TEST SCALING AND VALUE-ADDED MEASUREMENT

Dale Ballou

Department of Leadership,
Policy and Organizations
Vanderbilt University
Peabody College
Nashville, TN 37205-5721
dale.ballou@vanderbilt.edu

Abstract

Conventional value-added assessment requires that achievement be reported on an interval scale. While many metrics do not have this property, application of item response theory (IRT) is said to produce interval scales. However, it is difficult to confirm that the requisite conditions are met. Even when they are, the properties of the data that make a test IRT scalable may not be the properties we seek to represent in an achievement scale, as shown by the lack of surface plausibility of many scales resulting from the application of IRT. An alternative, ordinal data analysis, is presented. It is shown that value-added estimates are sensitive to the choice of ordinal methods over conventional techniques. Value-added practitioners should ask themselves whether they are so confident of the metric properties of these scales that they are willing to attribute differences to the superiority of the latter.

1. INTRODUCTION

Models currently used for value-added assessment of schools and teachers require that the scale on which achievement is measured be one of equal units: the five-point difference between scores of fifteen and twenty must represent the same gain as the five-point difference between twenty-five and thirty. If it does not, we will end up drawing meaningless conclusions about such matters as the average level of achievement, relative gains of different groups of students, etc., in that the truth of these conclusions will depend on arbitrary scaling decisions.

A scale that possesses this property is known as an interval scale. It is clear that a simple number-right score is not an interval scale of achievement when test questions are not of equal difficulty. The same is true of several popular methods of standardizing raw test scores that also fail to account for the difficulty of test items, such as percentiles, normal curve equivalents, or grade-level equivalents normed to a nationally representative sample.[1]

The search for measures of achievement that are independent of the particular items included on a test led to the development in the 1950s of item response theory (IRT). IRT is now used to score most of the best-known and most widely administered achievement tests today, such as the CBT/McGraw-Hill Terra Nova series, the Stanford Achievement Test (SAT), and the National Assessment of Educational Progress. IRT was regarded as a significant advance over earlier scaling methods for the following reasons: (1) the score of an examinee is not dependent on the difficulty of the items on the test, provided the test is not so easy that the examinee answers all items correctly or so hard that he or she misses them all; (2) an examinee's score is not dependent on the characteristics of the other students taking the same test; and (3) according to some psychometricians, an examinee's score is an interval scaled measure of ability.[2] This last claim makes IRT scaling particularly appealing to those practicing value-added assessment.

However, not all psychometricians share these views, and the literature contains many confusing and contradictory statements about the properties of IRT scales. Because many social scientists using test scores to evaluate educational institutions and policies have little or no training in measurement theory, the first objective of this article is to review the issues. The next section describes IRT. It is followed by a summary of the controversy over scale type in section 3. While under the right conditions IRT yields interval scaled measures

[1] These last methods also have the disadvantage of being dependent on the distribution of ability in the tested population or must rely on arbitrary assumptions about this distribution.

[2] The psychometric literature uses the term *ability*. Other social scientists might prefer *achievement*. It represents the student's mastery of the domain of the test and should not be confused with innate ability as opposed to knowledge and skills acquired through education.

of achievement, these conditions are difficult to verify. Moreover, IRT scales are often at odds with common sense notions about the effects of schooling and the distribution of ability as students advance through school. I argue in section 4 that we are rightly suspicious of IRT scales when we see such results. Section 5 takes up the implications for value-added assessment, with particular attention to methods of ordinal data analysis. Concluding remarks appear in section 6.

2. ITEM RESPONSE THEORY AND ABILITY SCALES

In IRT, the probability that student i correctly answers test item j is a function of an examinee trait (conventionally termed ability) and one or more item parameters.[3] Thus

$$P_{ij} = \mathrm{Prob}[h(\theta_i, \delta_j) > u_{ij}] = F(\theta_i, \delta_j), \tag{1}$$

where θ_i is ability of person i, δ_j is a characteristic (possibly vector valued) of item j, and u_{ij} is an idiosyncratic person-item interaction, as a result of which individuals of the same level of ability need not answer a given item alike. The u_{ij} are taken to be independent and identically distributed random errors. The function h expresses how ability and item parameters interact. F is derived from h and assumptions about the distribution of u_{ij}. Common assumptions are that the u_{ij} are normal or logistic. θ_i and δ_j are estimated by maximum likelihood methods.

When there is a single item parameter, the assumption that the u_{ij} are logistic gives rise to the one-parameter logistic model (also known as the Rasch model):

$$P_{ij} = [1 + \exp(-D(\theta_i - \delta_j))]^{-1}, \tag{2a}$$

in which D is an arbitrary scaling parameter, invariant over items, that can be chosen by the practitioner. (If $D = 1.7$ it makes essentially no difference whether the model is estimated as a logistic or normal ogive model. Alternatively, D is set to 1.) The estimate of θ is known as the *scale score*. The scale score is the principal measure of performance on the exam, although other measures derived from it, such as percentile ranks, may also be reported.[4] The item parameter is conventionally termed difficulty. The functional form of equation 2a implies that it is measured on the same scale as ability.

[3] There are many lucid expositions of IRT, including Hambleton and Swaminathan (1985) and Hambleton, Swaminathan, and Rogers (1991).

[4] Another measure is the so-called true score, which is simply the expected number right = $\Sigma P_{ij}(\theta)$ on a test comprising the universe of items. This suffers from the usual defect of a number-right score: the metric depends on the composition of the universe.

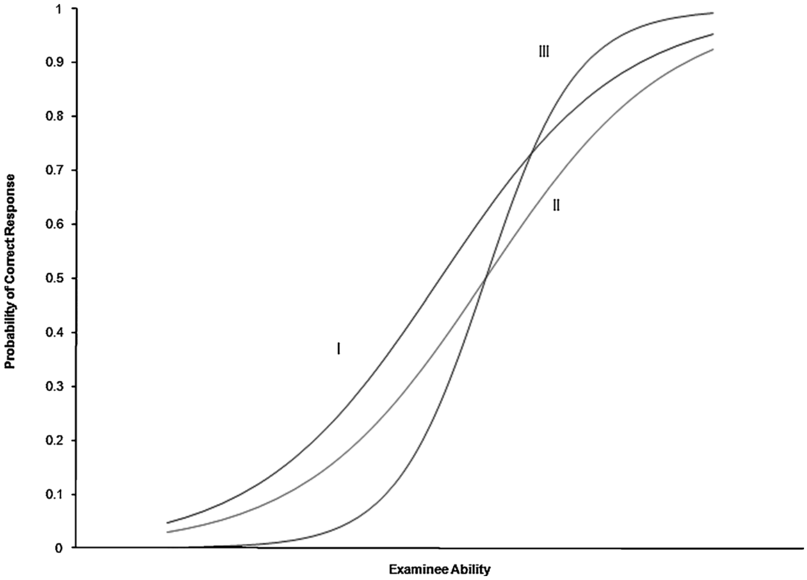


Figure 1. Item Characteristic Curves

More elaborate models introduce additional item parameters, as in the two- and three-parameter logistic models:

$$P_{ij} = [1 + \exp(-\alpha_j(\theta_i - \delta_j))]^{-1} \tag{2b}$$

$$P_{ij} = c_j + (1 - c_j)[1 + \exp(-\alpha_j(\theta_i - \delta_j))]^{-1}. \tag{2c}$$

Equation 2b contains a second item parameter, α_j, known as the item discrimination parameter because it enters the derivative of P_{ij} with respect to θ_i (thereby determining how well the item discriminates between examinees of different abilities). In both equations 2a and 2b, the limit of P_{ij} as $\theta_i \rightarrow -\infty$ is 0. This is not appropriate for tests where a student who knows nothing at all can answer an item correctly by guessing. Accordingly, c_j allows for a nonzero asymptote and is conventionally termed the guessing parameter.

The plot of P_{ij} against θ_i is known as the item characteristic curve (ICC). Three item characteristic curves using the two-parameter logistic IRT model are depicted in figure 1. Curve II differs from curve I by an increase in the difficulty parameter, holding constant item discrimination. Curve III corresponds to an item with the same level of difficulty as II but a doubling of the discrimination parameter. All three curves have a lower asymptote of 0. Observe that all three curves are steepest where $P_{ij} = .5$ and $\theta = \delta$.[5] At this point the slope equals α_j.

[5] Because these models are additive in θ and δ, ability and difficulty are expressed on the same scale.

In the IRT models in equations 2a–2c, θ and δ are underidentified. Modification of each by an additive constant obviously leaves P_{ij} unchanged. Multiplicative constants can be offset by changes to (or absorbed in) the discrimination parameter. Because an additive constant corresponds to a change in the origin of the scale while a multiplicative constant represents a change in the scale's units (e.g., from meters to yards), many psychometricians have concluded that θ and δ are measured on interval scales, which are characterized by the same conventional choice of origin and unit (see section 3). However, this opinion is by no means universal, nor is it firmly held. Many psychometricians, including some who state that these are interval scales, also regard the θ scale as arbitrary. Others caution that ability scales should be accorded no more than ordinal significance. These conclusions appear to be derived from the following considerations: (1) Because ability is a latent trait, it is impossible to verify by physical means that all one-unit increments in θ represent the same increase in ability. This (the argument goes) confers an inherent indeterminacy on the scaling of any latent trait.[6] (2) Replacing θ with a monotonic transformation $g(\theta)$ while making offsetting changes to the function F yields a model that fits the data just as well.[7] Thus the θ scale rests on arbitrary assumptions regarding functional form.[8] (3) The notion that θ measures the amount of something misconceives the entire enterprise. There is no single trait (call it ability, achievement, or what have you) to be quantified

[6] "When the characteristic to be measured cannot be directly observed, claims of equal-interval properties with respect to that characteristic are not testable and are therefore meaningless" (Zwick 1992, p. 209).

