THE RASCH MODEL AS A FOUNDATION FOR THE LEXILE FRAMEWORK

Benjamin D. Wright and John M. Linacre

This paper explains the measurement model used to construct the Lexile Framework. (Stenner, 1996)

The Measurement Model

Measurement

"Measurement is the assignment of numbers to material things to represent the relations existing among them with respect to particular properties. The number assigned to some particular property serves to represent the relative amount of this property associated with the object concerned.

The object of measurement is twofold: first, symbolic representation of properties of things as a basis for conceptual analysis: and second, to effect the representation in a form amenable to . . . mathematical analysis." (Eisenhart 1963, p. 163)

In practice, the conceptual and mathematical goals of measurement are achieved simultaneously. Linear measures, *i*.e., measures expressed as numbers with equal-interval units, are the easiest with which to think and also the simplest arithmetically. Indeed, so much does the observer prefer to think in terms of linear measures, that linearity is often imputed to numerals even when such an imputation is undoubtedly erroneous (Wright & Linacre 1989).

An earthquake of 6.0 on the Richter scale, for instance, is commonly thought to be twice as severe as one of 3.0, but, in fact, 6.0 is 10 times as severe as 5.0 and 1000 times as severe as 3.0. Similarly, on a rating scale with a range of 1 to 4, the distance from 2 to 3 is often thought to be the same as from 3 to 4. But, since 4 represents the upper extreme of the observable scale, its implied performance range is infinite. The distance from 3 to any particular rating of 4 is, therefore, unknown and, indeed, unknowable. In general, it is certainly larger than the distance from 2 to 3.

"The very idea of measurement implies a linear continuum of some sort such as length, price, volume, weight, age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind. We judge in a similar way qualities such as mechanical skill, the excellence of handwriting, and the amount of a man's education, as though these traits were strung out along a single scale, although they are, of course, in reality scattered in many dimensions. As a matter of fact, we get along quite well with the concept of a linear scale in describing traits even so qualitative as education, social and economic status, or beauty. A scale or linear continuum is implied when we say that a man has more education than another, or that a woman is more beautiful than another, even though, if pressed, we admit that perhaps the pair involved in each of the

comparisons have little in common. It is clear that the linear continuum which is implied in a "more or less" judgment may be conceptual, that it does not necessarily have the physical existence of a yardstick." (Thurstone 1928, p. 532)
Measures do not exist in nature, however. They must be constructed:

"Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing - a number that indicates how much of this property the thing has, for someone to use for a specific purpose." (Eisenhart 1963, p. 165).

In physical science, the developmental and definitional steps of this "sequence" have long since been established so that what we now think of as a measurement procedure is actually only the last step at which we apply an established technique to our particular situation. The essential prior steps involve:

1) a useful **idea about a variable** to measure on,

2) a **theory of measurement** construction and

3) the **construction and calibration** of measuring instruments.

Physicist Norman Campbell (1920) founds physical measurement theory on the physical concatenation of equal lengths. This theory explains the basis for constructing and calibrating yardsticks and other physical measuring balances. In social science, however, we are still at the beginning of the "sequence of steps". We are still unclear as to which ideas about variables might be worked into useful measures. The empirical bases and techniques of the measurement process are still elusive. Physical concatenations are not helpful. A more abstract, but, nevertheless, equally useful, measurement theory must be devised and applied. After an idea for a variable is developed conceptually, it must be successfully operationalized in terms of particular observations which indicate in a useful way the conceptual property intended. Then a sample of relevant observations must be collected and successfully coordinated by the measurement theory into enduring calibrations of measuring instruments and objective measures of known quality and precision for the objects under examination.

In order to articulate the necessities of such a measurement theory, we will work through a sequence of questions about test scores and the measures they might imply. Each question will bring out a basic concern of test users, a property of test data, or a numerical requirement for useful measurement.

1. Why would we ask a person to answer a reading item?
   To **find out** how well they can read!

2. Why would we ask them to answer more than one item?
   To **make sure** the answers we get are **typical of** the person rather than accidents of the moment!

3. Why do we count right answers and get raw scores?
   To estimate **how much** reading the person knows!

4. But what can a test score tell us?
   Only how many of these particular items this person answers correctly, this time!

5. Is that all we want to know?
   Not really. We ask the items **now**, because now is where we are. But our real interest is not in what the person knows about these items, now. What we want to know is **how much reading ability** the person has **in general**, any time. We want to **predict** their performances on all kinds of reading tasks. We want to **infer** a general measure of their reading ability!

To infer a measure we need a practical way to use the person's specific answers to particular items as the data from which to construct a generally objective measure of how much they know. (Stenner, 1997)

When we look to the future, we realize that the best we can do is to predict what is likely to happen. To manage that, we need a probability formulation which implements a useful connection between the answers we can see right now and what these answers might imply about probable future ability. To infer the future from the present in a systematic way requires a probability model for the answers a person might give to a particular set of items, a model which specifies how a raw score obtained from a few items today could predict how well that person might be able to do on whatever similar items might arise, tomorrow.

