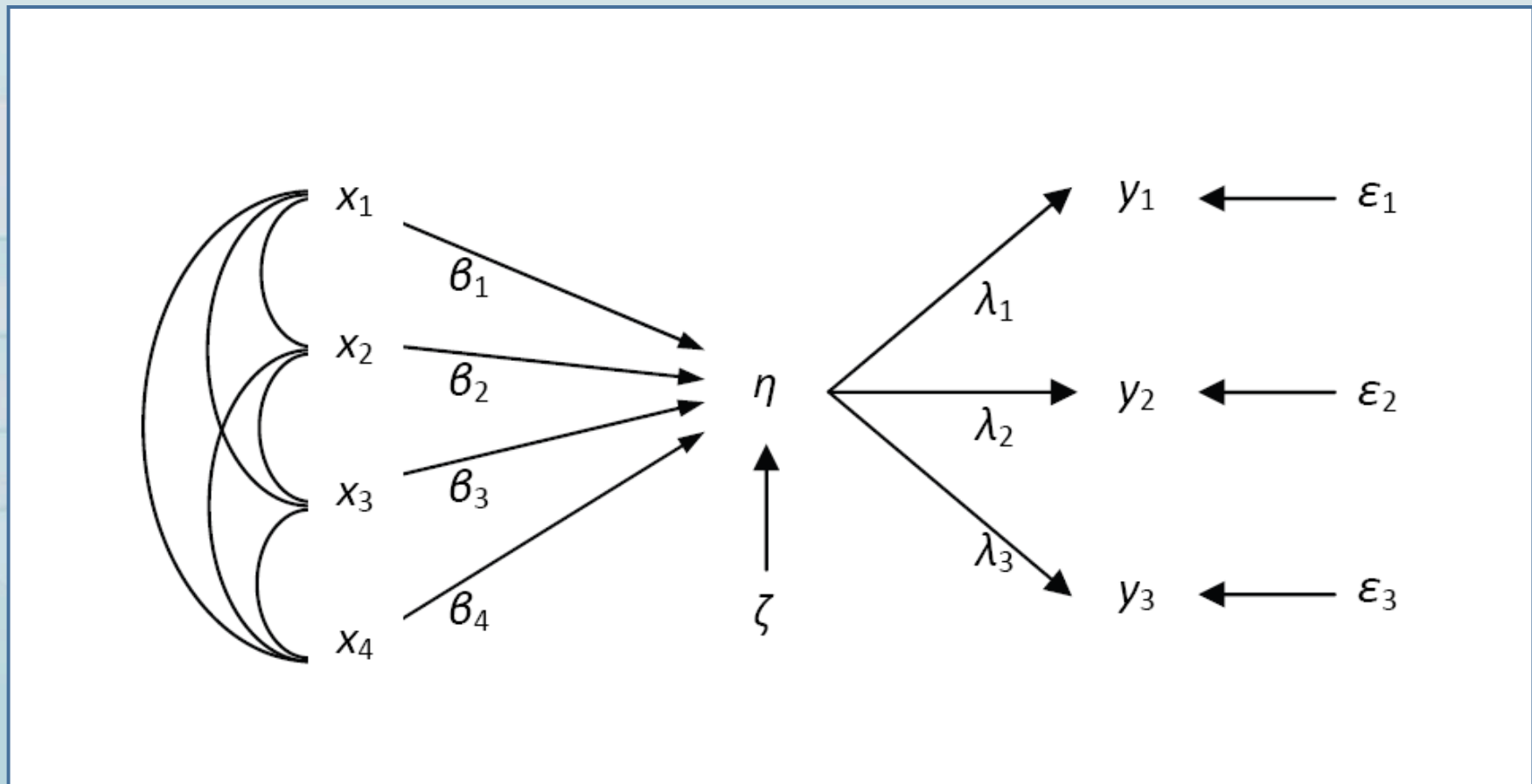
# A Schematic of a Concept Determined by Four Indicators