[7] An example is Lord's (1975) transformation of the θ scale to eliminate correlation between item difficulty and item discrimination, in which θ_i was replaced by the regression of α_j on δ_j and higher powers of δ_j evaluated at $\delta_j = \theta_i$. The result was a modification of the three-parameter logistic model:

$$P^*_{ij} = F^*_{ij}(\omega) = c_j + (1 - c_j)[1 + \exp(-\alpha_j(\omega^{-1}(\omega(\theta_i))) - \delta_j)]^{-1},$$

where $\omega(\theta)$ is proportional to $-.27 + 1.1694\theta + .2252\theta^2 + .0286\theta^3$. As a nonlinear transformation of the original ability scale, the ω scale differs from the θ scale by more than a change of origin and unit. Lord (1975, p. 205) saw this as no drawback: "There seems to be no firm basis for preferring the θ scale to the ω scale for measuring ability."

[8] "A long-standing source of dissatisfaction with number-right and percent-correct scores is that their distribution depends on item reliabilities and difficulties. Radically different distributions of true scores (expected percents correct) can be obtained for the same sample of examinees when they take different tests. The obvious restriction of such scores to their ordinal properties casts doubt upon their use for problems that require interval scale measurement, such as comparing individuals' gains. . . . Item response theory appeared to offer a general solution, since the same underlying θ scale could explain the different true-score distributions corresponding to any subset of items from a domain. But this line of reasoning runs from model to data, not from data to model as must be done in practice. Suppose that a given dataset can be explained in terms of a unidimensional IRT model with response curves of the form $F_j(\theta)$. Corresponding to any continuous, strictly increasing function h there is an alternative model with curves $F^*_j(\theta) = F_j(h(\theta))$ that fits the data in precisely the same manner (Lord, 1975). That a particular IRT model fits a dataset, therefore, is not sufficient grounds to claim scale properties stronger than ordinal" (Mislevy 1987, p. 248).

by this or any other means.[9] As such, the question of scale type is essentially meaningless: there are just various models for reducing the dimensionality of the data, some more convenient than others.[10] I will refer to this view of IRT scales in the ensuing discussion as the *operational perspective*.

To summarize, there are some psychometricians who consider θ to be interval scaled, others who think it is ordinal, still others who regard the choice of scale as arbitrary, even if it is an interval scale, and finally some who are unsure what it is. Clearly it is disconcerting to find this divergence of views on a question of fundamental importance to value-added assessment. Is the IRT ability trait measured on an interval scale or not? Indeed, how does one tell?

3. IRT MODELS AND SCALE TYPE

In the natural sciences, measurement is the assignment of numbers to phenomena in such a way that relations among the numbers represent empirically given relations among the phenomena. Technically, there is a homomorphism between the empirical relations among objects in the world and numerical relations on the scale. One such relation is order: if objects can be ordered with respect to the amount of some attribute, that order needs to be reflected in increasing (or decreasing) scale values. However, order is not the only attribute to matter. If objects A and B can be combined (or concatenated) in such a way that in combination they possess the same amount of some attribute (length, mass) as object C, a scale for that attribute needs to reflect the results of that operation. Thus, in an additive representation, the scaled value of the attribute in A plus the scaled value of the attribute in B equals the scaled value assigned to C.

The importance of convention relative to empirical phenomena turns out to be the key to scale type. At one end of the spectrum are scales in which there

[9] "The claim that a particular unidimensional scaling method is right must be based on the assumption that achievement is unidimensional, that it can be linearly ordered, and that students can be located in this linear ordering independently of performance on a particular achievement test. However, no one has succeeded in identifying or defining a linearly ordered psychological trait in educational achievement, and no one has demonstrated that a particular measurement scale is linearly related to such a trait. A serious obstacle to the establishment of truly (externally verifiable) equal-interval achievement scales is the fact that achievement is multidimensional and qualitatively changing. The nature of what is being learned is constantly being modified. Use of an objective-based approach to achievement highlights the difficulties in hypothesizing and verifying a continuous achievement trait. The student is learning the names of letters of the alphabet one day, associating sounds to those letters another day, and attaching meaning to groups of letters a third day. How can such qualitative changes be hypothesized to fall so many units apart on one particular trait?" (Yen 1986, p. 311–12).

[10] "It is important for educators and test developers to acknowledge that until the achievement traits are much more adequately defined, it is not possible to develop measurement scales that are linearly related to such traits. In fact, it appears impossible to provide such trait definition. Test users are therefore left to use other criteria to choose the best scale for a particular application; choosing *the* right scale is not an option" (Yen 1986, p. 314).

is no role for convention; at the other are scales that are entirely conventional. Scales of the first type are called *absolute*. An example is counting: one is not free to change units if the information to be conveyed by the scale is the number of discrete objects under observation.[11] At the other end of the spectrum are *nominal* scales, where the number assigned to an object is no more than a label and the information conveyed could just as well be represented by a nonnumerical symbol. Scales for most physical quantities, such as length and mass, have a degree of freedom for the conventional choice of a unit. Such scales are known as *ratio* scales because the ratio of two lengths or two weights is invariant to the choice of unit: ratios are convention free. If there is no natural zero, so the origin of a scale is also determined conventionally, ratios are no longer convention-free magnitudes. However, ratios of intervals are invariant under change of unit and change of origin. Such scales are therefore known as *interval* scales: they are characterized by two degrees of freedom. Between interval scales and nominal scales lie *ordinal* scales. Any increasing function of an ordinal scale conveys the same information.[12]

With respect to psychological variables, there is less agreement about the nature of measurement. There appear to be three prevalent views (Hand 1996). In one view (sometimes called *classical* measurement), it is simply assumed that the psychological variable of interest exists and that there is a ratio (or at least interval) scale on which it is measurable. The task of measurement is to discover those values. This appears to be the view of some psychometricians with respect to IRT scales. In the second view, a psychological variable exists only by virtue of its presence in some model. The model effectively defines the variable and, when the model is fit to data, provides a means of determining the scaled value of the variable. This has been called *operational* measurement and is compatible with the operational perspective on IRT scales described above, in which IRT models are devices for reducing the dimensionality of data. Two different models may both contain the term *ability*. There is no basis for deciding between them on grounds that one better represents “true ability”—each

[11] It should be clear that scale type is as much a matter of how numbers are interpreted as of formal relations among the items being scaled. For example, considered from a purely technical standpoint, number right on a test can be regarded as an absolute scale. Like any scale that counts items (e.g., the score of a football game), number right does not admit of a change of units without a loss of information. However, when test items are not of equal difficulty, number right cannot be regarded as a measure of achievement that generalizes beyond performance on the particular test in question, and as such is not an absolute scale, a ratio scale, or an interval scale. Except in special circumstances it is not even an ordinal scale.

[12] Scale type is therefore closely related to the notion of an *admissible* scale transformation (Stevens 1946). An admissible transformation preserves the empirical information in the original scale by altering only those elements that are purely conventional. For ratio scales, these are the similarity transformations, corresponding to a choice of unit (for example, substituting centimeters for meters). In the case of interval scales, the admissible transformations consist of affine transformations, $g = \alpha \cdot f(a) + \beta$.

is just more or less useful for the purposes to which they are put. Finally, there is a third view that holds that measurement of psychological variables is essentially the same as measurement in the physical sciences—the view sketched at the start of this section, known as *representational* measurement. Psychological attributes are postulated to explain empirically given relationships (such as the pattern of examinees' responses to the items on a test). It is the structure of those relationships that determines the properties of scales that measure these attributes. Some relationships are so lacking in structure that the attributes may not be scalable at all—the most we can do is to name them. In other cases it may be possible to say that there is a larger quantity of some attribute in one person than another, supporting ordinal scaling. In still other cases there may be sufficient structure to permit interval scaled comparisons: the difference in the amount of the attribute between A and B equals the difference between C and D.

We can aspire to resolve disputes about scale type only if the third of these views is correct. Classical measurement simply assumes an answer, whereas the question of scale type is either meaningless or of no importance in operational measurement. Virtually all discussions of scale type nod in the direction of representational measurement by invoking Stevens's classification of admissible transformations. To the extent that a case can be made that IRT scales are interval scales, it has to be made in terms of representational measurement theory.

The argument that θ and δ are interval scaled is found in the analysis of conjoint additive structures. We begin by assuming that the P_{ij} are given—or, more precisely, that we are given the equivalence classes comprising examinees and items with the same P_{ij}. Let $A_1 = \{a, b, \ldots\}$ denote the set of examinees and $A_2 = \{p, q, \ldots\}$ the set of test items, and let $\gtrsim$ represent the order induced on $A_1 \times A_2$ by P_{ij}. That is, $(a,p) \gtrsim (b,q)$ if $P_{ap} \geq P_{bq}$. The sets A_1 and A_2 and the relation $\gtrsim$ are known as an *empirical relational system.* If several exacting conditions are met, requiring that the relations between A_1 and A_2 exhibit still more structure than the ordering of equivalence classes, the resulting empirical relational system is called a *conjoint additive structure* and the following can be proved: (1) there are functions φ_1 and φ_2 mapping the elements of these sets into the real numbers (i.e., examinees and items can be scaled); (2) the relation ordering examinee-item pairs can be represented by an additive function of their scaled values; that is, $(a,p) \gtrsim (b,q) \Leftrightarrow \varphi_1(a) + \varphi_2(p) \geq \varphi_1(b) + \varphi_2(q)$; (3) only affine transformations of φ_1 and φ_2 preserve this representation. These transformations correspond to a change of origin and a change of units; hence φ_1 and φ_2 are interval scales (Krantz et al. 1971).