Consider a **raw score**. What factors affect its size? To begin with, raw scores vary with the number of items asked, with the number of opportunities to achieve right answers. We can adjust for that source of variation by dividing raw scores by the number of items asked. Moving to a percent correct gets rid of variation in test length. But that is not all that governs the magnitude of a raw score.

Figure 1 shows what can happen when a person at a particular ability level takes five tests all of which measure on the same variable but each of which differs in level and spread of item difficulties. The difficulties D_i of the eight items in each test are marked by their position on the line of the variable. In order to see each test separately we have redrawn the line of the variable five times, once for each test.

The ability B of the person on the measure is also marked by its position on each line so that we can see how this person stands with respect to each test. While each test has a different position on the variable depending on the difficulties of its tests items, this person's position, of course, remains the same on each line. **Figure 1** also shows the scores we would expect this person most often to get on these five tests.

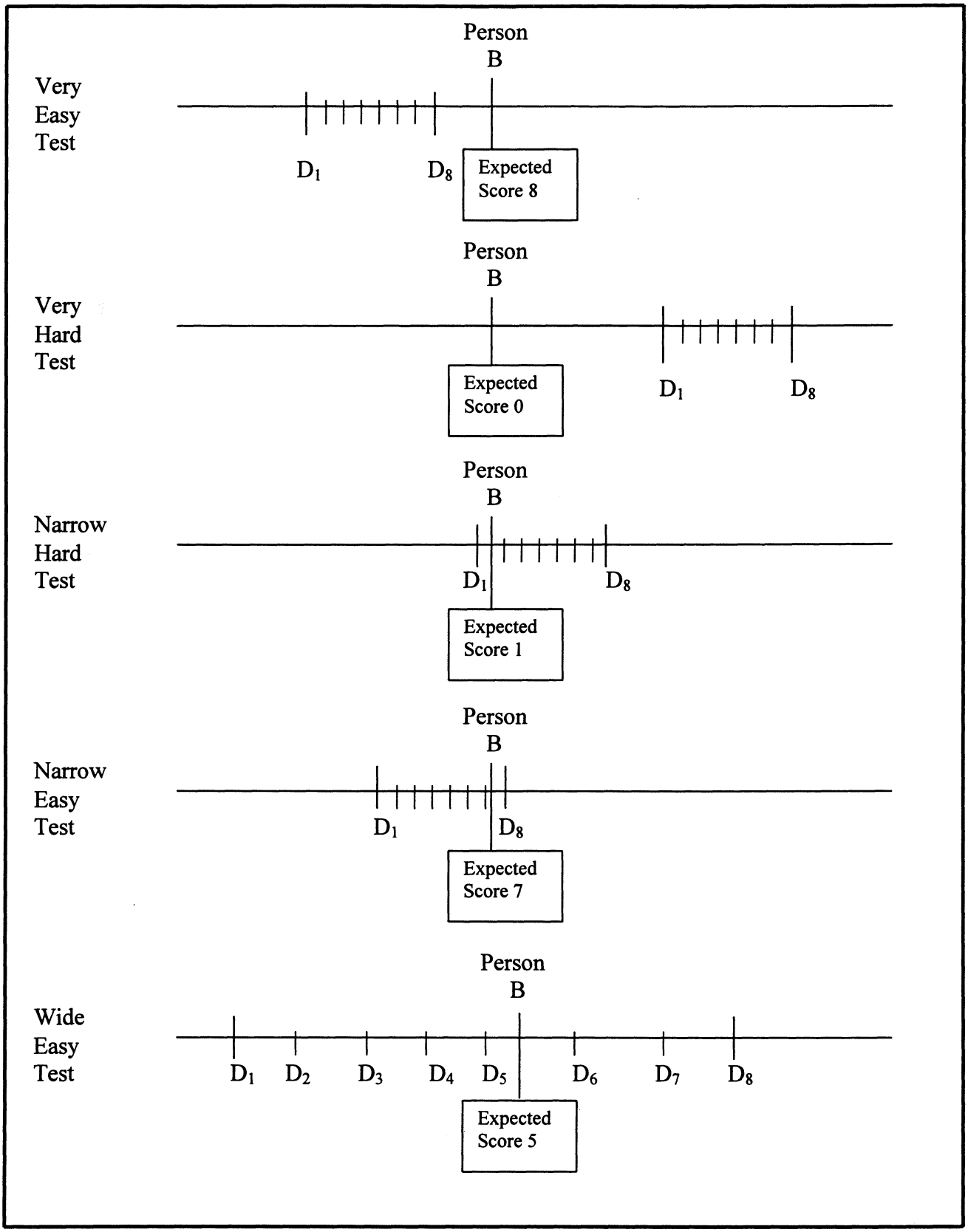

Figure 1. How scores depend on the level and spread of item difficulties.

The **Very Easy Test** has items so easy for this person that we expect a test score of eight. The **Very Hard Test** has such hard items that we expect a score of zero. The **Narrow Hard Test** has seven of its items above the person's ability and only one below. In this situation the most probable score would be a one. The Narrow Easy Test has seven of its items below the person's ability and so the most probable score is seven. Finally the **Wide East Test** is centered at the same position on the variable as the Narrow Easy Test and so has the same average difficulty level. Because of its greater width in item difficulty, however, there are only five items which should be easy for this person and so the most probable score is five rather than seven.

