Measuring Growth with Vertical Scales

Derek C. Briggs
University of Colorado, Boulder

September 2, 2012

Abstract

A vertical score scale is needed to measure growth across multiple tests in terms of absolute changes in magnitude. Since the warrant for subsequent growth interpretations depends upon the assumption that the scale has interval properties, the validation of a vertical scale would seem to require methods for distinguishing interval from ordinal. In taking up this issue, two different perspectives on educational measurement are contrasted, a metaphorical perspective and a classical perspective. Although the metaphorical perspective is more predominant, at present it provides no objective methods whereby the properties of a vertical scale can be validated. In contast, when taking a classical perspective, the axioms of additive conjoint measurement can be used to test the hypothesis that the latent variable underlying a vertical scale is quantitative (supporting ratio or interval properties) rather than merely qualitative (supporting ordinal or nominal properties). The application of such an approach is illustrated with both a hypothetical example and by drawing upon recent research that has been conducted on the Lexile scale for reading comprehension.

Pre-print, forthcoming in *Journal of Educational Measurement*

Acknowledgements: This work is the culmination of research initially supported by a National Academy of Education Postdoctoral Fellowship funded by the Spencer Foundation, and then continued under the auspices of a grant funded by the Carnegie Corporation. I would like to thank the reviewers of this manuscript for the constructive comments that led to improvements in the underlying argument.

Introduction

Intuitively, the concept of growth does not seem terribly complicated. When I was a child I used to visit my grandmother in Austria once a year during the summer. Upon seeing me, she would invariably exclaim "Look how much you have grown!" And so we would. She would march me over to the designated spot in her hallway and I would stand straight while she marked my height against the wall. Then we would compare the most recent mark to the mark that had been left the summer before using a ruler. At that point my grandmother's qualitative observation could quantified with respect to the number of centimeters I had grown over a one year time span. This little ritual captures what most people have in mind when they speak of "measuring growth." It begins with a qualitative assessment over at least two points in time, ("you look taller to me") and becomes measurement after a magnitude can be established relative to an agreed upon standard unit ("you have grown 4 centimeters since I last saw you").

Where complications arise is in the shift from measuring growth in height to measuring growth in knowledge, skills and abilities (i.e., learning). It might stand to reason that the latter activity should involve the same basic elements as the former: assessments that have been made at two points in time through some standardized procedure, and the use of a common scale to transform qualitative observations into quantitative magnitudes. When measuring what students have learned from grade to grade while in school, if there is a psychometric analog to the ruler then it would appear to be the vertical[1] score scale. For example, according the technical manual that accompanies CTB-McGraw Hill's *TerraNova* Test Battery, the vertical scale "can be viewed as a developmental continuum…scale scores are units of a single, equal-interval scale applied across all levels of TerraNova regardless of grade or time of testing. (CTB-McGraw Hill, 2001, 322)". When invoking

[1] Such scales are also commonly described as "developmental" score scales. I use the more neutral term "vertical" throughout.

an argument-based approach to test validation (Kane, 2006), such a claim would appear to be a central warrant that would need to be supported empirically. This is because the purpose of a vertical scale is to facilitate the measurement of growth in student learning. This growth is to be expressed in terms of absolute changes in magnitude. If the same growth magnitudes have different interpretations as a function of a student's starting point, this threatens the validity of intended test use. Hence it follows that a robust validity argument in support of vertically scale tests would require evidence that the resulting scale has equal interval properties.

When one has a physical referent available, it is often easy to show that the distinction between an interval and non-interval scale is important. Consider the following example to illustrate the point. According to the National Center for Health Statistics, the height of an American adult male is, on average, 5.4 inches greater than the height of an American adult female. In 2009, while teaching a doctoral seminar in a class with 7 female and 5 male students, I decided to find out whether the average difference in height among the males and females in this class was close to 5.4. To accomplish this, I created two different measuring sticks ("A" and "B") illustrated in Figure 1.

--Insert Figure 1 about here--

The reader will note that while both measuring sticks have the same total number of units (12), only in stick A do the units represent a standard sequence (in this example, inches). The first stick can be used to measure differences in length in terms of magnitudes that are intrinsically meaningful because any given object is measured as the ratio of that object's length to the designated standard unit. So using distinctions in scale properties first introduced by Stevens (1946), stick A represents a *ratio* scale. When differences are taken between two measures based on a ratio scale, the numbers that result will have *interval* properties: a difference of X units will have the same intrinsic meaning no matter which two measures were the basis for that difference. In contrast to stick A, stick B can only be used to rank objects according to their length because the units on stick B were purposefully

chosen so that they would have no consistent meaning. Thus, the numbers that result from application of stick B represent an *ordinal* scale. The differences between two objects with lengths measured using stick B are best not interpreted as having equal interval properties, as became readily apparent when students were asked to measure one another with each stick. For the 12 students combined, the use of sticks A and B resulted in average measurements of 68.2 and 70.6 units respectively. When average height differences were measured using stick A, the result was 5.5—very close to the value reported by the National Center for Health Statistics. When measured using the ordinal stick, the result was 21.2. When the differences were expressed in effect size units after dividing each difference by the overall standard deviation (SD) in heights as measured by each stick, use of the ordinal scale (stick B) relative to the interval scale (stick A) inflated the difference in male and female heights by 0.4 SDs (1.9 vs. 1.4). As this example shows, in the physical sciences, the practical consequence of performing arithmetic computations on a numeric scale with ordinal properties relative to one with ratio or interval properties is significant. Is it less so in the social sciences? When a vertical score scale is used to communicate the growth of a student from grade 3 to grade 4, how does one know whether the difference observed is more akin to the 5.4 found using stick A or the 21.2 found using stick B?

This was the question recently posed by Ballou (2009) in the context of value-added statistical models. Although not all value-added models require the availability of tests on a vertical scale, they do implicitly assume that test scores have equal-interval properties. In reviewing the psychometric research literature on this issue, Ballou pointed to many conflicting answers: "There are some psychometricians who consider theta to be intervally 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…Is the IRT [Item Response Theory] ability trait measured on an interval scale or not? Indeed, how does one tell?

(Ballou, 2009, 356)" This question can only be addressed if one is willing to wrestle with some profound philosophical issues regarding the meaning of educational measurement. In doing so, one quickly encounters a critique of modern psychometrics in the form of a series of publications over the past decade by Michell (1997, 2000, 2004, 2008a, 2008b). In short, Michell has argued that the field of psychometrics represents a "pathological science" because an assumption is routinely made about the quantitative nature of what is being measured without putting this assumption to empirical test—or even recognizing that the assumption has been made at all.