There are two critical steps in the proof. First, from the relation ordering examinee-item pairs we must be able to derive relations ordering the elements

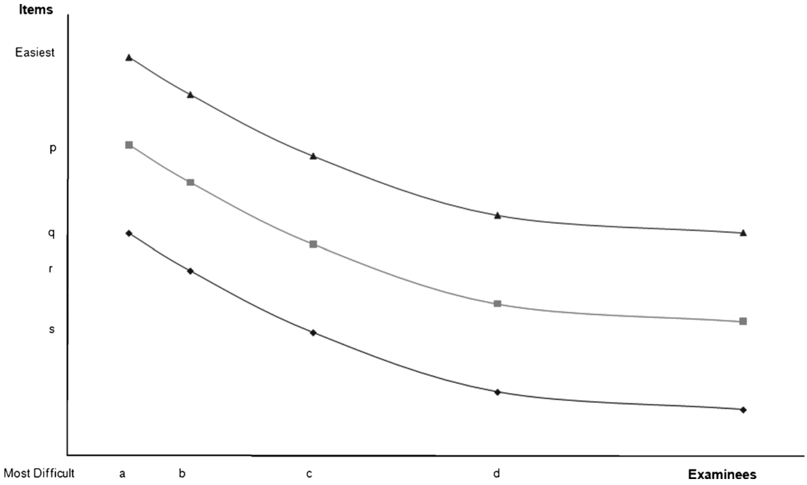


Figure 2. Equiprobability Curves, Conjoint Additive Structure before Scaling

of each set separately. For this we require a monotonicity condition (also known as independence): if $P_{aq} > P_{bq}$ for some item q, then $P_{ar} > P_{br}$ for all r. An analogous condition holds for items: if item q is harder for one examinee to answer, it is harder for all examinees. Monotonicity establishes orderings of A_1 and A_2 separately. Without it, not even ordinal scales could be established for examinees and items.

To obtain interval scales we need further conditions, as illustrated in figure 2, which depicts three "indifference curves" (literally, isoprobability curves) over examinees and items: that is, three equivalence classes determined by the P_{ij}. Examinees and items are arrayed along their respective axes according to the induced order on each set, but no significance should be attached to the distance between a pair of examinees or a pair of items: the axes are unscaled apart from the ordering of examinees and items. However, note that there is an additional structure to the equivalence classes in figure 2. They exhibit a property known as equal spacing: as we move over one examinee and down one item, we remain on the same indifference curve. Equal spacing is probably the simplest of all conjoint additive structures, but it is not a necessary condition for conjoint additivity.

To derive an equal interval metric, we establish that the distance between one pair of examinees is equal to the distance between another pair by relating both to a common interval on the item axis. As shown in figure 3, the interval between examinees a and b can be said to equal the interval between examinees c and d in that it takes the same increase on the item scale (from q to r) to offset both. Thus the item interval qr functions as a common benchmark for defining a sequence of equal-unit intervals on the examinee axis. The scaled

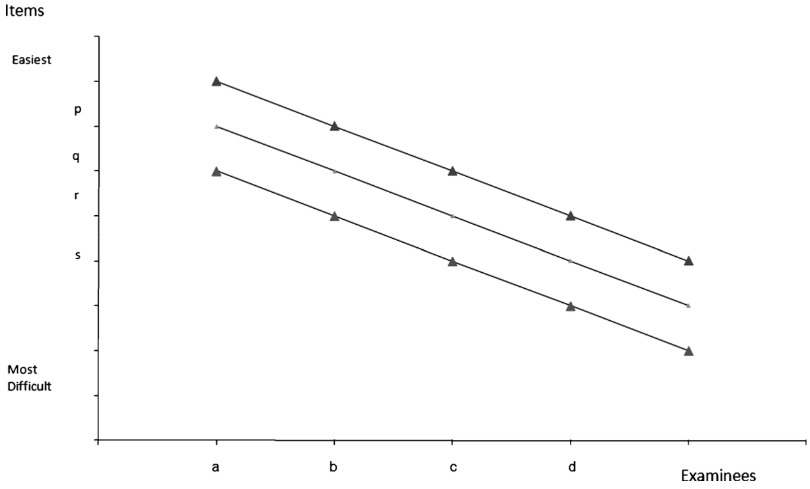


Figure 3. Equiprobability Curves, Conjoint Additive Structure after Scaling

value of any individual i is then obtained by counting the number of such intervals in a sequence from some arbitrarily chosen origin to person i: ab, bc, cd. . . . This scaling of the heretofore unscaled axes renders the equiprobability contours linear (and, of course, parallel).

In order that the conclusion ab = bc = cd = . . . not be contradicted by other relations in the data, it must be the case that the same conclusion follows no matter which interval on the item axis is selected as the benchmark. A like condition must hold for intervals on the examinee axis to serve as a metric for items. In addition to these consistency conditions, other technical conditions must be met when equal spacing does not obtain.

To summarize, conjoint measurement scales a sequence of intervals in one factor by using differences of the complementary factor as a metric. By moving across indifference curves, these sequences can be concatenated to measure the difference between any two examinees (or items) with respect to the latent factor they embody. Because the measurement of intervals reduces to counting steps in a sequence, there is an essential additivity to this empirical structure, on the basis of which we obtain additive representations of examinee and item traits.

Conjoint additivity uses only the P_{ij} equivalence classes and not the values of P_{ij} to determine the φ_1 and φ_2 scales. The final step from conjoint additivity to IRT requires a positive increasing transformation F from the positive reals to the interval [0, 1], such that $F(\varphi_1(a) + \varphi_2(p)) = P_{ap}$. Because any increasing transformation of $\varphi_1(a) + \varphi_2(p)$ preserves the representation of $A_1 \times A_2$ by φ_1 and φ_2, it is clear that a function F satisfying this condition can be found.

Of the three IRT models, only the one-parameter model is consistent with this simple conjoint additive structure. The two- and three-parameter IRT models, in which item discrimination varies, violate the monotonicity assumption. This is easily seen with the aid of figure 1. The ICC labeled II represents an item with discrimination parameter = 1, while the curve labeled III has a discrimination of 2. The ICC crosses where θ equals the common value of the difficulty parameter. An individual whose θ lies to the left of this intersection finds item II more difficult than item III; an individual whose θ is to the right of the intersection ranks the items the other way around. Because different individuals produce inconsistent rankings of items, the ranking of examinee-item pairs on the basis of P_{ij} does not yield orderings of examinees and items, and the derivation of the φ scales breaks down. Indeed, it is alleged that because only the one-parameter IRT model produces a consistent ordering of items for all examinees (ICCs in the Rasch model never cross), only the θ in a one-parameter model can be considered an interval scaled measure of ability (Wright 1999).

This goes too far. A straightforward modification of conjoint additivity accommodates structures with a third factor that enters multiplicatively, such as the IRT discrimination parameter. The relevant theorem appears in Krantz et al. (1971, p. 348). The empirical relations between examinees and items are termed a *polynomial conjoint structure.* The extra conditions on examinees and items ensure that we can obtain separate representations by discrimination classes in the manner just described. When these conditions are met, we find that there are functions φ_1, φ_2, and φ_3 such that $(a,p) \gtrsim (b,q) \Leftrightarrow \varphi_3(p)[\varphi_1(a) + \varphi_2(p)] \geq \varphi_3(q)[\varphi_1(b) + \varphi_2(q)]$. φ_1 and φ_2 are unique to linear transformations, and φ_3 is unique to a similarity transformation. Obviously this representation has the structure of the two-parameter IRT model.[13]

[13] This representation does not include anything that corresponds to the guessing parameter in the three-parameter IRT. While I have not seen an analysis of such a conjoint measurement structure, extending polynomial conjoint measurement to include the three-parameter IRT model would proceed similarly to the extension of conjoint additivity to cover the two-parameter IRT model. For sets of items with the same discrimination parameter, the two-parameter IRT model reduces to the one-parameter model. Proof of scale properties for the two-parameter model involves selecting one such set of items and scaling examinees with respect to it. Because the choice of items is arbitrary, the resulting scale is unique only up to the change of units that would result from the selection of an alternative set with a different value of the discrimination parameter. (For details, see Krantz et al. 1971, chapter 7.) Incorporating a guessing parameter in the structure would involve following the same logic. Examinees would be scaled using a subset of the data (i.e., for a particular choice of discrimination and guessing parameters). As with the polynomial conjoint structure, the fact that another choice could have been made introduces a degree of freedom into the representation, though in the case of the guessing parameter this degree of freedom affects only the function mapping $(\varphi_3(q)[\varphi_1(b) + \varphi_2(q)])$ to P_{ij} and not the relations among φ_1, φ_2, and φ_3. The properties of the φ_1 and φ_2 would therefore be precisely those established for the polynomial conjoint structure, inasmuch as these properties depend on relations between examinees and items and not on the function mapping equivalence classes to numerical values of P_{ij}.

To summarize: If the empirical relational system is one of conjoint additivity and the function F correctly specifies the relationship between $\theta_I - \delta_i$ and the P_{ij}, the IRT measure of latent ability, θ, and the IRT difficulty parameter, δ, are interval scaled variables. If the empirical relational system is a polynomial conjoint structure and F is again correct (e.g., the two-parameter logistic IRT model, or possibly the three-parameter model, if extra conditions entailing a lower asymptote on P_{ij} are met), θ and δ are again interval scaled.