Thus, for one person we have five quite different expected scores: zero, one, five, seven and eight! Although the person's ability does not change, they produce five different scores. Were we to mistake a test score for an ability measure, these data would suggest five different abilities for this one person. Test scores obviously depend as much on the item difficulties as on the ability of the person taking the test.

If the meaning of a test score depends on the difficulties of the test items, then before we can determine a person's ability from their test score we must adjust their score for the effects of the particular test items from which that particular score arose. This adjustment must be able to turn test-bound scores into measures of person ability which are test-free.[1]

The dependence of test scores on item difficulty is a problem with which most test users are familiar. Almost everyone realizes that fifty percent correct on an easy test does not mean as much as fifty percent correct on a hard test. Some test users even realize that seventy-five percent correct on a narrow test does not imply as much ability as seventy-five percent correct on a wide test. But there is still another problem in the use of test scores rather than measures which is usually overlooked.

It is common practice to compute differences in test scores to measure growth, to combine test scores by addition and subtraction in order to compare groups and to add and subtract squares and cross-products of test scores in order to do correlation and regression analysis. But when these simple arithmetic operations are applied to test scores the results are always slightly distorted and can be substantially misleading. Although test scores can estimate the local order of persons' abilities when all persons complete exactly the same test, test scores do not estimate the

---

[1]Unfortunately, with test scores like zero, in which there is no instance of success, and the eight of our example, in which there is no instance of failure, there is no satisfactory way to settle on a finite measure for the person. All we can do in those extreme situations is to observe that the person who scored all incorrect or all correct is somewhere below or somewhere above the operating level of the test they have taken. If we are serious about wanting a finite measure for such a person, then we will have to find a test for them which is more appropriate to their level of ability.

We might be tempted to interpret perfect scores as "complete mastery". But unless the test in question actually contained the most difficult items that could ever be written for this variable there would always be the possibility of other items which were even more difficult. These more difficult items might produce incorrect answers, even with our perfectly scoring person, revealing that mastery was not complete after all.