A first purpose of this paper is to demonstrate a methodological approach, rooted in a classical conception of measurement, that could be applied empirically to validate (or invalidate) the use of vertically scaled tests to measure growth. It is in the context of vertical scaling that the distinction between quantitative and qualitative, ordinal and interval, can be expected to have the most dramatic practical consequences, and as noted above, establishing this distinction is critical to any serious attempt at test validation. This is because in contrast to the score scales created for tests administered at a single point in time, which are often only used to rank students or to make predictions about a student's likelihood of answering a given item correctly, the raison d'être of the vertical score scale is to measure of growth in student learning in terms of changes in magnitude. A second purpose of this paper is to compare a classically oriented approach to establishing a vertical scale with the more pragmatic approach typical in mainstream psychometrics (at least as practiced in the American testing industry). Under this approach, the term measurement is understood and used metaphorically, and in this sense, the premise of Michell's critique may not apply. But embracing the "measurement as metaphor" perspective can lead to scenarios in which it becomes difficult, if not impossible, to establish whether one vertical scale is in some sense "better" than another. I argue that irrespective of one's (perhaps tacit) philosophical orientation towards educational measurement, the

science behind vertical scaling will only improve to the extent that explicit criteria can be established for the validation activities that accompany and follow the creation of a vertical scale.

Early Arguments over Vertical Scale Interpretations

Confusion over the interpretability of scores deriving from a vertical scale can be traced back to an invited address given by H. D. Hoover at the annual meetings of the American Educational Research Association (Hoover, 1984a; 1984b). Hoover's address was primarily intended as a defense of the use of the grade-equivalent metric to represent trends in growth across grades, a practice that had a longstanding history associated most notably with the Iowa Test to Basic Skills (see Peterson, Kolen & Hoover, 1989 for a detailed description of "Hieronymous" scaling). In the process, Hoover had taken issue with the claim that the scale scores resulting from the application of Thurstone's method of absolute scaling (Thurstone, 1925), or the application of the more recently implemented methods based on the use of Item Response Theory (IRT; Lord & Novick, 1968), were somehow preferable to grade-equivalents as a theoretical basis for subsequent arithmetic computations because they possessed interval properties that grade-equivalents did not. Hoover contrasted the patterns of growth for three vertical scales created in the domain of English Language Arts. Two that had been created using the Thurstone approach, and one that had been created using the Three Parameter Logistic Model (3PLM; Birnbaum, 1968). In all three cases, Hoover was able to point to published claims by the test developers that the resulting scales were "equal-interval" (Hoover, 1984a, pp. 9-10). Yet when Hoover examined the grade to grade growth patterns in reading comprehension implied for students at the 10^{th}, 50^{th} and 90^{th} percentiles on each test, he found that the results led to conflicting interpretations about student development. On the two Thurstone-based scales, there was evidence of rapid growth in the elementary school grades (2-5), slower growth in the middle school

grades (6-8), and then more rapid growth again in the high school grades (9-12). Furthermore, on both scales students at the 90th percentile of the score distribution in a previous grade appeared to grow significantly faster than students at the 10th percentile in the previous grade. In contrast, for the scale created using the 3PLM, after the elementary grades there was a dramatic deceleration of growth for all students with the apparent exception of those at the 10th percentile, for whom growth continued at a rapid clip up until high school. In short, when using the Thurstone approach, the variability in scores increased over time as higher-achieving students appeared to acquire new skills and master new content more rapidly than lower-achieving students. According to the more recently implemented IRT approach, it was the opposite, with the variability in scores decreasing over time as lower-achieving students appeared to catch up to their higher-achieving peers.

Hoover was not alone in his skepticism about the interval properties of vertical score scales (see Philips & Clarizio (1988), Camilli (1988), Clemans (1993), and Camilli, Yamamoto, & Wang (1997)). A common thread to these critiques was the observation that the growth trends implied by the vertical scales under examination were counterintuitive. Not only did the score scales appear to "shrink" in a manner than had never been observed previously, but by a dramatic order of magnitude. Given that existing theory (Hoover cited Anastasi (1958)) and intuition supported the opposite trend, this led to a prevailing sentiment that the growth trends being observed were at least in part an artifact of either the data collection design, the use of IRT, or both.

In a series of publications, Wendy Yen and George Burket, who had been responsible for the vertical scaling of test batteries under critique (the California Test of Basic Skills and the California Achievement Test), defended the use of IRT to create the vertical scales. In Burket (1984), Yen, Burket & Fitzpatrick (1995/1996) and Yen (1986), the IRT approach was defended primarily on the grounds that it represented an improvement over the Thurstonian approach. In Yen & Burket (1997), evidence from a simulation study was presented to argue that to the extent that the achievement

construct is unidimensional and the true scale has constant variance across grades, there is nothing inherent to the use of the IRT that would lead to scale shrinkage as an artifact. On the other hand, Yen (1985) had previously demonstrated through simulation that a violation of the assumption of unidimensionality *could* theoretically lead to scale shrinkage. Since the potential for violations of unidimensionality is quite plausible for a scale spanning 12 grades, this would seem to present a critical problem, and one would have expected scale shrinkage to be a rule rather than an exception in subsequent vertical scales created after the early 1980s. Yet by the mid-1990s Yen & Burket (1997) had noted that the dramatic scale shrinkage evident in early IRT-based vertical scalings was no longer evident in the later editions of these tests, which showed "minimal scale shrinkage or modest scale expansion, depending on the subtest." (Yen & Burket, 1997, 307). A similar finding was reported in a study by Williams, Pommerich & Thissen (1998).

To date, no satisfactory explanation has been given in regard to the anomalous growth trends found on the tests that precipitated Hoover's critique in 1984. In hindsight, a remarkable aspect of the defense of vertical scaling offered by Yen and Burket in their publications was that at no point did they seem interested in arguing that the approach produces a scale with equal-interval properties, even though this was the proposition at the crux of the critiques written by Hoover, Philips & Clarizio, and Clemans. For example, while Yen, Burket & Fitzpatrick (1995/1996) responded quite forcefully to many of the specific elements of the Clemans critique at no point did they respond to the central issue he had raised: when and under what conditions does a vertical scale have interval properties? And if it does not have interval, but only ordinal properties, then how is it useful?

What is Measurement?

In order to resolve whether it is possible to measure growth in student ability with vertical scales that possess equal-interval properties, we can begin by revisiting (and slightly reconceptualizing) a framework established by Michell (1986) in which distinctions are drawn between what it means to measure something. Michell focused on three theories of measurement he referred to as operationalism, representationalism and classicism. To this mix I will add instrumentalism. When motivated by a belief in operationalism or instrumentalism, the notion of measurement in education is best viewed as metaphorical. When motivated by classicism, the notion of measurement in education can be viewed as the act of distinguishing quantity from quality. I will refer to the operationalist and instrumentalist perspectives as "metaphorical" conceptions of measurement. A key distinguishing feature is that under metaphorical conceptions of measurement, the assumption that a scale has interval properties cannot be directly falsified; under the classical conception, it can.

Metaphorical Conceptions of Measurement

Operationalism is typically attributed to the writing of Bridgman (1927), and summarized by the slogan "In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations (p. 5)." In this sense, once numbers have been attached to objects and the resulting variable has been named, it has been "operationalized"—that is, is has been operationally measured. From an operational school of thought test scores are measurements "because they are reasonably consistent numerical assignments that result from a precisely specified operation" (Michell, 1986, 404).