Notwithstanding the fact that these propositions were proved in the 1960s, one continues to find a wide range of opinions about the properties of IRT scales in the psychometric community (as we have seen). At least some of that diversity of opinion is due to the following three misconceptions about scales: (1) because arbitrary monotonic transformations of θ and δ can be shown to fit the data equally well, θ and δ cannot be interval scaled. At best they are ordinal variables; (2) because θ is interval scaled, no scale of achievement related to θ by anything other than an affine transformation can be an interval scale. In particular, if $\psi = g(\theta)$, where g is monotonic but not affine, the ψ scale is ordinal; and (3) using an IRT model (or specifically the Rasch model) to scale a test produces an interval scale of ability. Each of these beliefs is wrong. Before quitting this discussion of representational measurement theory, we need to understand why.

The first of these views derives from Stevens's (1946) stress on the role of admissible transformations in determining scale type. The problem is that Stevens's formulation of the matter fails to make clear just what makes a transformation "admissible." There is a sense that information must not be lost when a scale is transformed—but precisely what information? All transformations rest on the implicit assumption that there is something we can alter freely—something, in other words, that is not "information," at least not information we care about. Stevens's criterion for determining scale type remains empty until this something is specified.

Misconception (1) rests on the following argument: we can replace θ with $g(\theta)$ for arbitrary monotonic function g and still fit the data (the P_{ij}) equally well, provided we make an offsetting change to the function F relating θ and δ to P. An example appeared in footnote 7, in Lord's (1975) modification of the three-parameter logistic model:

$$P^*_{ij} = F^*_{ij}(\omega) = c_j + (1 - c_j)[1 + \exp(-\alpha_j(\omega^{-1}(\omega(\theta_i))) - \delta_j)]^{-1}.$$

Because this model fits the data as well as the original three-parameter logistic model (as it must, being mathematically equivalent), it is argued that θ and $\omega(\theta)$ contain the same information about ability. Because the ω function is not affine but an arbitrary monotonic function, the conclusion is drawn that θ and ω must both be ordinal scales.

The flaw in this argument is the supposition that the only information that matters is the order over equivalence classes of examinees and items induced by the P_{ij}. But the proof from conjoint additivity does not conclude that θ and δ are interval scales merely because these mappings from examinees and items to real numbers preserve order among the P_{ij}. The empirical relational system between classes of examinees and items contains more structure than the ordering of equivalence classes. It is this additional structure, as illustrated by the example of the equally spaced conjoint structure above, that underlies the claim that certain mappings from examinees and items are interval scales. An arbitrary monotonic transformation of these scales no longer reflects the relations holding among the scaled items and examinees shown in figure 3. Such a transformation loses the information that the distance between examinees a and b equals the distance between b and c in the sense that both offset the same substitution of one pair of test items for another. Preserving that information restricts the class of admissible transformations to affine functions.

The preceding comment notwithstanding, it does not follow that there cannot be another way of assigning numbers to examinees, on the basis of some other property, producing a new scale $\xi = h(\theta)$ for a non-affine function h that can also be regarded as an interval scaled measure of ability. Clearly ξ and θ cannot be interval scales of the same property of examinees. (In the language of representational measurement theory, they cannot represent the same empirical relational system.) That we might have reason to regard both as measures of achievement is due to the vagueness and imprecision with which the term *achievement* is used, not just in ordinary discourse but in social science research.

I illustrate with an example from economics. Consider a firm that employs thirty workers on thirty machines. Workers are rotated among machines on a daily basis. The only information the firm has on the productivity of either factor of production is the daily output of each worker-machine combination. Suppose, for purposes of rewarding employees or scheduling machines for replacement, the firm decides it needs measures of the productivity of individual workers and machines. In other words, it wants to scale these heretofore unscaled entities. (The parallel with testing should be obvious.) A measurement theorist is called in who observes that the data satisfy the conditions of a conjoint additive structure, inasmuch as workers and machines can be scaled such that the resulting isoquants in worker-machine space are linear and parallel. The measurement theorist confidently announces that these scaled values represent worker and machine productivity, up to an arbitrary choice of unit of measurement and origin.

Some time later the firm calls in a production engineer, who prowls around the shop floor with a stopwatch for a week and then reports the following discovery: the output of a worker-machine pair is a simple function of the downtime of the machine (beyond a worker's control) and the amount of time the worker is goofing off. That is, output $Q_{wm} = \pi_w \pi_m$, where π_w is the proportion of time worker w is attending to his work and π_m is the proportion of time machine m is running properly. There is no difference between workers otherwise: during their time on task with a functioning machine, all are equally productive. The engineer therefore proposes π_w as the natural (and ratio scaled) measure of employee productivity.

It is then noted that the engineer's productivity measure is not a linear transformation of the measurement theorist's measure in that the two variables differ by more than a choice of origin and unit. (Indeed, the measurement theorist's $\theta_w = \ln \pi_w$.) Yet each expert swears that his measure has at least interval scale properties, and each is right. The two measures capture different properties of the relations between workers and machines. The engineer doesn't care that his metric ignores the information in the conjoint additive structure because he relies on an alternative empirical relational system (one that includes the position of hands on a stopwatch) to scale the entities on the shop floor. Both experts have produced interval scales, and it is up to the firm to decide which scale captures properties of the relations between workers and machines that most matter to it.

This example shows why the second of the three misconceptions cited above is false. The parallel with testing is maintained if there is some other empirical relational system into which examinees and test items fit, affording an alternative metric. Suggestive examples are found in the education production function literature, in the form of back-of-the-envelope calculations of the benefits of various educational interventions (e.g., the dollar value of higher achievement associated with smaller class sizes). Whether this can be done with sufficient rigor to provide an interval scale for measuring student achievement (and whether it is desirable to construct one along these lines, if feasible) is a question to which I return in section 5.

Finally, the notion that the use of an IRT model (or a particular IRT model, such as the Rasch model) confers interval scale status on the resulting θ places the cart before the horse: it overlooks the requirement that the empirical relational system be a conjoint additive or polynomial conjoint structure. Applying an IRT model willy-nilly to achievement test data does not of itself confer any particular properties on the scale score metric.[14]

[14] Though an obvious point, this is not always appreciated by researchers. Consider the following statement in a report of the Consortium on Chicago School Research. Having rescaled the Iowa Test of Basic Skills using a Rasch IRT model, the authors claim to have produced a metric with interval scale properties: "A third major advantage of the Rasch equating is that, in theory, it produces a 'linear test score metric.' This is an important prerequisite in studies of quantitative

4. ACHIEVEMENT SCALES IN PRACTICE

To summarize the argument of the preceding section, under stringent conditions an IRT measure of ability can be shown to be an interval scaled variable. But there are two important caveats. First, do the data meet these stringent conditions? Second, might there be some other set of relations holding between examinees and test items that provides an alternative, and perhaps more satisfactory, basis for constructing an achievement measure with interval (or even ratio) scale properties? I take up these questions in turn.

It is highly unlikely that real test data meet the exacting definition of a conjoint structure. At best they come close, though it is not easy to say how close. The theory places restrictions on the equivalence classes of $A_1 \times A_2$ (examinees crossed with items), but these classes are not given to us. Instead, the data consist of answers to test items, often binary indicators of whether the answer was correct or not. From these data the membership of the equivalence classes must be inferred before one can ascertain whether these classes can be represented by a set of linear, parallel isoprobability curves. Because the theory stipulates restrictions that hold for every examinee, while the amount of data per examinee is small, there is little power to reject these hypotheses despite anomalies in the data. Many restrictions might be accepted that would be deemed invalid if the actual P_{ij} were revealed.

IRT assumes that conditional on θ and δ, the P_{ij} are independent across items and examinees. Although correlated response probabilities attest to the existence of additional latent traits that affect performance, unidimensional models are fit to the data anyway. Sometimes it is clear even without statistical tests that the model does not fit the data. Consider multiple-choice exams, where the lower asymptote on the probability of a correct response is not zero. The lower asymptote is more important for low-ability than high-ability examinees. A conjoint structure for these data must take into account the lower asymptote as another factor determining P_{ij} or the scaled values of examinee ability will be wrong. Indeed, if responses to a multiple-choice exam are tested to see whether the definition of conjoint additivity is met, data that appear to meet this definition will no longer do so once guessing is factored in unless all items are equally “guessable.”

Even when the data do fit the model, the extent to which they have been selected for just this reason should be kept in mind. Indeed, it is not clear that the word *data* is the right one in this context because the tests are designed by test makers who first decide on a scaling model and then strive to ensure that

change. This allows us to compare directly the gains of individual students or schools that start at different places on the test score metric” (Bryk et al. 1998, p. 46).

their test items meet the assumptions of that model.[15] Even if they are wholly successful in this endeavor, the data represent a selected, even massaged, slice of reality and not a world of brute facts.

Thus, notwithstanding the support given by representational measurement theory to IRT methods, IRT can fail to produce satisfactory interval scales. Verifying that the conditions of the theory are met is difficult even for test makers, let alone statisticians and behavioral scientists who use test scores for value-added assessments.