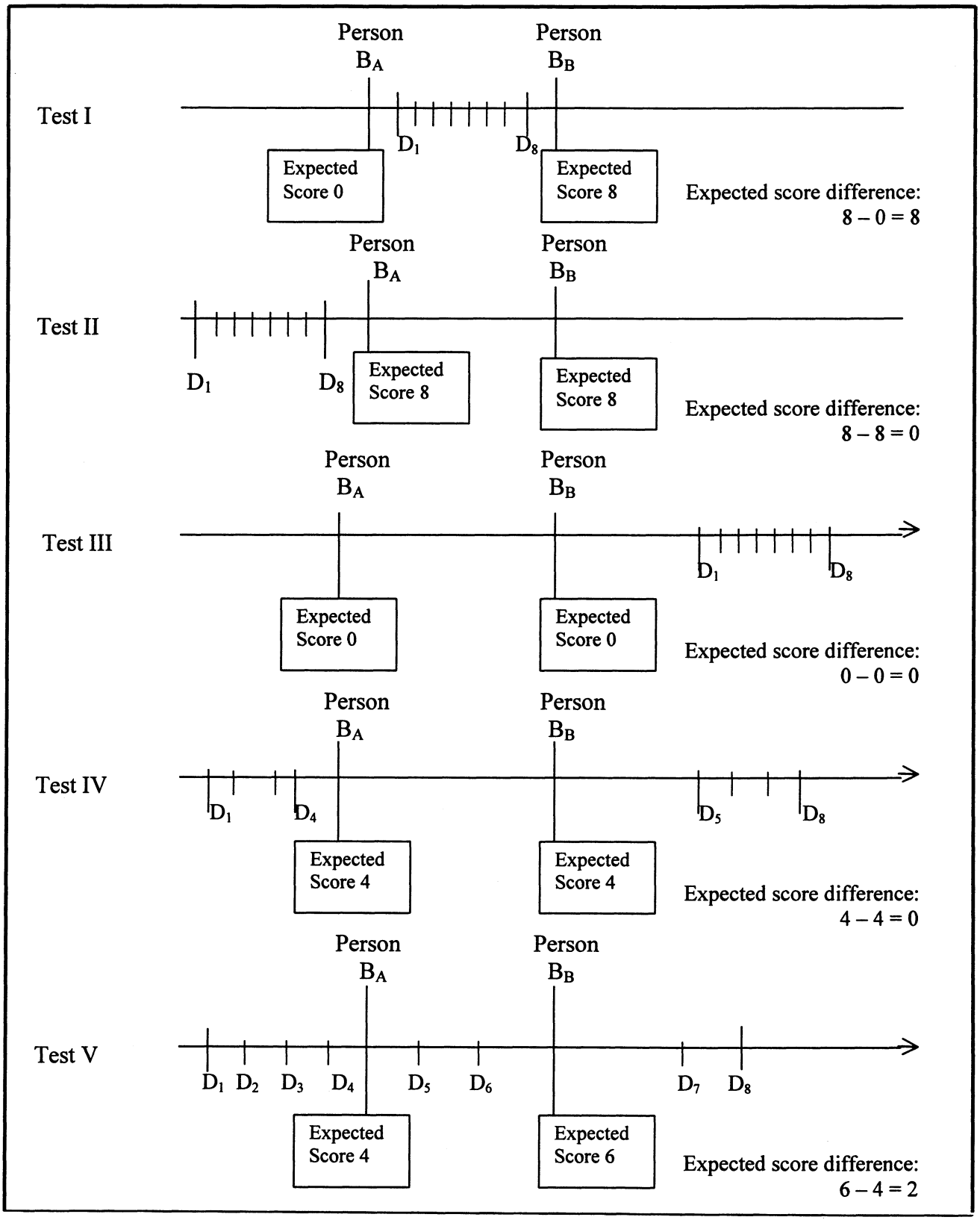

Figure 2: The non-linearity of scores.

spacing between persons satisfactorily. This is because test scores are not linear in the measures they imply and for which they are misused.

In the statistical use of test scores, the numerical consequences of floor and ceiling effects are occasionally recognized. But they are almost never adjusted for. These boundary effects cause any difference of score points to vary in the ability difference it implies over the score range of the test in a way that is specific to that particular test. The distance on the variable a particular difference in score points implies is not the same from a low score to a high score. A difference of five score points, for example, implies a larger difference in ability at the ends of a test than in the middle.

Figure 2 illustrates this problem with test scores. We show two persons with measures B_n and B_m who are a fixed distance apart on the same variable. Both persons are administered five different tests of eight items, all measuring on this variable. The persons' locations and hence their measure difference on the variable remain the same from test to test, but their most probable scores vary widely. This is because the five tests differ in their item difficulty level, spread and spacing. Let's see how the resulting expected scores manifest the fixed difference between these two persons.

Test I has eight items all of which fall between person *n* and person *m*. We expect person *n* to get none of these items correct for a most probable score of zero while we expect person *m* to get all eight items correct for a most probable score of eight. On this test their difference in ability will most often appear as a difference of eight score points. That is as far apart in ability as it is possible to appear on this test.

Test II has eight items all of which are well below both persons. We expect both persons to get scores of eight because this test is too easy for both of them. Now their expected difference in test scores is zero and the usual conclusion will be that their abilities are the same!

Test III has eight hard items well above both persons. Now we expect both persons to get scores of zero because this test is too hard for them. Once again their most probable score difference is zero. Once again the usual conclusion will be that their abilities are the same.

Test I was successful in separating persons *n* and *m*. Tests II and III failed because they were too far off target. Perhaps it is only necessary to center a test properly in order to observe their difference. **Test IV** is centered between person *n* and person *m*. But its items are so spread out that there is a wide gap in its middle into which person *n* and person *m* both fall. The result is that both persons can be expected to achieve scores of four because four items are too easy and four items are too hard for both persons. Even for this test which is centered on their positions, their most probable score difference is zero and we will once again conclude, once again incorrectly, that their abilities are the same.

Test V is both wide and fairly well centered on both persons. It contains two items which fall between their positions and therefore separate them. We expect person *n* to get the four easiest items correct for a most probable score of four. We expect person *m* to get the same four items correct but also the next two harder items because these two items are also below person *m*'s

ability level. Thus on Test V the most probable difference in scores between persons *n* and *m* becomes two. On this test their abilities will usually appear to be somewhat, but not extremely, different.

These simple experiments show us what can happen when we deal with test scores directly without taking into account the particular difficulties of the items involved. Test scores by themselves can show any difference between persons *n* and *m* from none to the maximum allowed by the test, regardless of how different the persons actually are! Persons *n* and *m* appear equally able on Tests II, III, and IV, somewhat different on Test V and as different as possible on Test I. If differences between the test scores of the same two persons can be made to vary so widely merely by changing the difficulties of the items in the test, then how can we use differences in test scores to study ability differences on a variable? The answer is, we can't. Not as they stand.

Test scores always contain a potentially misleading distortion. In order to use test scores to analyze differences we must find a way to transform the non-linear test scores into linear measures which are equated for differences in item difficulties. If we intend to use test results to study growth and to compare groups, then we must use a method for making measures from test scores which marks locations along the variable in an equal interval, linear way.

The magnitudes of **percent corrects** are also governed by the difficulties of the items asked. 50% on a reading test can imply anything from kindergarten to college depending on whether the items are extremely easy or extremely hard. Dependence on item difficulty requires an adjustment to any particular percent correct to liberate it from its inherent item difficulty bias. But "adjustment" involves subtraction. And subtraction requires linear numbers, on an interval scale, along which any given difference remains fixed no matter where on the scale it occurs.

A 6 inch gain on a yardstick is the same amount of movement, whether from 10 to 16 inches or 36 to 42 inches. Do percents subtract like inches? Does a difference of 6% imply the same movement everywhere on a percent scale? Unfortunately not! When we reach 95 %, for example, we can no longer increase by 6% because 100% is the top. The same constriction occurs at the bottom near 0%. Subtraction does not work with percents! Because percents and raw scores are bound by 0% at "none" and 100% at "all", neither percents nor raw scores can be linear or interval or suitable for subtraction.