Under instrumentalism (Duheim, 1954), a test score is a "measure" to the extent that it is useful, where utility in this context depends upon whether a measure can successfully classify and

predict observational statements (Niiniluoto, 2011). One of the best known modern day examples of the instrumentalist school of thought can be found in the writings of Stephen Toulmin (1958), work that has had considerable influence on contributions to test validation theory (Kane, 2006). The early influence of instrumentalist thinking on psychometrics is apparent in the opening chapter of Lord & Novick's *Statistical Theories of Mental Test Scores*:

> At various times in this book, however, we shall treat a measurement as having interval scale properties and the theory underlying it yield only a nominal or, at best, an ordinal scale....*from a pragmatic point of view, the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale* [emphasis added]. If we construct a test score by counting up the correct responses and treating the resulting scale scores as having interval properties, the procedure may or may not produce a good predictor of some criterion. *To the extent that this scaling produces a good empirical predictor the stipulated interval scaling is justified* [emphasis added] (pp. 20-21).

The conception of measurement as a metaphor calls to mind a psychometrician who is agnostic as to the existence and structure of one or more latent variables that may or may not underlie a test-taker's sum score so long as the resulting score or score transformation can be shown to be the result of a thoughtful operationalization (e.g., follow from the process of sampling and coding items systematically) and/or useful (e.g., predictive of some external criterion).

The Representational and Classical Conceptions of Measurement

In representational theory, measurement occurs through the process of mapping empirically observable, qualitative phenomena into numerical relationships. The central principle is that measurement concerns the numerical representation of empirical facts. According to Michell, the theory can be traced most directly back to the writings of Stevens and Suppes (Stevens, 1946, 1951; Suppes, 1951; Suppes & Zinnes, 1963). It was Stevens who provided the definition of measurement that has become most ubiquitous in the social sciences: "Measurement is the assignment of numerals to objects or events according to rule" (Stevens, 1946; Michell, 1999). It is interesting to note that this definition, when broadly interpreted (which appears to have been Stevens's intent) and taken out

of historical context, is consistent with the meaning of measurement under operationalism. What most distinguishes representationalism from operationalism are subsequent efforts to undergird Stevens's definition by formalizing the necessary and sufficient conditions (i.e., axioms) that would need to hold before it would be deemed sensible to "assign" numbers to any given empirical relational system (c.f., Krantz et al., 1971). Depending upon which axioms could be satisfied, the resulting numerical relational system could be distinguished with respect to the ratio, interval, ordinal or nominal categories and corresponding admissible statistical procedures that Stevens had popularized[2].

The classical theory of measurement predates representationalism, and can be traced back to ideas held by the Aristotle and Euclid. Under the classical theory, measurement is nothing more nor less than the assessment of quantity. As defined by Michell, "a quantity is a class of properties (such as length) or a class of relations (such as temporal durations), the elements of which stand in additive relations to one another rich enough to sustain numerical ratios" (Michell, 1999, 26). Therefore, measurement is the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity. For the classicist, measurement proceeds by hypothesizing the existence of an attribute of some object, and then seeking to test this hypothesis through experimentation. Unlike representational theory, there is no "assigning" of numerals to objects; objects with quantitative attributes are assumed to exist, and it is the objective of the measurer to discover them. Examples of research on measurement in psychology consistent with the classical perspective would include Thurstone (1927), Coombs (1950), and Rasch (1960). From the classical perspective, there is no such thing as an ordinal (or nominal) "measurement" because an ordinal scale is not quantitative, and the process of measurement is all about going from the qualitative to the quantitative.

---

[2] This motivation is evident in the first chapter of the *Foundations of Measurement* series when Krantz et al write "Stevens has not provided any argument that the procedure of magnitude estimation can be axiomatized so as to result in a ratio-scale representation; he has neither described the empirical relational structure, the numerical relational structure, nor the axioms which permit the construction of a homomorphism" (1971, p. 11)

Representationalism and classicism are distinct traditions with different philosophical perspectives on measurement. Classicists are wedded to scientific realism, while representationalists tend to be skeptical of any meta-physical reasoning about latent variables. They are grouped together here only because for both camps there is an empirical method available to evaluate whether data have a structure that would support an interval over an ordinal scaling. This empirical method is the theory of conjoint measurement[3], which I describe in more detail shortly.

Contrasting Different Conceptualizations of Measurement to Growth Interpretations from Contemporary Vertical Scales

The advent of the No Child Left Behind Legislation of 2002 and the subsequent expansion of state-level testing across grades 3 through 8 has led to a mushrooming of state-specific vertical scales in math and reading. Nationally, the two predominant test contractors that have been responsible for the development of state-specific vertical scales have been CTB-McGraw Hill (CTB) and Harcourt Educational Measurement (Harcourt)[4]. This is in large part because CTB and Harcourt have had a longstanding history as developers of vertical scales. Their respective commercial test batteries, the *TerraNova* and the *Stanford Achievement Test*, were created using nationally representative samples of American students and a common item nonequivalent groups linking design (Kolen & Brennan, 2004; Briggs & Weeks, 2009). A majority of states with vertical scales in math and reading have

[3] In my presentation of additive conjoint measurement that follows I draw upon the classically oriented presentation of the approach found in Michell (1990). That is, I assume that latent attributes and numbers exist a priori and that the purpose of measurement is to discover and describe quantitative structure numerically. For a primer cast in the language of representationalism, see Borsboom (2005) and Kyngdon (2008). The most complete presentation is found in Krantz et al, 1971. While the theory of conjoint measurement was formulated within a representational framework, Michell (1990; 1999) has shown that application of the axioms—in particular, cancellation—is also compatible with the classical theory of measurement. In this sense, while the representational and classical theories are philosophically incompatible, the theory of conjoint measurement serves as a bridge between the two.

[4] In 2008, Pearson Educational Measurement acquired Harcourt. So states that had previously contracted with Harcourt became Pearson clients. However, the vertical scale scores that were the basis for the results that follow derive from technical reports that were written by Harcourt staff, so I reference Harcourt rather than Pearson.

established those scales by contracting with CTB or Harcourt and then embedding TerraNova or Stanford Achievement Test items into their state-specific tests. The parameters for these items are treated as known, calibrated together with unique items in an IRT model, and this, in principle, serves to anchor the scale of the state tests to the underlying vertical scale from which the embedded items originated.

For each of 16 states with a vertical scale during the 2007-08 school year, an implied growth trajectory in reading[5] can be formed by comparing mean scale scores across grades 3 through 8. An effect size metric is sometimes used to depict grade to grade gains as a proportion of the average standard deviation of the scale scores across adjacent grades. In the present context, expressing grade to grade gains as effect sizes makes it possible to compare patterns of growth across states in the same plot, as can be seen in Figure 2.

--Insert Figure 2 here--

The large dark circles in Figure 2 represent the average effect size across states, the bars extending from these circles represent the SD across states. The light dots and lines represent the effect sizes and implied growth trajectories for each of the 16 states. What stands out in this figure is the considerable variability in growth patterns within and across grade pairs. Mean growth in student performance for any pair of adjacent grades ranges from a low of .30 SDs (grade 3 to 4) to a high of .65 SDs (grades 5 to 6).