As the worker-machine example shows, even if test data satisfy the conditions of a conjoint structure, there might be some other scale with a claim to measure what we mean by achievement. IRT scales have the peculiarity that the increase in ability required to raise the probability of a correct response by any fixed amount is independent of the difficulty of the question. That is, raising the probability of answering a very difficult question from .1 to .9 takes the same additional knowledge as it does to raise the probability of answering a very simple item from .1 to .9. (Observe that in this argument, .1 and .9 can be replaced by any other numbers one likes, for example .0001 and .9999.) That it takes the same increase in ability to master a hard task as an easy one follows directly from conjoint additivity (specifically, the parallel equiprobability contours that result when items and examinees are scaled to represent conjointness), but it may be difficult for many readers to square this notion with other ideas they entertain about achievement, based on the use of the term in other contexts: how long it takes to accomplish these two tasks, how hard instructors must work, the extent to which a student who has mastered the more difficult item is in a position to tackle a variety of other tasks and problems compared with the student who has mastered the easier item, and so forth. In short, we may find ourselves in the position of the production engineer in the worker-machine example: taking into account the other information we possess, we might find the scale derived from conjoint additivity lacking, notwithstanding its impeccable pedigree on purely formal grounds as an interval scale.

At this point it may be useful to examine some of the scales that have been produced using IRT methods. If larger increments of ability (as measured by an alternative metric grounded in some of the aforementioned phenomena) are in fact required to produce the same change in P_{ij} as questions become more

[15] For example, in the Rasch model, items have a common discrimination parameter (their item characteristic curves do not cross). Test makers using the Rasch model plot ICCs and discard items that violate this assumption. IRT also assumes that conditional on θ, probabilities of a correct response are independent over examinees. Violations of this assumption lead items to be discarded on grounds of potential bias. This is probably a good idea, but it further weakens the sense in which we are dealing here with the fit between a model and "data."

Table 1. Scale Scores for Comprehensive Test of Basic Skills, 1981 Norming Sample

	Reading/Vocabulary			Mathematics Computation		
Grades	Mean Score	Std. Dev.	Mean Between-Grade Change	Mean Score	Std. Dev.	Mean Between-Grade Change
1	488	85	–	390	158	–
2	579	78	91	576	77	186
3	622	65	43	643	44	67
4	652	60	30	676	35	33
5	678	59	26	699	24	23
6	697	59	19	713	20	14
7	711	57	14	721	23	6
8	724	54	13	728	23	7
9	741	52	17	736	17	8
10	758	52	17	739	16	3
11	768	53	10	741	18	2
12	773	55	5	741	20	0

Source: Yen 1986. Reprinted with permission of Blackwell Publishing Ltd.

difficult, IRT scaling will compress the high end of the scale, diminishing mean gains between the upper grades and reducing the variance in achievement. Evidence that something of this kind occurs is found in developmental scales used to measure student growth across grades. Students at different grade levels are typically given different exams, but by including a sufficient number of overlapping items on forms at adjacent grade levels, performance on one test can be linked to performance on another test, facilitating the creation of a single scale of ability spanning multiple grade levels.

Table 1, reproduced from Yen (1986), displays scores for the norming sample for the Comprehensive Test of Basic Skills (CTBS), Form Q, developed by CTB/McGraw-Hill and scaled using IRT methods for the first time in 1981.[16] In both subjects, mean growth between grades drops dramatically, and almost monotonically, between the lowest elementary grades and secondary grades. In addition, the standard deviation of scale scores declines as the mean score rises.

[16] Prior to 1981, the CTBS was scaled using an older method known as Thurstone scaling. Although it has been claimed that Thurstone scaling produces scores on an interval scale, there is no basis for this claim comparable to the proofs provided for conjoint measurement structures in representational measurement theory: scale properties are not based on the structure of empirical relations holding among test items and examinees but on ad hoc assumptions about the distribution of ability in the population tested.

Table 2. Mean Between-Grade Differences, Terra Nova

Subject Year	2nd–3rd	3rd–4th	4th–5th	5th–6th	6th–7th	7th–8th	8th–9th
Mississippi, 2001							
Language arts	30.2	20.4	17.8	6.9	9.9	10.9	–
Reading	23.8	21.5	15.5	10.5	9.2	12.4	–
Math	47.3	24.7	19.0	21.7	10.7	17.0	–
New Mexico, 2003							
Language arts	–	14.8	12.6	5.1	3.7	4.5	8.3
Reading	–	14.4	13.8	5.0	4.7	11.4	4.9
Math	–	21.1	14.6	19.0	5.5	16.5	5.4
Science	–	20.2	13.9	9.0	11.6	11.9	7.1
History/SS	–	13.3	6.2	11.2	11.5	3.6	3.6

Source: Author's calculations from school-level data posted on www.schooldata.org, maintained by the American Institutes for Research. Data for each school are weighted by enrollment and aggregated to the state level.

CTB/McGraw-Hill has since superseded the CTBS with the Terra Nova series. The decline in between-grade gains seen in the CTBS norming sample is still evident, though less pronounced, in Terra Nova results for Mississippi and New Mexico, as shown in table 2.[17] Mean gains in all subjects tend to decrease with grade level, though there is a break in the pattern between grades 7 and 8.

Between-grade gains can be affected by the differences in the content of tests and by linking error. It is particularly instructive, therefore, to see the patterns that emerge when there are no test forms specific to a grade level and no linking error. The Northwest Evaluation Association uses computer-adaptive testing in which items are drawn from a single, large item bank. Results for all examinees in reading and mathematics in the fall of 2005 are presented in table 3. As in the previous tables, we again find that between-grade differences decline with grade level. Within-grade variance in scores is stable (reading) or increases moderately (mathematics).

Because this is at base a dispute about how to use words, we need to be careful in discussing these phenomena. If the test data in fact exhibit a conjoint structure (let us concede the point for now), the IRT θ is an interval scaled variable. Yet this scale commits us to the conclusion that the variance of

[17] Data are from www.SchoolData.org maintained by the American Institutes for Research. New Mexico and Mississippi were two of a handful of states in these data that reported mean scale scores for vertically linked tests produced by test makers known to employ IRT scaling methods.

Table 3. Scale Scores, Northwest Evaluation Association, Fall 2005

	Reading			Mathematics		
Grade	Mean	Std	Change	Mean	Std	Change
2	175.57	16.22	–	179.02	11.81	–
3	190.31	15.56	14.74	192.96	12.06	13.95
4	199.79	14.95	9.48	203.81	12.80	10.85
5	206.65	14.60	6.86	212.35	13.92	8.53
6	211.49	14.76	4.84	218.79	15.00	6.44
7	215.44	14.82	3.96	224.59	15.99	5.80
8	219.01	14.76	3.56	229.38	16.79	4.79
9	220.93	15.28	1.92	231.76	17.42	2.38

Source: Author's calculations from data provided by NWEA.

reading ability is no greater among high school students than second graders. Most of us, I suspect, would respond that this scale fails to capture something about the word *ability* (or *achievement*) that causes us to recoil from such a conclusion.

Readers wondering whether their own pre-theoretical notions of these terms accord with the usage implied by IRT are invited to consider the sample of mathematics test items displayed in figure 4, taken from the Northwest Evaluation Association (2008) Web site. The items are drawn from a larger chart providing examples of test questions at eleven different levels of difficulty. The two selections represent items scored at the 171–80 level and the 241–50 level, respectively.[18] Consider the following question: if student A is given the items in the first set and student B the items in the second set, and if initially each student is able to answer only two of seven items correctly, which student will have to learn more mathematics in order to answer all seven items correctly? Student A has basics of addition and subtraction to learn, as well as (perhaps) how to read simple charts and diagrams. None of the required calculations is taxing; all could be done by counting on one's fingers. By contrast, student B must make up deficits in several of the following areas: decimal notation, fractions, factoring of polynomials, solving algebraic equations in one unknown, solid geometry, reading box plots, calculating percentages. The calculations required are more demanding. However, the

[18] According to the data in table 3, the average second grader tested in 2005 would have answered slightly more than half the questions in the first set correctly, while the average ninth grader would have responded correctly to slightly less than half the questions in the second set.

Student A

68 equals

- ✓A. 60 + 8
- B. 60 + 80
- C. 6 + 8
- D. 600 + 8
- E. 6 + 80

99
−56

- A. 34
- B. 155
- C. 53
- ✓D. 43
- E. 42

14 □ 6 = 8
□ =

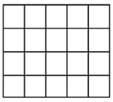
- A. +
- ✓B. −
- C. ÷
- D. <
- E. >

Which makes us think of a circle?

- A. Block
- B. Pen
- C. Door
- D. A football field
- ✓E. Bicycle wheel

What is the area of the figure?

- A. 18 square units
- B. 9 square units
- ✓C. 20 square units
- D. 16 square units
- E. 5 square units



Student Council Election Results	
Student	**Number of Votes**
Ann	卌 卌 卌 卌 ||
Mark	卌 卌 卌 |
Sue	卌 卌 卌 卌

How many votes did Mark get?

- A. 20
- ✓B. 16
- C. 22
- D. 17
- E. 19

10 Students will play basketball after school. They ask 5 more students to play with them.

If you want to find out how many students will play in all, what method should you use?

- A. Simplify
- B. Subtract
- ✓C. Add
- D. Multiply
- E. Divide

Source: Northwest Evaluation Association RIT Charts, Sample Mathematics Items, 2008.

Student B

43,000 equals:

- A. 4.3×10^3
- ✓B. 4.3×10^4
- C. 4.3×10^5
- D. 43×10^4
- E. 43×10^5

$\frac{31}{6} - (-3\frac{3}{8}) =$

- ✓A. $8\frac{13}{24}$
- B. $4\frac{1}{7}$
- C. $1\frac{19}{24}$
- D. $17\frac{7}{16}$
- E. $-17\frac{7}{16}$

Factor $x^2 - 5x - 36$

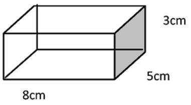
- A. $(x - 6)(x - 6)$
- B. $(x + 9)(x - 4)$
- C. $(x - 36)(x + 1)$
- ✓D. $(x - 9)(x + 4)$
- E. $(x + 6)(x + 6)$

Using the Pythagorean Theorem, $a^2 + b^2 = c^2$,
When a = 9 and b = 13, then c = ?