Raw score non-linearity becomes especially clear when we compare the raw scores we expect from persons who take two tests measuring the same variable, but differing in difficulty. **Figure 3** shows what must happen. Any person earning 0% on the easier test must be expected to earn 0% on the harder. But a person earning 50% on the easier test must be expected to do less well on the harder, say only 20%. While a person earning 50% on the harder test must be expected to do better on the easier, say as much as 80%. Finally, however, any person earning 100% on the harder test must be expected to also earn 100% on the easier. The expected score curve in **Figure 3** is the result. This curve shows that neither raw scores nor percent corrects can serve as linear measures of the variables they imply. Before we can do arithmetic with percents or raw scores, before we can adjust percents for the effects of item difficulty bias, we must change their numerical scale into one that is linear (Wright & Linacre 1989).

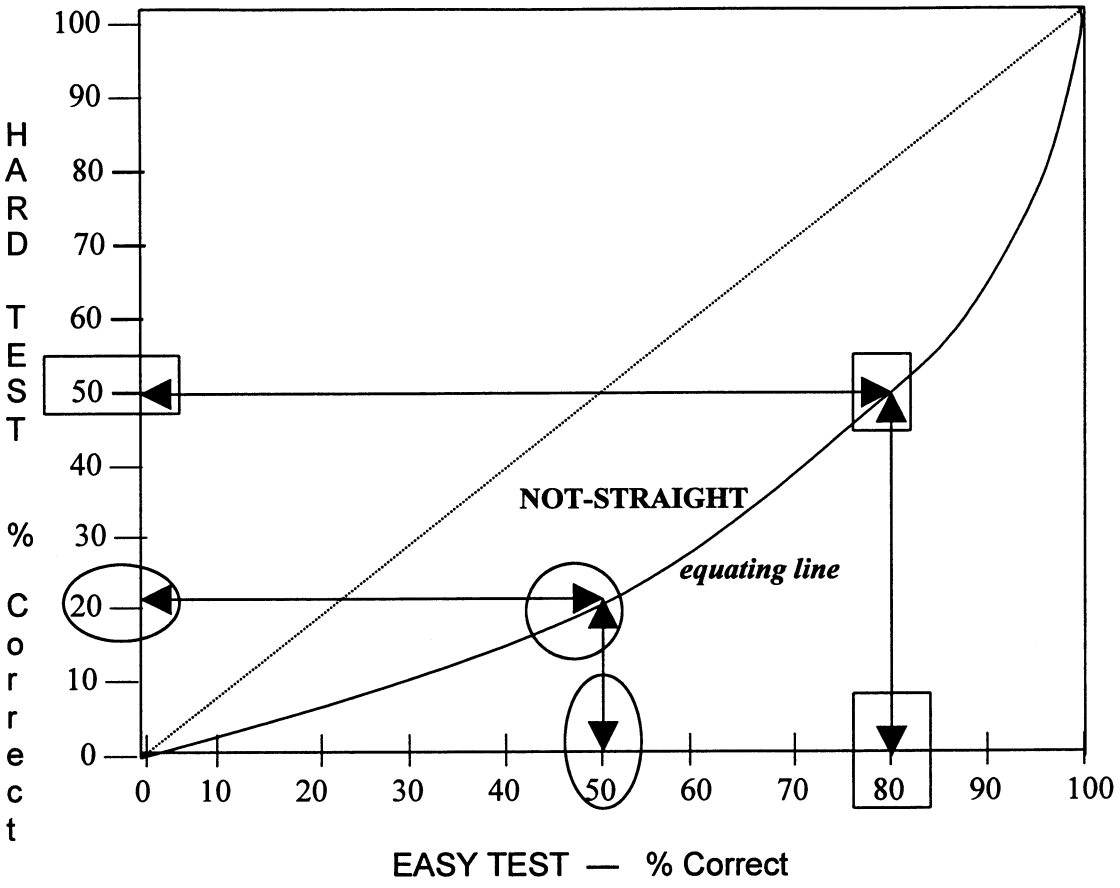

Figure 3: Equating a Hard and Easy Test. Test Scores in Percent Correct.

There is another problem with percents and raw scores. What happens when we want to compare abilities among persons or measure amounts of individual growth from year to year? Unless we can persuade all persons to take the same set of items in the same frame of mind, we cannot compare percents or raw scores because they will not share a common basis. Unless we give a person the same test twice and can force them to remember nothing of the first testing, we cannot measure their growth.

As we think about raw scores and percent corrects we realize an overwhelming necessity for finding a method that abstracts a general inference about probable future performance from present, specific, incomplete raw response data. The method must replace concrete, specific test-bound, non-linear raw scores with abstract, general, test-free, linear measures.

1. All we can **observe** is a concrete set of **nominal** answers to specific items.

2. All we can **count** are person responses which we have decided ought to point to knowledge and so might be interpreted as **ordinal** "correct" answers

3. The **quantitative meaning** of an ordinal raw score, however, varies with **test length** and **item difficulty**. The quantitative meaning of a percent correct varies with item difficulty.

4. Neither raw scores nor percent corrects are **linear** enough to enable item difficulty adjustments based on subtraction or, in fact, any statistical analyses employing summations, means or variances.