From the metaphorical perspective, the variability observed in Figure 2 should come as little surprise because each test, in principle, is a uniquely operational or instrumental measure of "reading ability." Grade to grade growth only has meaning as a function of the average differences in the common items answered correctly for students in the lower and upper grades of any pair of adjacent grades. The magnitude of growth that is observed will depend largely upon the developmental and

[5] Similar plots have been made for math vertical scales, but have been omitted due to space constraints. They are available from the author upon request. For details see Dadey & Briggs (2012).

instructional sensitivity of the common items that have been selected. So unless all states were using the same common items, there would be no reason to expect the patterns of growth for students from two different states to look the same—even if the students in both states had comparable demographic backgrounds and had received comparable instruction from a common curriculum. To complicate matters further, if the vertical scales have not been maintained from year to year using the same horizontal equating design (i.e., common items linking scores across the same grades in different years), this would also lead to further differences in operationalized growth interpretations across the grades of the vertical scale.

In contrast, recall that from a classical perspective, measurement is the process of turning qualitative observation into a quantitative relationship via testable hypotheses. In this sense, the results shown in Figure 2 are notable and somewhat surprising, because each state's test should be interpreted as a measure of the same latent psychological attribute ("reading ability"), an attribute that has been hypothesized to be quantitative. If the hypothesis were true, and each state's assessment system could be said to have produced measures of reading ability, then the observed variability across grades and states would be something of scientific interest in the same way that variability in effect sizes is a key source of interest in the meta-analytic literature. The next step would be to look for substantive factors that would explain why students' growth in reading comprehension from grades 3 to 4 in one state is, on average, twice as much as the growth observed for students in another state. A competing or complementary reaction would be to question whether the quantity hypothesis for the psychological attribute of reading ability is plausible. If it can be rejected empirically, then there would little point in attempting to interpret the differences in growth magnitudes for any pair of grades across states, any more than it would be sensible to interpret the magnitude of mean differences in height between makes and females based on the use of stick B in Figure 1.

In making this contrast between the metaphorical and classical orientations to measurement, I will not argue that one is superior to the other in terms of ontological coherence[6]. What I do wish to argue is that the metaphorical conceptualizations do not appear to lend themselves to empirical validation in the context of vertical scaling. If the claim of interest is that test scores placed onto a vertical scale can be used measure growth, then the warrant for this claim is that the scale has interval properties. I can see no way to establish a backing for this warrant when the act of educational measurement is metaphorical. When adopting this conceptualization, distinctions between interval and ordinal scales are either meaningless a priori (operationalism), or meaningless in the absence of some external criterion for utility (instrumentalism). Since there is, to my knowledge, no such criterion available for growth in student achievement, if a test-maker were to establish two different vertical scales for the same state using, for example, two different IRT models, there would be no way to evaluate if one approach led to growth interpretations that were more valid than the other[7]. This strikes me as unacceptable science. The same problem does not emerge when a vertical scale has been established in a manner consistent with a classical conceptualization of measurement, as I illustrate in what follows.

Using the Axioms of Additive Conjoint Measurement to Evaluate Scale Properties

Arguably the most important contribution of measurement theorists in the representational tradition has been to provide a framework whereby the hypothesis that a variable has quantitative

---

[6] For some insight on this issue see Borsboom, 2005; Mislevy, 2008; 2009; Michell, 2008b; Dooremalen & Borsboom, 2009.

[7] Some might be tempted to argue that this could be settled by choosing the model with the best fit to the data for any given grade-specific test (c.f., Skaggs & Lissitz, 1986). If this where a criterion for tests comprised of dichotomously scored items, an IRT model such as the 3PLM would usually for the data better than the Rasch Model. But this in itself does nothing to establish whether linking multiple tests vertically leads to a scale with interval properties. In fact, the superior within grade fit of the 3PLM relative to the Rasch Model may well constitute evidence *against* an interval scale interpretation.

structure can be falsified—even if that variable is latent. This framework is known collectively as the *theory of conjoint measurement*, and the simplest version of it—additive conjoint measurement—was first introduced by Luce & Tukey (1964)[8]. In the most general sense, conjoint additivity implies that two variables can be scaled such that their additive combination forms a third variable. A famous example of this (Krantz et al, 1971; Andrich, 1988; Michell, 1999) is the relationship between force (f), mass (m) and acceleration (a) in Newton's second law of motion. After taking logarithms, A = F + M where A = log(a), F = log(f) and M = -log(m). The remarkable result of additive conjoint measurement is that even if distinctions between different values of force and mass could only be made in terms of order, if the values of acceleration that resulted from their combination could be shown to follow certain rules, then it could be proven that all three variables have quantitative structure. Although the mathematical proofs of additive conjoint measurement can be hard to follow, the key conceptual features of the theory and its usefulness in testing hypotheses about the quantitative structure of a latent variable are easy to illustrate. I do this first in the abstract and then follow this with a specific example pulled from Angoff's (1971) discussion of the difficulties of establishing that a scale has equal-interval properties.

Testing the Quantity Hypothesis

Assume the existence of two variables, *X* and *Y*. For each variable respectively there are *J* and *K* observed values, $\{x_1, x_2, \ldots, x_j, \ldots, x_J\}$ and $\{y_1, y_2, \ldots, y_k, \ldots, y_K\}$. While it is not necessary to assume a priori that the values of each variable are ordered, we will do so here to simplify the illustration. Given this, we can say that $x_j \leq x_{j+1} \leq x_{j+2} \leq \ldots \leq x_J$ and similarly $y_k \leq y_{k+1} \leq y_{k+2} \leq \ldots \leq y_K$. The theory of additive conjoint measurement is premised upon a situation in which a third variable, *Z*, can be expressed as a function of *X* and *Y* such that $Z = f(X,Y)$. In other

[8] Although as an anonymous reviewer of this manuscript pointed out, the seeds for this theory were already visible in an even earlier paper by Cliff (1959).

words, values of Z are observed empirically as a consequence of different combinations of *X* and *Y*. When it can be demonstrated that the order relationships amongst values of Z satisfy certain axioms, it follows that *X*, *Y*, and *Z* have been *conjointly* established as quantitative variables, with *f* as a noninteractive function (e.g., $Z=X+Y$). The key axioms of additive conjoint measurement are *cancellation*, *solvability* and the *Archimedian condition*. The axioms are easiest to visualize when presenting a subset of a conjoint system for two variables *X* and *Y* as a 3 by 3 matrix as in Figure 3.

--Insert Figure 3 about here--

The solvability axiom essentially says that there must be enough combinations of *X* and *Y* to produce any desired value of *Z*. The Archimedean condition ensures that the difference between two values of *X* or *Y* will never be infinitely larger than any other two values of *X* or *Y*. While neither solvability nor the Archimedean condition can be directly falsified, Michell (1990) has argued that evidence in support of them can be established indirectly to the extent that the cancellation axioms can be satisfied. I describe this process in detail to give the reader some sense for what is required to evaluate the cancellation axioms. For any $n \times n$ conjoint matrix, there will be $n-1$ cancellation conditions that can be tested. In the case of the 3 by 3 matrix shown in Figure 3 there are two: single and double cancellation. Single cancellation (sometimes referred to as the independence assumption), asserts that the ordering of the values of *Z* (cells in the matrix) remain the same when the values of *X* (the rows) are changed and the value of *Y* (the columns) is fixed, and vice-versa. If single cancellation can be established, the main diagonal of the matrix shown in Figure 3 must have an ordering such that $z_{33} > z_{22} > z_{11}$. The axiom of double-cancellation is used to establish the relative orderings of the off-diagonal cells. Under double cancellation,

if $z_{12} \geq z_{21}$
and $z_{23} \geq z_{32}$
then $z_{13} \geq z_{31}$.