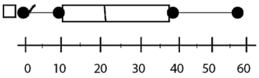
- A. 8
- B. 21
- ✓C. 15
- D. $\sqrt{21}$
- E. 225

Calculate the surface area of this rectangular solid.

- A. 79 cm^2
- B. 110 cm^2
- C. 120 cm^2
- D. 128 cm^2
- ✓E. 158 cm^2



Ages of the First 37 People to Enter the Museum



What percentage of these people are less than 10 years old? (Round to the whole percent)

- A. 12%
- B. 33%
- ✓C. 23%
- D. 40%
- E. 19%

A $30.00 pair of jeans is discounted 20%. If sales tax is 5%, what will be the final price for the jeans?

- A. $22.80
- B. $24.00
- C. $24.40
- ✓D. $25.20
- E. $28.35

Figure 4. Math Items at Two Levels of Difficulty

correct answer, according to IRT, is that both require the same increase in mathematics ability.[19]

One might conclude from this evidence that psychometricians should not attempt to construct overarching developmental scales for mathematics and reading ability that span so many age levels. But other concerns are also raised. Between-grade gains begin to decline at the lowest grade levels in tables 1–3: gains between third and fourth grades are markedly lower than gains between second and third. Between-grade comparisons may matter little for value-added assessment if most instructors, particularly in the elementary grades, teach only one grade level. However, there are also implications for within-grade assessments. If third graders are not in fact learning less than second graders—if instead IRT methods have compressed the true scale—then the true gains of higher-achieving students within these grades are understated.[20]

5. OPTIONS FOR VALUE-ADDED ASSESSMENT

What options are available to the practitioner who wants to conduct value-added assessment but is unwilling to accept at face value claims that the IRT ability trait is measured on an interval scale? Broadly speaking, there are three available courses of action.

1. Use the θ scale anyway;
2. Choose another measure of achievement with an interval scaled metric; or
3. Adopt analytical methods suited to ordinal data.

[19] Strictly speaking, this is the correct answer only if items have equal discrimination parameters, that is, the one-parameter IRT model fits the data. In fact, NWEA does use the one-parameter model and has gone to considerable lengths to verify that the items meet the assumptions of that model. When I put this question to faculty and graduate students of my department in the School of Education at Vanderbilt, 13 of the 108 respondents chose A, 47 chose B, 15 said the amounts were equal, and 33 said the answer was indeterminate. Obviously this was not a scientifically conducted survey, nor is it clear just what respondents meant by their answers. Conversations with some revealed that they converted the phrase “more mathematics” into something more readily quantified, such as the amount of time a student would need to acquire these skills. Persons who said the answer was indeterminate may have meant it was indeterminate in principle (“more mathematics” is meaningless) or simply that that answer could not be determined from the information given. Nonetheless, it is striking how few gave the psychometrically correct response.

[20] The phenomena of decreasing gains and diminishing or constant variance with advancing grade level have been treated in the psychometric literature under the heading “scale shrinkage.” While a number of explanations have been advanced, these explanations typically assume that there is a true IRT model that fits the data (with ability, perhaps, multidimensional rather than unidimensional) and that various problems (e.g., a failure to specify the true model, the small number of items on the test, changing test reliability within or across grades, ceiling and floor effects) prevent practitioners from recovering the true values of θ. Psychometricians have disagreed about the extent to which these explanations account for the phenomena in question. (For notable contributions to this literature, see Yen 1985; Camilli 1988; Camilli, Yamamoto, and Wang 1993; Yen and Burket 1997.) By contrast, the point I am making here is that even in the absence of these problems, IRT scales are likely to exhibit compression at the high end when compared with pre-theoretical notions of achievement grounded in a wider set of empirical relations.

In this section I argue that neither (1) nor (2) is an attractive option. I then go on to demonstrate the feasibility of (3).

(1) Using the θ Scale

Even if the IRT ability scale possesses at best ordinal significance, one might continue to use it if reasonable transformations of θ all yield essentially the same estimates of value added. (Compare the claim that statistics calculated from ordinal variables are generally robust to all but the most grotesque transformations of the original scale.)[21] The practice of many social scientists who are aware that achievement scales may not be of the interval type but proceed with value-added assessment anyway suggests that this view may be widespread.

Unfortunately, to test whether reasonable transformations of θ yield essentially the same measures of value added requires some sense of what is reasonable. Absent that, there may be a tendency for researchers conducting sensitivity analyses to decide that the alternatives that are reasonable are those that leave their original estimates largely intact. The compression of scales displayed in tables 1–3 suggests one possibility: assume an achievement scale in which average gains are equal across grades. I have examined the consequences of adopting this alternative scaling for data from a sample of districts in a Southern state.[22] Data are from mathematics tests administered in grades 2–8 during the 2005–6 school year. The tests were scaled using the one-parameter IRT model. Results were similar to those we have seen in tables 2–4, with near-monotonic declines in between-grade gains. A transformation $\psi = g(\theta)$ was sought that would equalize between-grade gains for the median student.[23] It turned out that this could be closely approximated by a quadratic function increasing over the range of observed scores. Figure 5 depicts the relationship between the transformed and original scaled values for the median examinee. While one might object to the ψ scale on the grounds that median student gains are not "really" equal across grades, tapering off as students approach adolescence, figure 5 shows that the transformation from θ to ψ is not driven by the upper grades—essentially the same curve would be found if data points for grades 7 and 8 were dropped. Moreover, $g(\theta)$

[21] Cliff (1996) attributes this remark to Abelson and Tukey (1959), but it does not appear in that paper.

[22] Anonymity has been promised to both the state and the test maker. Data are available for only a portion of the state. In addition, because between-grade growth is central to the analysis, only those districts that tested at least 90 percent of students in each grade are included. The final sample comprised 98,760 students in grades 2–8 during the 2005–6 school year.

[23] To accomplish this, the original median scores were replaced with a new series in which the between-grade gain was set to the overall sample median gain across all grades. This left unchanged the second-grade values but altered the values in subsequent grades, as shown in figure 5. A quadratic function of θ was then fit to the new series. The fit, as shown, was exceedingly close, though on the resulting scale, median gains can vary by ±.5 points.

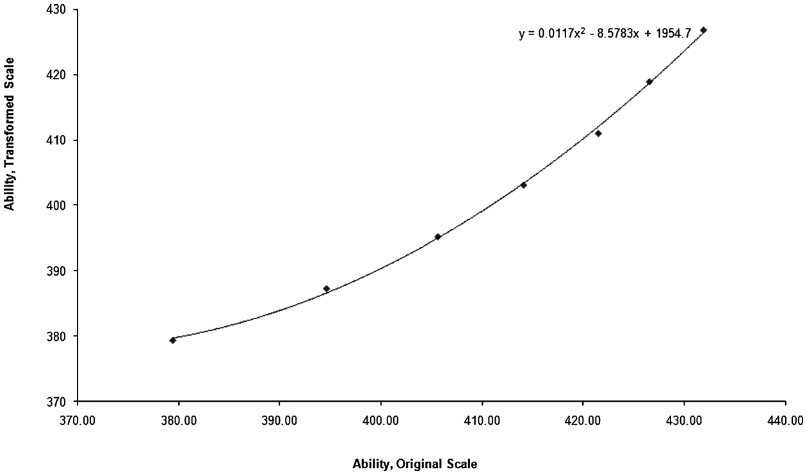


Figure 5. Scale Transformation Equating Median Between-Grade Gains

exhibits only a modest departure from linearity. It seems unlikely that many researchers would regard it as a grotesque transformation of the data.

Nonetheless, the consequences of this transformation for the distribution of growth are pronounced (table 4). At each grade level, I have calculated the change in θ (alternatively, ψ) required to remain at the 10th, 25th, 50th, 75th, and 90th percentiles of the achievement distribution when advancing to the next grade. The absolute changes in θ and ψ that meet this criterion are affected by the magnitude of the median student growth (and are therefore scale dependent—i.e., they depend on choice of units). However, relative changes are invariant to the choice of units. Accordingly, column 1, top panel, presents the ratio of $\Delta\theta_{25}$ to $\Delta\theta_{10}$ (the change required to remain at the 25th percentile over the change required to stay at the 10th percentile). Comparable ratios for the 50th, 75th, and 95th percentiles appear in the other columns. The lower panel contains the same ratios for the ψ scale.

In the original scale, differences in growth at various points of the distribution are not pronounced. The ratios are greatest in the fifth and sixth grades, but even here there is not much difference between a student at the median and one at the 90th percentile. By contrast, the ratios are much greater using the transformed scale and increase monotonically as we move to the right, from $\Delta\psi_{25}/\Delta\psi_{10}$ to $\Delta\psi_{90}/\Delta\psi_{10}$. While the direction of the change is what we would expect on the basis of the preceding discussion—less compression at the high end of the ψ scale compared with the θ scale—the magnitude of the difference is surprising. The impact on value-added assessment depends on how students are distributed over schools and teachers. Clearly, changes of the magnitude shown in table 4 can make a great difference to teacher value added when that distribution is not uniform.