5. A raw score is not, in the end, what we want to know. The moment we have a raw score, it has become a local description of a bygone event. What we want to know is what any particular raw score **implies** about the person's general, on-going ability to answer items of this kind.

Measurement Theory

Numerals can be understood in three ways: as **numbers**, in which case they are linear and so arithmetical, as **ranks**, ordinal indicators of a property, in which case linearity is not present since the distances between ranks is unknown, or as **labels**, like social security numbers, in which case not even ordinality is present. Were the numerals on a yardstick rearranged, the numerals themselves, in the abstract, would still be linear numbers. But as measures of length the numerals would have lost their linearity and hence their measurement utility.

The linearity of numbers in an application can only be obtained by definition and construction. Merely reporting some numerals, as though they were measures, does not produce linear measurement. We need a method which can use a person's answers to a set of items to make an inference about the person's probable amount of ability. We need a method which uses raw scores to construct ability measures on linear, equal interval scales - like yardsticks.

Consider a yardstick. Because bodily navigation depends on continuous "measures" of distance, distance is our prototypical variable and the yardstick is our prototypical measuring device. A yardstick "draws a line" in our mind's eye from a particular "here", at one end, to a particular "there", at the other. The distance between ends is marked off in roughly equal inches. Numerical labeling begins with an implied "0" at the "low" end, followed, after a one inch distance, by a "1" and continues inch by inch to an implied "36" at the "high" end. When we use a yardstick, we set the "0" end at one end of the object and define the direction in which we decide to measure length by aiming the yardstick in that direction. Then we note the numerical label of the location along the yardstick at which the other end of the object coincides. That number becomes our measure. While we usually measure a 6 inch object by the difference between "0" and "6", we can, as well, start one end of the object at "12" and note that its other end coincides with "18" . In either case we read the difference, distance or length of the object, as 6 inches.

Yardsticks are calibrated to mark out a linear scale. Because "measures" from yardsticks are linear, they enable useful arithmetic. Since we know how to make yardsticks work, it no longer matters where we begin measuring on a yardstick or which yardstick we use. We have at long

last developed methods for making and using yardsticks which guarantee that measures from different yardsticks, used with different starting points by different persons are, nevertheless, all quantitatively comparable. Of course, any particular measure depends on someone actually making some concrete observations with some particular yardstick. But the **quantitative meaning** of these measures is, nevertheless, completely **independent** of which yardsticks are used, of who uses them, or at which inch markers the measures are "started". We have learned how to separate the incidental particulars of our length-measuring agent, the yardstick, from the measures it makes. As a result, we have learned how to make measures of length **objective**.

A similar "objectivity" is what we need for measuring reading ability. Indeed, should we be unable to construct and maintain reading ability "yardsticks" which work the way length measuring yardsticks do, we would be unable to make quantitative studies of reading. We would be unable to compare readers and we would be unable to measure reading growth.

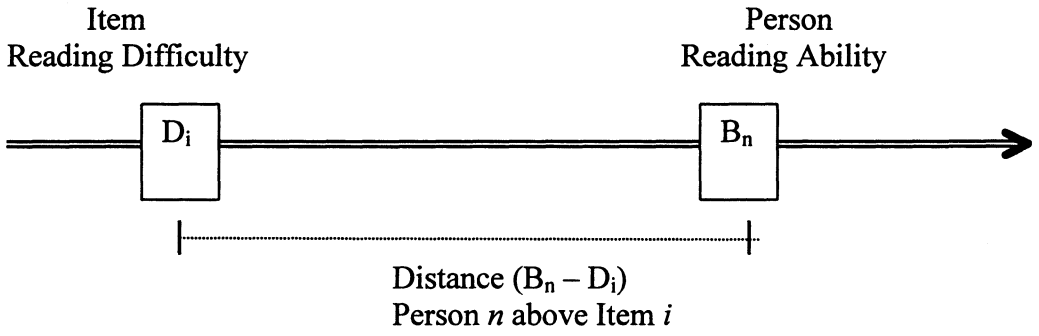

Figure 4: Reading Yardstick in operation.

Figure 4 shows two positions on a "reading yardstick". Location D_i marks the reading difficulty calibration of item *i*. Location B_n marks the reading ability measure of person *n*. The **difference/distance** (B_n . D_i) between these yardstick positions measures the power of person *n* to overcome the resistance of item *i*. The size of this difference governs what will **probably** happen when person *n* is challenged by items of the same difficulty as item *i*. The difference is the person-by-item parameter which determines the probability that persons like person *n* will get items like item *i* correct.

Differences like (B_n . D_i) have no natural bounds. They can become negative or positive to any extent. The only bounds we can specify for differences are - ∞ and + ∞ The item response data from which we must estimate these differences, however, is always bound between the least and most we can observe. Raw scores and percent corrects are bound between nothing right at 0% and everything right at 100%.