The double cancellation hypothesis is illustrated by the arrows in Figure 3, where the two solid arrows represent the antecedent conditions, and the dashed arrow represents the consequence that must follow. The consequence of double cancellation comes from the fact that if the variables *Z*, *X* and *Y* form a conjoint system, then it must be the case that we can express Z as an additive combination of *X* and *Y*, such that $z_{jk} = x_j + y_k$. Given this, it follows that the conditional relationship above, the antecedents can be re-expressed as

$$x_1 + y_2 \geq x_2 + y_1 \quad (1)$$

and

$$x_2 + y_3 \geq x_3 + y_2. \quad (2)$$

Summing (1) and (2) produces

$$x_1 + y_2 + x_2 + y_3 \geq x_2 + y_1 + x_3 + y_2. \quad (3)$$

Since x_2 and y_2 are common to both sides of equation 3, they cancel (hence the term "double" cancellation). Recalling again that $z_{jk} = x_j + y_k$, and given that the antecedents in equations 1 and 2 hold, it follows that equation 3 reduces to $z_{13} \geq z_{31}$.

From Theory to Practice

Now we consider a specific example of how the cancellation axioms of additive conjoint measurement could be used to test the hypothesis that a latent variable has quantitative structure with equal-interval interpretations. Angoff (1971) pointed to the equivocal nature of such an endeavor when he wrote "…there is no assurance that equal differences between scores in different regions on the scale of a psychological test represent equal differences of ability" (p. 509). To illustrate the problem, Angoff used an example that had been shared with him informally by Frederic Lord, in which the latent attribute in question was typing ability. Lord had imagined a scenario in which typing ability was operationally measured by the number of words a person could type correctly in a minute. Angoff noted that one might be tempted to conclude that the difference between two people

able to type, respectively, 20 and 30 words per minute, is equivalent to the difference between two typists able to respectively type 50 and 60 words per minute. He then pointed out that such a conclusion would be equivocal because the amount of practice required for a typist to improve from 50 to 60 words per minute would surely be an order of magnitude higher than the amount of practice required to improve from 20 to 30.

--Insert Figure 4 about here--

Interestingly, in the scenario described by Angoff and Lord the hypothesis that typing ability has a quantitative structure *could* in fact be tested under the theory of conjoint additivity. This can be shown by re-expressing Figure 3 in terms of the Angoff/Lord typing scenario. Let the variable *X* now denote the number of weeks of typing practice to which a student has been exposed. Let the variable *Y* denote a task consisting of some number of words a student is given to type in a minute. Consider the conjoint matrix shown in Figure 4 that results from an experiment in which three levels of *X* (1, 2 and 3 weeks) are crossed with three levels of *Y* (20, 30, and 40 words). In other words, in this experiment, a sample of N students is randomly assigned to one of nine possible cells—some students practice typing for one week and then are given a list of 30 words to type correctly; others practice for two weeks and are given a list of 20 words to type correctly, etc. The values observed in each cell would represent the proportion of students in each condition that successfully completed the typing task. If typing ability is to be interpreted as a quantitative variable measured conjointly as an additive function of *X* and *Y*, a necessary condition is that it must be the case that the single and double cancellation axioms of conjoint measurement hold when the results of this experiment are evaluated. In the fictitious results shown in Figure 4, both single and double cancellation axioms would hold, providing provisional support to the hypothesis that typing ability can be measured quantitatively.

It has been well-established that the typical logistic formulation of the Rasch Model[9], $\log\left[\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right]=\theta_p-\delta_i$ is analogous to the sort of situation presented in the theory of additive conjoint measurement because it involves the linear and noninteractive combination of person "ability" (i.e., the rows in Figure 4) and item "difficulty" (i.e., the columns in Figure 4) to predict the log odds of a correct response (Brogden, 1977; Perline, Wright & Wainer, 1979; Wright, 1997; Michell, 2008c). The left side of the Rasch Model equation above represents the log odds ("logit") of a correct item response and the right side of the equation consists of parameters for a person's ability (θ, indexed by the subscript *p* for each respondent) and the item's difficulty (δ indexed by the subscript *i* for each test item). It is in this sense that one can attempt to justify the logit scale that results from the application of the Rasch Model as possessing interval properties—if the model can be shown to adequately fit the data at hand. As Ballou (2009, 358-360) has noted, under conjoint measurement, the interval property of a scale comes from the ability to express differences between any two levels of one variable (i.e., *Y*) in terms of a designated reference interval (i.e., a standard unit) on the second variable (i.e., *X*). So in the Rasch Model, the person ability scale can be given meaning with respect to a defined interval of the item difficulty scale, and vice-versa. It is important to recognize that in the ideal scenario, the person ability scale does not have equal-interval properties because of some distributional assumption, but through its relationship to the item difficulty scale.

An Example of the Classical Approach in the Context of an Existing Vertical Scale

The Lexile Theory and Scale

[9] In particular, Rasch's (1960) emphasis on the concept of specific objectivity has a clear parallel with the necessary condition of single cancellation in the theory of conjoint measurement.

The research that has been conducted on the Lexile Test Battery for reading comprehension demonstrates that it is possible the create vertically scaled tests with falsifiable scale properties. According to the Lexile theory, the ability of a student to comprehend the meaning of a reading passage is a function of two variables: (1) syntactic complexity, estimated by the ratio of the total number of words to the number of sentences; and (2) semantic complexity, estimated by the average frequency that the words in the passage are used in a text corpus of over 5 million words sampled from a broad range of school materials (Carrol corpus; Carrol, Davies & Richman, 1971). Stenner and colleagues have fine-tuned this theory over many years of empirical investigation (Stenner et al., 2006; Stenner et al., 2010). Much of this research draws upon on the concept of item difficulty modeling (c.f., Stenner, Smith & Burdick, 1983, Fischer, 1983; Gorin & Embretson, 2006), which requires a test developer to hypothesize, in advance, the manipulable variables that would make a student more or less likely to answer an item correctly. In this particular context, the items under investigation are known as cloze items, so-called because they consist of a series questions embedded within a reading passage. At different junctures of the passage a word from a sentence is omitted and the reader is prompted to choose between four options that would "cloze" the sentence.

In an initial exploratory stage of research, Stenner and colleagues (Stenner, Smith & Burdick, 1983) found empirical examples where cloze items taken from the Peabody Individual Achievement Test had been administered and then calibrated with the Rasch Model. Item difficulty was regressed on a collection of up to 50 variables that, along with sentence length and word complexity, consisted of factors such as parts of speech, content classifications of words, number of syllables, etc. Stenner (1996) reports that the estimates of syntactic and semantic complexity were the strongest predictors, by themselves explaining up to 85% of the observed variability in item difficulty.