Table 4. Effect of Scale Transformations on Between-Grade Growth

	Relative to students at the 10th percentile, growth by students at the:			
	25th percentile	**Median**	**75th percentile**	**90th percentile**
		Original Scale		
2nd–3rd	0.97	1.06	1.03	0.95
3rd–4th	1.10	1.03	1.16	1.34
4th–5th	0.96	1.15	1.35	1.23
5th–6th	1.61	2.12	2.17	2.24
6th–7th	1.21	1.39	1.43	1.62
7th–8th	1.16	1.17	1.16	1.13
		Transformed Scale		
2nd–3rd	2.63	5.06	6.83	7.90
3rd–4th	1.69	2.10	2.93	3.95
4th–5th	1.30	1.94	2.74	2.87
5th–6th	2.12	3.48	4.25	4.91
6th–7th	1.60	2.29	2.79	3.50
7th–8th	1.51	1.89	2.17	2.35

Source: Mathematics test results furnished to author from a southern state, 2005–6.

(2) Changing the Scale

Clearly nothing is gained by substituting percentiles, normal curve equivalents, or standardized scores for θ. While all are used in the research literature, no one claims that they are interval scaled. However, two approaches merit discussion, not because they work any better but because the view seems to be gaining currency that they do, or could.

The first, which I will refer to as *binning*, assigns every examinee to a group defined by prior achievement (e.g., deciles of the distribution of prior scores). Gain scores are normalized by mean gains within bins, either through dividing by the mean gain or standardizing with respect to the mean and standard deviation within each bin. The normalized gain scores are then used as raw data for further analysis, which could include value-added assessment of schools and teachers. Examples of this approach are Springer (2008) and Hanushek et al. (2005). In the latter study, binning is explicitly motivated by concern that the metric in which test results are reported does not represent true gains uniformly at all points of the achievement distribution.[24]

[24] A practice similar to binning has been used in the Educational Value-Added Assessment System (EVAAS) of the SAS Institute, wherein gain scores are divided by the gain required to keep an examinee at the same percentile of the post-test distribution that he or she occupied in the pretest

Binning does not solve the problem of scale dependence; it merely declares that a normalized gain in one bin is to count the same as one that takes the same normalized value in another bin. The declaration does not ensure that these two gains are equal when measured on the true scale (if such a thing exists). Rather than solving the problem, binning is simply the substitution of a particular transformation for the θ scale.

The notion that binning represents a solution may derive from the fact that normalizing within bins removes much of the effect of any prior transformation of scale (for example, the substitution of the ψ scale for the θ scale). Differences between bins have no effect on value-added measures. Only differences within bins matter (though these are still affected by the transformation whenever the slope of $g(\theta) = \psi$ at a bin mean differs from 1). Hence the results of the binned analysis are less sensitive to prior transformations of the original scale. This may have led some to believe that scale no longer matters as much. This is not so, given that the normalized-within-bin scores are themselves just another transformation of θ. Moreover, if invariance of this kind were the sole desideratum, percentiles could be used in place of θ. Percentiles are invariant to any increasing transformation of θ, but that does not make percentiles an interval scale of achievement.

Suppose we decide that the problem of finding interval scales for academic achievement is intractable. We still might be able to conduct value-added assessment if achievement—by whatever metric—could be mapped into other variables whose measurement poses no such difficulties. This mapping could go backward to inputs or forward to outcomes. In the former, academic achievement would be related to measurable inputs required to produce that achievement. Thus, instead of worrying whether one student's five-point gain was really the equal of another student's five-point gain, we would concern ourselves with measuring the educational inputs required to produce either of these gains. If those inputs (e.g., teacher time) turn out to be equal, then for all purposes that matter the two gains are equivalent. If those inputs turn out not to be equal, the teacher or school that has produced the gain requiring the greater inputs has contributed more and should be so recognized by value-added analysis.

Forward mapping treats test scores as an intermediate output. Value-added assessment would proceed by tying scores to long-term consequences.

distribution (Ballou 2005). Thus for exams like the Iowa Test of Basic Skills that exhibit increasing variance at higher grade levels, the transformation pulls up gains of examinees whose pretest scores were below the mean and reduces gains of examinees whose pretest scores were above the mean. There is no reason that this should be regarded as superior to using the original scale.

> An important methodological issue . . . is the problem of choosing the correct metric with which to measure academic growth. Because the metric issue is so perplexing, almost all researchers simply use the particular test at their disposal, without questioning how the test's metric affects the results. . . . The only solution I see to the problem of determining whether gains from different points on a scale are equivalent is to associate a particular test with an outcome we want to predict (say, educational attainment or earnings), estimate the functional form of this relationship, and then use this functional form to assess the magnitude of gains. For example, if test scores are linearly related to years of schooling, then gains of 50 points can be considered equal, regardless of the starting point. If the log of scores is linearly related to years of schooling, however, then a gain of 50 points from a lower initial score is worth more than a gain of 50 points from a higher initial score. This "solution" is, of course, very unsatisfactory, because the functional form of the relationship between test scores and outcomes undoubtedly varies across outcomes. (Phillips 2000, p. 127)

As the final sentence of this passage suggests, we are very far from being able to carry out either of these programs. Only some educational inputs are easily quantified. For such inputs as the clarity of a teacher's explanations or the capacity to inspire students, the challenges to quantification are at least as great as for academic achievement. Indeed, the low predictive power of those inputs that are easily quantified is largely responsible for the current interest in value-added assessment.

The practical difficulty mentioned in the last sentence of the quoted passage is not the only problem facing the forward mapping of test scores to long-term educational outcomes. Given the variation in test results in tables 1–3, it would almost certainly be the case that the functional form of the relationship between test scores and outcomes would vary across tests as well as outcomes. The introduction of each new test would require additional analysis to determine how scores on its metric were related to long-term objectives like educational attainment and earnings. In many cases, the data for such analysis would not be available for years to come, if ever. In the interim we would have to make do with very imperfect efforts to equate the new tests with tests already in use (for which we would hope this mapping had already been done).

These are the technical issues. There is in addition the difficult normative question of how to value various outcomes for different students in order to assign a unique social value to each θ_i.It is not obvious how we would come by these weights. Even if future earnings were the only outcome that mattered, we would require the relative values of a marginal dollar of future income for all examinees (whose future incomes are, of course, unknown at the time the assessment is done). Attaching a price to the nonpecuniary benefits of an education is still more difficult. Education is a transformative enterprise. The ex ante value placed on acquiring an appreciation of great literature is doubtless very different from the ex post value. In these circumstances there is likely to be a great discrepancy between the compensating and equivalent variations associated with a given educational investment. Accordingly, the weights in our index would have to reflect the values of "society" rather than the still-unformed persons to be educated.

It is not clear that we should subject decisions about education to this kind of utilitarian calculus because it fails to respect the autonomy of individuals. There is a strong tradition in our polity of regarding educational opportunity as a right. Individuals have a claim on educational resources not because distributing resources in this manner maximizes a social welfare function but because they are entitled to the chance to realize their potential as individuals. If we take this seriously, the notion that teachers and schools are to be evaluated by converting test scores into the outputs that matter and weighting these outputs according to their social value is wrongheaded. The point of education is to provide students with skills and knowledge that as autonomous persons they can make of what they will. Teachers should therefore be judged on how successfully they equip their students with these tools—regardless of anyone's views of the merits of the final purposes, within wide limits, for which students use them.

(3) Analyzing Test Scores as Ordinal Data

The final option for researchers uncertain of the metric properties of ability scales is to treat such scales as ordinal, thus forgoing any analysis based on the distance between two scores. On the assumption that θ scales contain valid ordinal information about examinees, statistics based on the direction of this distance remain meaningful. There are a variety of closely related statistics of this kind (known generally as measures of concordance/discordance) employed in the analysis of ordinal data. In this discussion I will not attempt to identify a particular approach as best. Rather, my objective is to demonstrate the feasibility of such methods for value-added assessment.

Suppose we want to compare the achievement of teacher A's class to the achievement of other students at the same grade level in the school system. If A has n students, and teachers elsewhere in the system have m students, there are nm possible pairwise comparisons of achievement. Because only ordinal statements are meaningful, each pair is examined to determine which student has the higher score. If it is A's student, we count this as one in A's favor (+1); if it is the student from elsewhere in the system, we count this as one against A (−1). Ties count as zeros. The sum of these counts, divided by the number of pairs, is known as *Somers' d statistic.* Somers' d can be considered an estimate of the difference in two probabilities: the probability that a randomly selected student from A's class outperforms a randomly selected student from elsewhere in the system, less the probability that A's student scores below the outsider.

This procedure suffers from the obvious defect that no adjustment has been made for other influences on achievement. In conventional value-added analyses, this might be accomplished through the introduction of prior test scores as covariates in a regression model or by conditioning on prior scores in some other manner. An analogous procedure in the ordinal framework would be to divide students into groups on the basis of one or more prior test scores, compute separate values of Somers' d by group, and aggregate the resulting statistics using the share of students in each group as weights. In principle there is no reason to restrict the information used to define groups to test scores. Any student characteristic could be used to define a group. Only data limitations prevent the construction of ever-finer groups.

Results from an application of this approach are presented in table 5. Because the data used in the previous example do not contain teacher identifiers, for this application I have used a different data set provided by a single large district that contains student-teacher links. Two sets of value-added estimates are presented for teachers of fifth-grade mathematics. The first is the weighted Somers' d statistic, based on pairwise comparisons of each teacher's students to other fifth graders in the district. To control for prior achievement, students were grouped by decile of the fourth-grade mathematics score. Students without fourth-grade scores were dropped from the analysis. The second value-added measure is obtained from a regression model in which fifth-grade scores were regressed on fourth-grade scores and a dummy variable for the teacher in question. Separate regressions were run for each teacher so that each teacher was compared with a hypothetical counterpart representing the average of the rest of the district, preserving the parallel with the Somers' d statistic. The coefficients on the dummy variables represent teachers' value added. Statistical significance was assessed using the conventional *t*-statistics in the case of the regression analysis and jackknifed standard errors in the case of the ordinal analysis.