What does this difference in boundaries require of the relationship between specific raw data and the general measures they might imply? Let R_{ni} be the "score" person *n* earns against item *i* when

he tries it L_n times. Then when difference $(B_n - D_i)$ is $-\infty$, percent correct $F_{ni} = 100(R_{ni}/L_n)$ must be 0%. When difference $(B_n - D_i)$ is $+\infty$, F_{ni} must be 100%. In the middle, when $(B_n - D_i) = 0$, so that $B_n = D_i$, then F_{ni} might as well be set at 50 %, so that when person ability equals item difficulty the expected result is a draw. As $(B_n - D_i)$ increases so too must percent corrects. Therefore the relationship between percent corrects and person-item differences must be smoothly increasing.

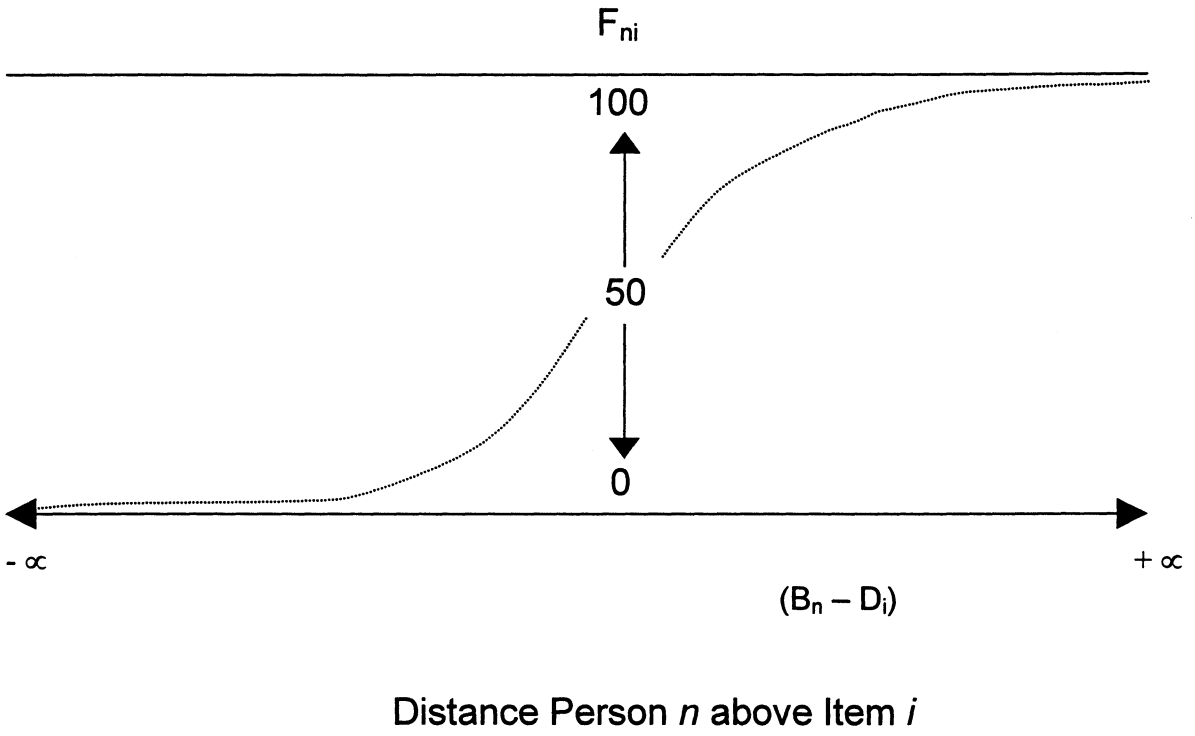

Figure 5: Percent Success: Person *n* over Item *i*

The smoothly increasing correspondence between $(B_n - D_i)$ and F_{ni} requires that the line connecting the reference points $\{(B_n - D_i), F_{ni}\}$ at $\{0\%,-\infty\}$, $\{50\%,0\}$ and $\{100\%,+\infty\}$ follow the S-shaped ogive in **Figure 5**. Observable performance $F_{ni} = 100(R_{ni}/L_n)$ is the vertical ordinate. The measure difference between person *n* and item *i* $(B_n - D_i)$ is the horizontal abscissa. Since the differences which define the abscissa are, by construction, linear, the ogive in **Figure 5** shows the way in which **scores and percents are not linear**. Non-linearity is most severe near 0% and 100%. This can be seen, in detail, in **Figure 6** where an observed increase of 10% from 88% to 98% is shown to be worth 2.8 times as much in linear gain as an observed increase of 10% from 45% to 55%.