On the basis of such studies, the Lexile developers established a prediction equation that makes it possible to predict the difficulty of a cloze item *before* it has been administered. The

equation takes the form $\delta_i = a + bLMSL - cMLWF$, where δ_i is the "theoretical logit" for item difficulty, *LMSL* is the log of mean sentence length, *MLWF* is the mean of the log word frequency, and the parameters *a*, *b* and *c* are treated as known constants, having been previously estimated. All else being equal, a cloze item is expected to be more difficult (large positive value) if it is comprised of longer sentences and words that students encounter infrequently in their everyday reading. To the extent that an equation such as the one above can accurately predict item difficulty, it dramatically simplifies the process of creating a vertical score scale because it is no longer necessary to administer common items to students at different grade levels in order to estimate the relevant linking constants.

Because item difficulty parameters are known in advance, it is straightforward to estimate a student's reading comprehension level using the Rasch Model with logit difficulty values known (using estimates from the prediction equation) and ability parameters unknown. To make ability estimates interpretable in a criterion-referenced sense, logit values are transformed into "Lexiles" as a function of two anchor points: the text difficulty from seven basal primers (lower anchor, typical of first grade text) and text from The Electronic Encyclopedia (Grollier, 1986; typical of 12th grade text). A standard measurement unit—a single Lexile—is defined as 1/1000 of the difference in difficulty between these two anchor points. So if a student shows growth of 100 Lexiles from grade 1 to 2, this magnitude has an unambiguous criterion-referenced meaning.

Testing the Quantity Hypothesis

Does reading comprehension have a quantitative structure? If it does, then a gain of 100 Lexiles from first grade to second grade will the same meaning as a 100 Lexile gain from fifth grade to sixth grade: the Lexile scale has an equal-interval property. Stenner (1996) suggests this is the case when he writes: "Measurements for persons and text are now reportable in Lexiles, which are similar to the degree calibrations on a thermometer." There are at least two reasons for being skeptical of such a claim on the basis of the evidence described above. First, we may argue that the theory behind

the Lexile equation is flawed. For example, Stenner et al. (1996) found that the equation did not predict well for reading passages consisting of poetry or non-continuous prose. This limits the generalization of the Lexile scale. Others have argued that reading comprehension is far too complex a construct to quantify in the simple manner implied by the Lexile equation. Second, establishing a linear equation that is strongly predictive of item difficulty does necessarily imply that a quantity hypothesis can be supported. To do so the hypothesis would need to be put to a formal test.

Kyngdon (2008; 2011) demonstrated how the axioms of conjoint measurement could be used to perform such a test. Kyngdon conducted a small scale empirical evaluation of Lexile test data using a probabilistic approach to checking the cancellation axioms of additive conjoint measurement initially proposed by Karabatsos (2001). In doing so, Kyngdon failed to reject the hypothesis that the difficulty of reading items (as hypothesized by the Lexile theory), and the ability of persons (represented by total number of items answered correctly), satisfy an additive relationship that make them jointly quantitative rather than qualitative. Kyngdon also demonstrated how IRT models with a more complex parameterization than the Rasch Model could be expressed and tested with respect to extensions of additive conjoint measurement (e.g., polynomial conjoint measurement).

One criticism of Kyngdon's evaluation of the Lexile is that it involved only a single submatrix that had been drawn from the larger available conjoint data matrix of 39 score groups (rows) by 51 items (columns). As part of a more recent study, Domingue (2012a, 2012b) created and implemented the R package `ConjointChecks` to repeatedly sample 3 x 3 submatrices from a full conjoint matrix and check them against the cancellation axioms of additive conjoint measurement. To do this, Domingue simulated data that would satisfy the axioms probabilistically. He then kept track of the (small) proportion of cells found to violate the axioms in this simulated scenario for use as a normative baseline. Next, using the same Lexile data that was the basis for Kyngdon's study (a 39 x 51 data matrix), he performed checks of the axioms with random samples of 16,800 3 x 3

submatrices. Although Domingue was able to replicate Kyngdon's empirical finding in support of the Lexile when using the same single submatrix, the proportion of cells found to violate the cancellation axioms across the repeated samples of 3 x 3 submatrices were 10 times larger then the proportion detected under the baseline condition (i.e., data had been simulated to fit the axioms). Based on these findings it appears premature to conclude that the latent attribute of reading comprehension—at least as conceptualized with Lexile theory—has a quantitative structure. Knowing that the Lexile equation is highly predictive of item difficulty is only a boon if the grouping of items by difficulty and respondents by total scores can be shown to adequately satisfy the axioms of additive conjoint measurement. Although the Lexile scale appears to have considerable utility as a tool for generating criterion-referenced reading assignments with possible diagnostic advantages, empirical evidence suggests that changes in magnitude along its vertical scale cannot be given an equal-interval interpretation.[10]

A major takeaway from this example is that, in contrast to more pessimistic assertions (Yen, 1986; Cliff, 1992; Zwick, 1992; Ballou, 2008), testing an equal-interval hypotheses is not an impossible or an insurmountably difficult task. The theory of conjoint measurement provides the means by which such hypotheses could be investigated empirically. The approach first suggested by Karabatsos and more recently expanded upon by Domingue (2012) adjusts the axiomatic approach such that it takes measurement error into account, and the R package `ConjointChecks` provides researchers with an open source computational approach for implementing additive conjoint checks that can be readily applied to any matrix of item responses.

[10] One notable source of indeterminacy here is the use of reading passages with multiple-choice cloze items to elicit evidence of reading comprehension. The multiple response options may lead to unanticipated guessing; the nesting of items within passages may increase the dimensionality of the assessment. Would different instrumentation lead to a different conclusion about the structure of the hypothesized latent variable? More research would be necessary to find out.

Limitations

One obvious challenge with the classical approach sketched out above is that it requires test developers to establish hypotheses about manipulable variables that cause items to be harder or easier to answer, and test-takers to be more or less able to respond correctly to them. The Lexile is one of the only large-scale assessments of which I am aware in which at least an item side hypothesis (reading comprehension as a function of sentence length and word complexity) has been made explicit, *and* a research agenda has been undertaken to validate the larger assumption of interval scale properties. Unfortunately, there are no current examples along the lines of the Angoff/Lord illustration presented earlier where the conjoint hypothesis is premised upon an external manipulation of *both* item and person factors. Furthermore, the broader the domain of interest, the more difficult it will be make targeted and testable hypotheses. This would suggest that vertical scales could only be plausibly supported for more narrowly defined latent variables. In other words, it is more conceivable that one might be able to measure growth in a student's ability to add fractions, but not the more broadly defined "ability" to solve mathematical problems.

A second challenge when taking the classical approach is to establish criteria for how close is close enough. Just as the interval properties of a ruler begin to break down as the standard unit gets smaller and smaller relative to the objects of measurement, the same will be true of the measure of a latent variable as the differences in difficulty between items get smaller and smaller. It may be the case that scales can be created for which the equal-interval hypothesis holds—but only for a coarse level of granularity. For example, imagining a score scale ranging from 200 to 800 in increments of 10, perhaps the theory of conjoint measurement could be used to show that a 60 point change from 300 to 360 has the same meaning as the change from 700 to 760, but that measurement error prevents a similar assertion about score changes at various points on the scale that are 50 points or less. In any case, while there is surely no easy solution to the question of how close is close enough, at least the

axioms of additive conjoint measurement provide a criterion against which this can be evaluated. Hence it would be possible to determine, when faced with two competing vertical scales, that one is closer to the interval ideal than the other.