# Index Construction Where Indicators Determine the Concept

- Typically, the indicators are simply unit weighted
- But there could be cases where they are weighted based on theory, differential utilities, or other preferences (e.g., relative importance based on a community survey)
- One can estimate the weights of the indicators if there are multiple indicators and multiple causes (the MIMIC model)

# A Schematic Representation of a Multiple Indicator-Multiple Cause (MIMIC) Model



# MIMIC Models

- Allow one to estimate the  $\beta_i$  as well as the coefficients on the causal (right) side of the model
- Different specifications on the causal side of the model allow one to check for the stability of the  $\beta_i$
- Note that one does not actually construct an index with MIMIC models
- Bob Hauser showed the general utility of MIMIC models in a series of papers he did in the 1970s and 1980s

# The Role of Standards in Developing Common Metrics

- In the area of educational assessment, the National Assessment Governing Board set standards, called achievement levels, in various subject areas measured by NAEP
- These are set as cutpoints on the NAEP scale (0-500) for each grade and subject by a set of experts as what a student should know and be able to do if one is basic, proficient, or advanced

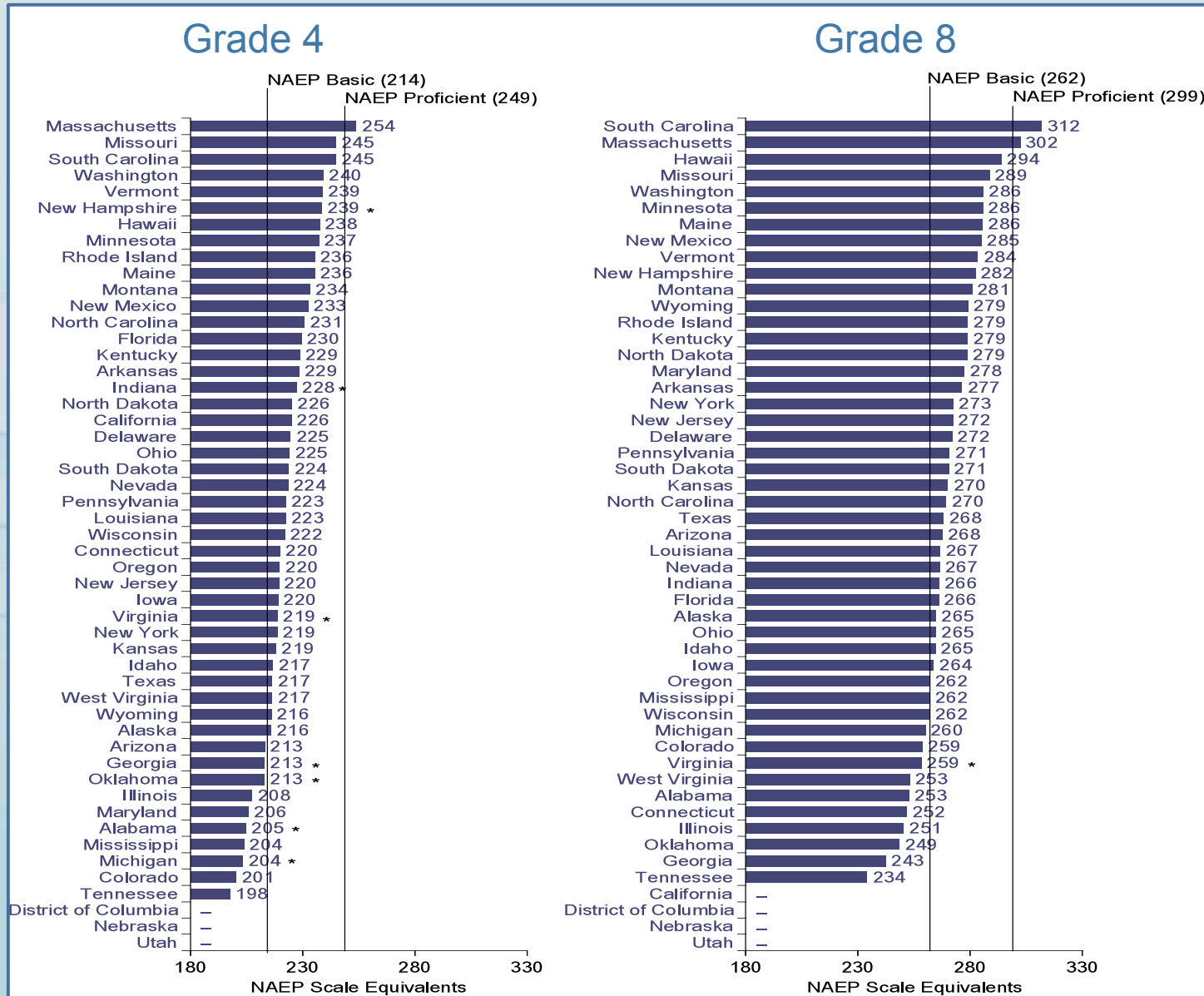
# The Role of Standards in Developing Common Metrics (cont.)