How much difference does it make to a teacher to be evaluated by one method rather than the other? The hypothesis that teachers are ranked the same by both methods is rejected by the Wilcoxon signed rank test ($p = .0078$). The proportion of statistically significant estimates is higher using the ordinal measure (which is less sensitive to noisiness in test scores): 86 of the 237 teachers are significant by this measure, compared with 64 by the other.

Table 5. Comparison of Conventional and Ordinal Measures of Teacher Value Added

Total number of teachers	237
Wilcoxon signed ranks test, *p*-value	0.0078
Maximum discrepancy in ranks	229
Absolute discrepancy in ranks, 90th percentile	45
Number of statistically significant teacher effects, fixed effect estimate	64
Number of statistically significant teacher effects, ordinal estimate	86
Number of significant effects by both estimates	55
Number ranked in the top quartile	59
Number ranked in the top quartile by both estimates	47
Number ranked in the bottom quartile by both estimates	48
Number ranked above the median by one estimate, below the median by the other	14
Number ranked in the bottom quartile by one, top quartile by the other	3

Source: Fifth-grade mathematics teachers, large southern district, 2005–6, author's calculations.

The maximum discrepancy in ranks is 229 positions (out of 237 teachers in all). In 10 percent of cases, the discrepancy in ranks is 45 positions or more.

Depending on the uses to which value-added assessment is put, the question of greatest importance to teachers may be whether they fall at one end of the distribution or the other. There are fifty-nine teachers in each quartile of the distribution. The two measures agree in forty-seven cases on the teachers in the top quartile and in forty-eight cases on teachers in the bottom quartile. Thus, if falling in the top quartile qualifies a teacher for a reward, twenty-four teachers (more than a third of the number of awardees) will qualify or not depending on which measure is used. A comparable figure applies to teachers placing in the bottom quartile, if that event is used to determine sanctions. In a small number of cases (three), the effect of using one measure rather than the other is great enough to move a teacher from the top quartile to the bottom quartile.

In principle it is possible to condition on multiple variables (additional prior test scores, student demographic characteristics, or socioeconomic status [SES]) by defining groups as functions of several covariates. In practice this is apt to exceed the capacity of the data. Consider, for example, a data set containing prior test scores in two subjects, plus indicators of race and participation in the free and reduced price lunch program. If groups are defined by deciles of the two test scores plus two binary indicators, the total number of groups is 400 (10 × 10 × 2 × 2). In a district of moderate size, there could be many cells with only one observation and therefore no matching pair.

A two-stage method circumvents this difficulty. In the first stage, multiple covariates are used to predict Prob($Y_i > Y_j$) $-$ Prob($Y_j < Y_i$) for all pairs of students i and j, where Y is the dependent variable of ultimate interest (e.g., end-of-year test scores in the current year). The prediction is of the form $\hat{\pi}_i = \Sigma w_k(X_{ki} - X_{kj})$ or $\Sigma w_k \text{sgn}(X_{ki} - X_{kj})$, depending on whether the X_k are themselves interval scaled or ordinal. The w_k are weights that reflect how informative the different covariates are about sgn($Y_i -$ Y_j). (For details, see Cliff 1996.)

In the second stage, the n students of one teacher are compared to the m students elsewhere in the same system, using $\hat{\pi}_i$ (or grouped values of $\hat{\pi}_i$) as the covariate. $\hat{\pi}_i$ is therefore analogous to a propensity score: it is a summary measure of the effects of the stage-one covariates on the probability that a student outranks other students, *before controlling for teachers*. The resulting measure of a teacher's value added is based on her students' performance relative to this prediction (in the ordinal sense, of course).

Given that many achievement tests are now administered to thousands of students throughout a state, it is worth noting that all the data can be used in the first stage to form an ordinal prediction based on prior achievement, demographics, and SES while continuing to rely on within-district comparisons in the second stage for the final measures of value added.[25]

6. CONCLUSION

Are IRT ability traits measured on an interval scale? It seems hazardous to assume so. Whether examinees and test items constitute a conjoint structure depends on the makeup of equivalence classes defined by the P_{ij}, but those are not given. Statistical testing can reveal whether the data are strongly inconsistent with this hypothesis, but moderate departures from the conjoint structure almost certainly go undetected. Moreover, even if these conditions hold in the norming samples used by test makers to calibrate item difficulties, this provides no assurance that they will hold in the population of students to whom the test is finally administered.

End users of the data, such as practitioners of value-added assessment, typically have no access to item-level data to test these assumptions themselves. Moreover, even if the assumptions are met, conjoint additivity may not capture everything we want in a scale of achievement. It would seem wise, then, to check the plausibility of the resulting scales. On this count IRT ability scales often do poorly. Gain scores frequently fall from one grade to the next. While some of this may reflect adolescents' declining interest in academic

[25] Within-district comparisons are preferred, given the impact of curriculum and other district policies on achievement.

achievement, the patterns set in as early as third grade and the drop between second and third grade is often the largest. In addition, IRT ability often shows a diminishing or constant variance from lower to higher grades.

What, then, is the practitioner of value-added assessment to do? It is no good hoping that the choice of scale makes little difference to estimates of value added. We have seen that reasonable transformations of the θ scale can have a substantial effect on relative gains across the distribution of achievement. No other scales with superior metric properties are at hand. We can, however, use methods of ordinal data analysis on the assumption that IRT scales (or any of their monotonic transformations) at least permit us to rank students.

Ordinal analysis changes the question we ask in value-added assessment. Instead of measuring mean achievement of a teacher's students vis-à-vis the students of a (hypothetical) average instructor, we ask what fraction of the former outperform the latter. In ordinal analysis, as in regression-based methods, it is possible to control for other influences in order to isolate the teacher's or the school's contribution. Clearly, if θ is an interval-scaled variable, ordinal methods throw away valuable information. Practitioners should ask themselves, however, whether they are so confident of the metric properties of θ scales that they are willing to attribute differences between conventional estimates of value added and estimates based on ordinal analysis to the superiority of the former.

Ordinal methods have other advantages. They are likely to be more robust to measurement error in test scores and to various model misspecifications (though the question of robustness is a complicated one; see Cliff 1996). The question they answer may be a more sensible way to evaluate educators, given that it attaches more value to spreading gains over a wider number of students, compared with larger but more concentrated gains. However, this article has considered ordinal methods from one standpoint only—that of finding appropriate value-added models when test scores are not expressed on an interval scale. Numerous questions have been raised about the assumptions and methods of more conventional regression-based analyses. Some of these concerns can be addressed through modifications of those models. It remains to be seen whether the same concerns arise with respect to ordinal methods and, if so, how readily they can be accommodated within the ordinal framework.

The author thanks Kurt Scheib and Warren Langevin for their expert research assistance.

REFERENCES

Abelson, Robert, and John Tukey. 1959. Efficient conversion of nonmetric information into metric information. In *Proceedings of the social statistics section*, pp. 226–30. Washington, DC: American Statistical Association.

Ballou, Dale. 2005. Value-added assessment: Lessons from Tennessee. In *Value added models in education: Theory and applications*, edited by Robert Lissitz, pp. 272–319. Maple Grove, MN: JAI Press.

Bryk, Anthony, Yeow Meng Thum, John Q. Easton, and Stuart Luppescu. 1998. *Academic productivity of Chicago public elementary schools*. Chicago: Consortium on Chicago School Research.

Camilli, Gregory. 1988. Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics* 13 (3): 227–41.

Camilli, Gregory, Kentaro Yamamoto, and Ming-wei Wang. 1993. Scale shrinkage in vertical equating. *Applied Psychological Measurement* 17 (4): 379–88.

Cliff, Norman. 1996. *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, Ronald, and H. Swaminathan. 1985. *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hambleton, Ronald, H. Swaminathan, and Jane Rogers. 1991. *Fundamentals of item response theory*. Newbury Park: Sage.

Hand, D. H. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society, Series A* 159 (3): 445–92.

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, and Gregory F. Branch. 2005. Charter school quality and parental decision-making with school choice. NBER Working Paper No. 11252.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of measurement, volume 1: Additive and polynomial representations*. New York: Academic Press.

Lord, Frederic. 1975. The "ability" scale in item characteristic curve theory. *Psychometrika* 40 (2): 205–17.

Mislevy, Robert. 1987. Recent developments in item response theory with implications for teacher certification. *Review of Research in Education* 14: 239–75.

Northwest Evaluation Association (NWEA). 2008. Mathematics. Available www.nwea.org/assessments/ritcharts.asp. Accessed 20 February 2008.

Phillips, Meredith. 2000. Understanding ethnic differences in academic achievement: Empirical lessons. In *Analytic issues in the assessment of student achievement*, edited by David W. Grissmer and J. Michael Ross, pp. 103–32. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Springer, Matthew G. 2008. Accountability incentives: Do schools practice educational triage? *Education Next* 81: 75–79.

Stevens, S. S. 1946. On the theory of scales of measurement. *Science* 103: 677–80.

Wright, Benjamin D. 1999. Fundamental measurement for psychology. In *The new rules of measurement*, edited by Susan E. Embretson and Scott L. Hershberger, pp. 65–104. Mahwah, NJ: Lawrence Erlbaum.

Yen, Wendy. 1985. Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika* 50 (4): 399–410.

Yen, Wendy. 1986. The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement* 23 (4): 299–325.

Yen, Wendy, and George R. Burket. 1997. Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement* 34 (4): 293–313.

Zwick, Rebecca. 1992. Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics* 17 (2): 205–18.