Discussion

Once some philosophical distinctions between different theories of what it means to measure are explicated, it becomes easier to make sense of the seemingly contradictory statements that have been made about the use of vertical scales to measure growth. For example, one can infer that to Yen and Burkett, because measurement has only a metaphorical meaning, the use of an IRT-based approach could be justified by arguing that it represented a superior statistical model to any other alternative. The resulting scales were no more or less equal-interval than any other score scales because such properties can never be internally justified. On the other hand, one can infer that to Hoover, Clemans, and Philips & Clarizio, the driving motivation for creating a vertical score scale using Thurstonian or IRT methods was to measure growth in a classical sense. Given this assumption, it is no surprise to observe their consternation over empirical findings that raised doubts about the plausibility of equal-interval scale properties. A fundamental problem with much of the research literature on vertical scaling is that it is largely premised on a metaphorical conception of measurement, yet communicated to test users through reference to a classical conception (c.f., Burket, 1984, p. 15). A coherent framework for validating a vertical scale can only be established if this contradiction is well-understood. If the distinction between ordinal and interval is to be regarded as meaningless, then the consumers of psychometric products should be placed under no illusions to the contrary.

There are some possible advantages to embracing the classical definition of measurement as a basis for vertical scale creation and validation, and this was illustrated vis-à-vis the research that has gone into the development of the Lexile Framework. In the classical approach, one aspires to measure growth relative to a standard unit with a criterion-referenced meaning. There is a great need and demand for vertically scaled tests because there is a great desire to make absolute statements about differences in the quantity of what students have learned. If this can be accomplished, it greatly simplifies the statistical task of modeling growth over time, because results can be communicated in terms or linear or nonlinear trajectories, which meshes nicely with the intuitive notion parents, teachers and policymakers have when they speak of growth. If it cannot be accomplished, then different statistical methods would need to be used to communicate inferences about growth.

The classical approach requires one to put forward testable and falsifiable hypotheses about the design factors that make items easier or harder to answer correctly, and students more or less able. Even if such assumptions could not be supported when scrutinized against the axioms of conjoint measurement (as was shown to be the case for the Lexile scale), it is hard to imagine that a process that thoughtfully invoked the principles of experimental design in this manner would not lead to stronger and more defensible testing programs (Briggs, . It is in this sense that there may be some common ground to be found in thoughtful renditions of the classical and metaphorical approaches to measurement. For example, in presenting measurement as a "narrative frame" for model-based reasoning and in his applications of "Evidence Centered Design", Mislevy has emphasized the need for assessments that leverage advances in cognitive psychology to form "student models" and "task models." (Mislevy, 2006; 2008) In so doing he focuses attention on some of the same sorts of a priori hypothesizing in that is at the heart of the classically oriented investigation illustrated above. To the metaphorical measurer, the central goal of test development is to elicit actionable evidence about

what students know and can do. One suspects that in many instances the classical measurer would develop tests that would elicit the same sort of evidence.

However, a context where this common ground breaks down is when tests are being designed for the specific use of measuring growth along a vertical scale. When taking the classical approach, there is a clear program of research that could be undertaken to validate this use. The nature of a competing program of validation research under the metaphorical approach has not been explicated and remains an open question. One possibility would be to take seriously the program of research implied by one of the founding fathers of the philosophy of pragmatism, William James. In describing the pragmatic method, James wrote:

> "The pragmatic method in such cases is to try to interpret each notion by tracing its respective practical consequences. What difference would it practically make to anyone if this notion rather than that notion where true? If no practical difference whatever can be traced, then the alternatives mean practically the same thing, and all dispute is idle. Whenever a dispute is serious, we ought to be able to show some practical difference that must follow from one side or the other's being right." (James, 1907)

From a pragmatic perspective, until one can demonstrate empirically that a violation of the quantity assumption (i.e., the "pathology" of psychometricians to use Michell's language) leads to significant practical consequences—for example, the estimated value-added effects of a large number of teachers or schools goes from positive to negative or from large to small—there will be little incentive to invest the time and effort into a research agenda focused on the discovery of psychological attributes that are measurable in a classical sense. In applying the pragmatic method to validate growth interpretation from a vertical scale, the challenge would be to demonstrate that there would be no practical differences to decisions based on these interpretations if the scale was in fact only ordinal and not interval.

In social science research in general, and educational research in particular, there is a tendency to use the term "growth" so loosely that almost any procedure whereby one number is

compared to another would qualify. In the same vein, measurement, when it is defined at all, is typically cast as some version of "the assignment of numerals of objects according to rule", a definition that rules out nothing but the random assignment of numbers to objects as a measurement procedure. The measurement of growth may be understood to be metaphorical by most psychometricians, yet the best metaphors are ones that can be firmly tethered to reality. So long as the term "measuring growth" remains a Rorschach Test, some will view the resulting picture as a work of art, and others will view it as the result of an underlying pathology in need of treatment.

References

Anastasi, A. (1958). *Differential psychology* (3rd ed.) New York: MacMillan.

Andrich, D. (1988). *Rasch models for measurement*. SAGE Publications.

Angoff, W. H. (1971) Scales, norms and equivalence scores. In R. L Thorndike (Ed.). *Educational measurement*, (2nd ed., 508-600). Washington, DC: American Council on Education.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351–383.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Boring, E. G., Bridgman, P. W., Feigl, H., Israel, H. E., Pratt, C. C., & Skinner, B. F. (1945). Symposium on operationism. *Psychological Review*, 52, 241–294

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Briggs, D. C. (2011) Cause or Effect? Validating the use of tests for high-stakes inferences in education. In N. J. Dorans & S. Sinharay (Eds.), *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*. New York: Springer.

Briggs, D. C. & Weeks, J. P. (2009) The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice,* 28(4).

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika, 42*(4), 631–634.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues & Practice*, 3(4), 15-16.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379.

Carroll, J. B., Davies, P. & Richman, B. (1971). *The word frequency book*. Boston: Houghton Mifflin.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment,* 1(4), 329–347.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, 66(1), 27-44.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 186–190.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 145-58.

Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment Research and Evaluation*.

Domingue, B. W. (2012a). Evaluating the equal-interval hypothesis with test score scales. Doctoral Dissertation, University of Colorado, Boulder.

Domingue, B. W. (2012b). Evaluating the equal-interval hypothesis with test score scales. Paper presented at the annual meeting of the National Council for Measurement in Education. Vancouver, B. C.

Dooremalen, H. & Borsboom, D. (2009). Metaphors in psychological conceptualization and explanation. In A. Toomela & J. Valsiner (Eds.), *Methodological Thinking in Psychology: 60 Years Gone Astray?* (pp. 121-144). Charlotte, NC: Information Age Publishing.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press.

Fischer, G. H. (1983) Logistic latent trait models with linear constraints. *Psychometrika*, 54, 599-624.

Gorin, J. S., & Embretson, S. E. (2006). Item Difficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement*, 30(5), 394.

Hoover, H. D. (1984a). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, 3(4), 8–14.

Hoover, H. D. (1984b). Rejoinder to Burket. *Educational Measurement: Issues and Practice*, 3(4), 16–18.

James, W. (1907). *Pragmatism: A New Name for some Old Ways of Thinking*, Cambridge MA: Harvard University Press, 1975.