- In addition, states were required to set proficiency standards on their own assessments as part of NCLB
- These are used to see if states are making adequate yearly progress toward their goals
- A natural question is whether the states' standards for proficiency are the same as or even similar to one another's

# The Role of Standards in Developing Common Metrics (cont.)

- Comparing the states requires a common standard against which to compare the states
- States required to take NAEP as a requirement to receive Title I funds
- One is able to use the data to map state standards onto the NAEP scale
- There is tremendous variation in states' standards for grade 4 and 8 mathematics

# Examining Variation in States' Proficiency Standards in Mathematics: 2007



# The Role of Standards in Developing Common Metrics (cont.)

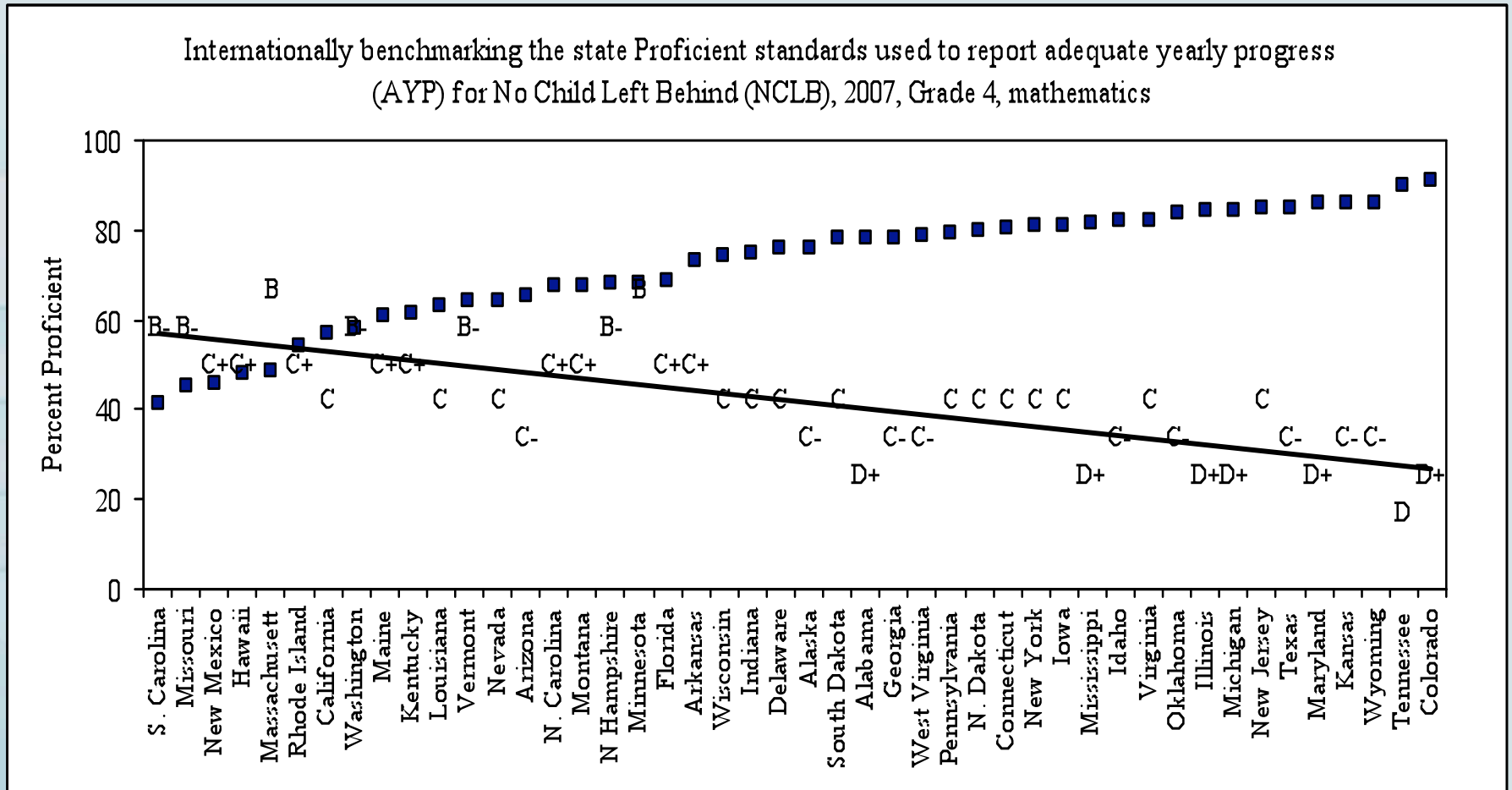
- Gary Phillips has taken these results and used them to benchmark how well the states' standards match international standards
- Five levels of performance on the Trends in International Science and Mathematics Study (TIMMS)
- They seem to comport well with grades of "A," "B," "C," etc.

# Using TIMMS Standards to Benchmark States' Standards

- Phillips took the NAEP mapping results shown in the figure above and projected them onto the TIMMS performance measures
- Clearly there is linking error in both the NAEP and the TIMMS projections, but the results seem informative nonetheless

# An Example of Benchmarking States' Standards against TIMSS Standards

Internationally benchmarking the state Proficient standards used to report adequate yearly progress (AYP) for No Child Left Behind (NCLB), 2007, Grade 4, mathematics



# Using TIMMS Standards to Benchmark State's Standards (cont.)

- The states with the highest standards are on the left side of the graph. Note that about half a dozen states set their standards at the B or B- level, but most are in the C range and about half a dozen are in the D range.
- The upper array in the figure is states' percent proficient on their state assessments. The data show clearly that those states that set the lowest standards had the highest percentage of students who were labeled as performing at the “proficient” level

# Mapping onto Common Standards

- Illustrates that even in the absence of perfect measures, there may be cases when arbitrary standards can be set against which judgments of performance can be made
- Worth exploring if there are other areas where such common metrics might have utility (e.g., risk of re-incarceration of prisoners eligible for parole)

# In Summary

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.
3. Science has a strong and central role to play in the development of standards.

## In Summary (cont.)

4. We have not 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. Our concepts lack strong axiomatic theories against which to evaluate our measurements and with which to generate good measures.

## In Summary (cont.)

7. Rasch and IRT models could be one answer.
8. Another might be using standards to generate common metrics, which, while not meeting the standards for real measurement, have utility nonetheless.
9. I hope I have generated some thoughts we might kick around in our various discussion periods as we seek to understand if and when common metrics are warranted.