Kane, M. (2006). Validation. In R. Brennan (Ed.) *Educational Measurement*, 4th Edition (pp. 17-64). Westport, CT: Praeger.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2(4), 389-423.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer Verlag.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. 1: Additive and polynomial representations*. New York: Academic Press

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, 64(3), 478-497.

Kyngdon, A. (2008). Treating the Pathology of Psychometrics: An Example from the Comprehension of Continuous Prose Text. *Measurement: Interdisciplinary Research and Perspectives*, 6.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability.* Addison-Wesley, Reading, MA.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal and Psychology*, 88, 355-383.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge University Press Cambridge, England.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology*, *14*(1), 121.

Michell, J. (2008a). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective, 6*(1), 7–24.

Michell, J. (2008b). Rejoinder. *Measurement: Interdisciplinary Research & Perspective, 6*(1), 125-33.

Michell, J. (2008c). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology, 18*(1), 119.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.) Educational Measurement, 4th Edition (pp. 257-306). Westport, CT: Praeger.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives*, Volume 6, Issue 1-2.

Niiniluoto, I. (2011) Scientific Progress, *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*, Edward N. Zalta (ed.), http://plato.stanford.edu/archives/sum2011/entries/scientific-progress.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 221-262.) New York: MacMillan.

Phillips, S. E., & Clarizio, H. F. (1988). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice,* 7(1), 8–15.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research,* 56(4), 495-529.

Stenner, J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20(4).

Stenner, J. (1996). Measuring reading comprehension with the Lexile framework. In Fourth North American conference on adolescent/adult literacy. Washington, D.C.: International Reading Association.

Stenner, J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are lexile text measures? *Journal of Applied Measurement*, 7(3), 307-322.

Stenner, J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: some corollaries. *Journal of Applied Measurement*, 11(3), 244-252.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. *Handbook of experimental psychology*, 1–49.

Suppes, P. (1951). A set of independent axioms for extensive quantities, *Portuguliae Mathematica*, 10, 163-72.

Suppes, P. & Zinnes, J. L. (1963) Basic measurement theory. In R. D. Luce, R. R., Bush, & E. Galanter (Eds). *Handbook of mathematical psychology.* New York, NY: John Wiley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433-451.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *The Journal of Educational Psychology*, 18, 505-524.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Williams, V. S., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, 35(2), 93-107.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *45*, 51–71.

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

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

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio. *Educational Measurement: Issues and Practice*, 7(4), 16–17.

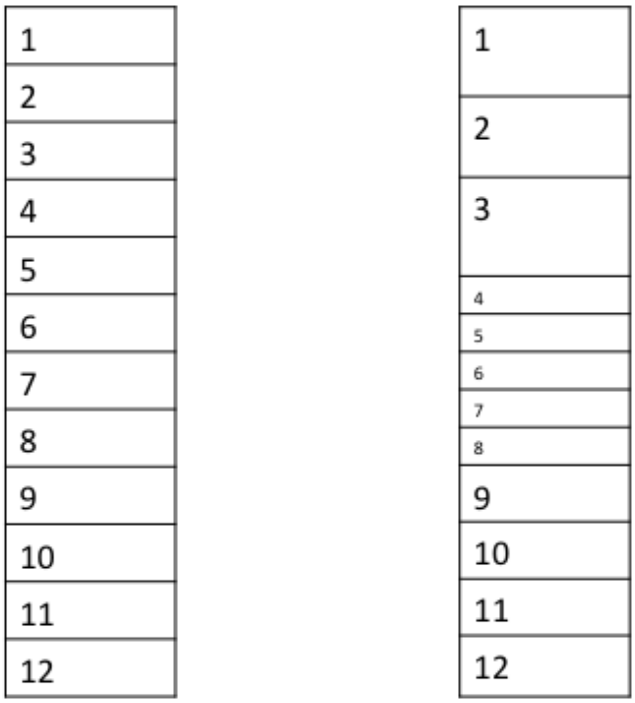
Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement*, 34(4), 293–313.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995a). Response to Clemans. *Educational Assessment*, 3(2), 181–190.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995b). Rejoinder to Clemans. *Educational Assessment*, 3(2), 203–206.

Zwick, R. (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, 205-18.

Figure 1. Two Rulers Used to Measure Length



Stick A Stick B

Figure 2. Reading Growth Trends in Effect Size Metric

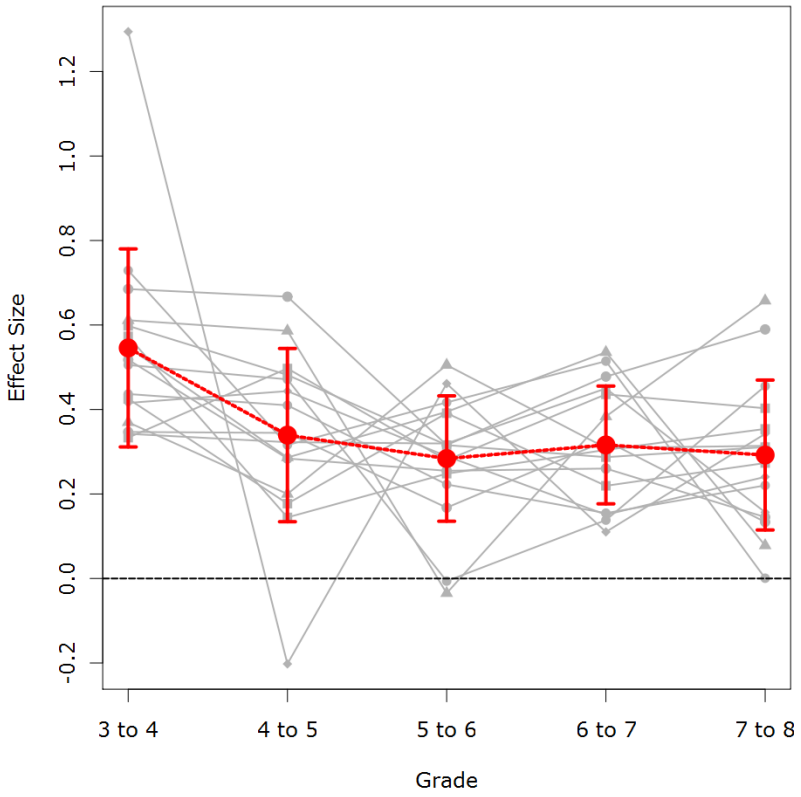


Figure 3. Hypothetical Conjoint Matrix

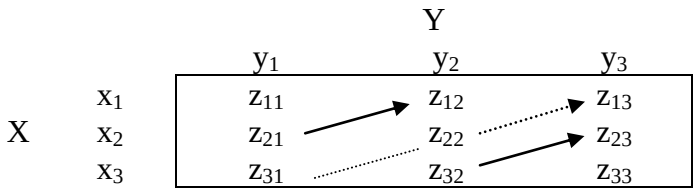


Figure 4. Evaluating the Angoff-Lord Typing Example by Testing Cancellation Axioms

		Difficulty of Task		
		40 words	30 words	20 words
Weeks of Practice	1	.00	.10	.60
	2	.10	.50	.80
	3	.15	.65	.95