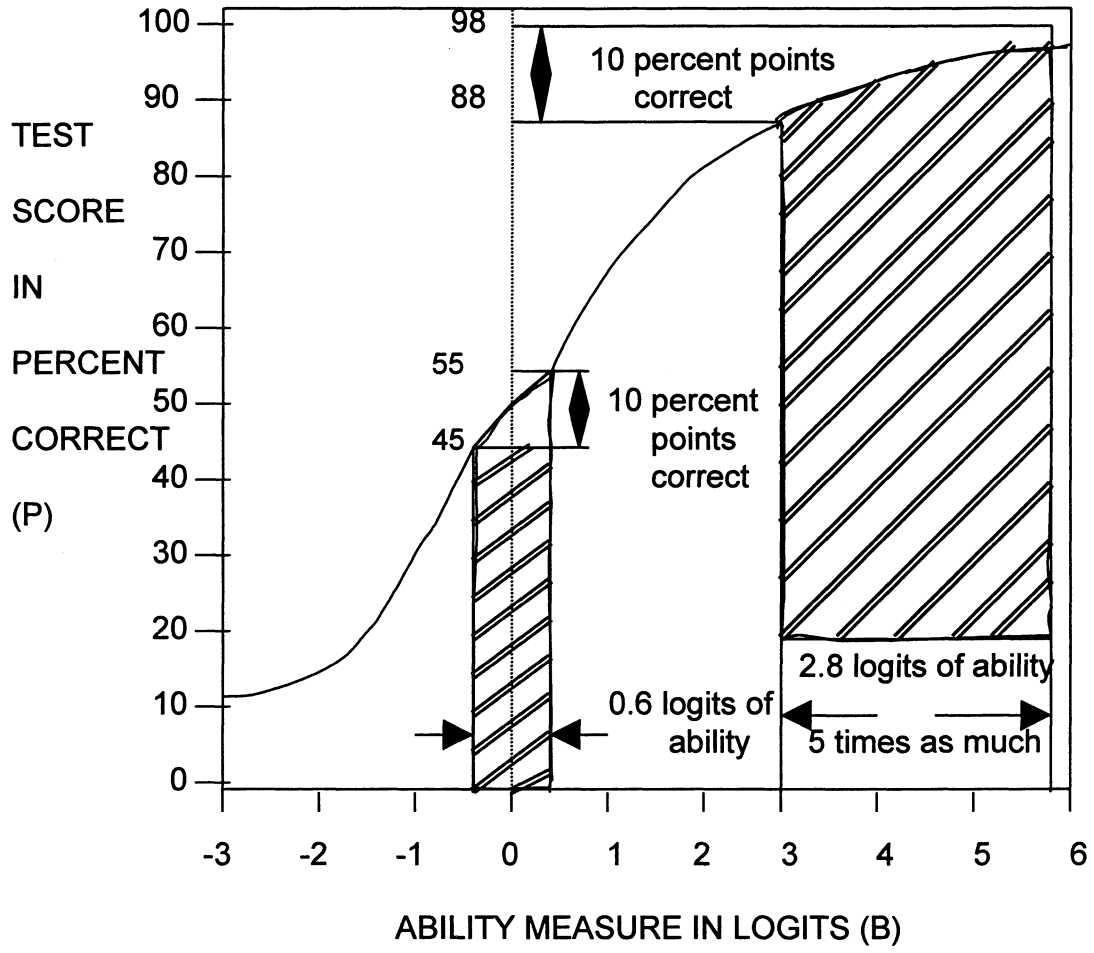

Figure 6: Logistic ogive illustrating nonlinearity of percents.

The most elementary observations of a property are qualitative dichotomies: "Is the property exhibited or not?" On a yardstick this becomes: "Does John exceed the six foot mark or not?" On a math test this becomes: "Does Mary provide a correct answer to 2+2= ? or not?" In physical science, John is compared with a pre-calibrated measuring instrument, a yardstick. In social science, Mary is compared with a test item for which a right answer has been specified but a calibration is, as yet, unknown.

Since, at the outset, both the measure of the object, Mary, and the calibration of the measuring agent, the test item, are unknown, the same set of ordinal observations must lead simultaneously to commensurable but separated estimates of a linear measure for Mary and a linear calibration for the item. When such a conjunction of calibration and measurement can be shown to produce linear results, then the process is called additive conjoint measurement and shown to be equivalent in measurement construction to Campbell's physical concatenation [Luce & Tukey 1964, Brogden 1977, Perline, Wright & Wainer 1979].

When John is very close to six feet tall, then observations as to whether he does or does not exceed the six foot mark will vary. Sometimes he will appear to exceed the mark. Sometimes he will appear not to. All observational processes entail an inevitable uncertainty, a stochasticity that becomes increasingly apparent as object measure and agent calibration approach the same value. Indeed, contradictory observations like this are useful because they are the evidence that object and agent are closely aligned. The possibility of their occurrence, however, requires us to collect more than one observation in order to estimate the precision of our measures.

To solve these problems we need a **measurement model** that:

1. Models raw data as **qualitative dichotomies** (pairs of ordered categories) which express the presence or absence of a property through a **stochastic interaction** between object and agent (readers and text).

2. Constructs **conjoint linear measures** from these qualitative dichotomies.

3. Estimates measures **independently** of one another.

4. Estimates the **quality and precision** of these measures.

The only model that meets these specifications is the one devised by Georg Rasch in 1953 (1960, 1966a, 1966b), since termed the "Rasch model".

The Rasch Model

The curves in Figures 5 and 6 can lead us to the solution of all our measurement problems. If we can deduce a function which reproduces these curves, which transforms finite, bounded, non-linear raw scores and percents into infinite, unbounded, linear differences and which also enables separate, independent estimations of person and item parameters, then that function will be the measurement model we need to estimate general linear measures from specific ordinal raw scores and percent corrects.

So far we have addressed data in their concrete observed percent form. Now we will abstract beyond particular data to the generality which they must imply to be useful. Instead of F_{ni}, a particular observed percent success for person n on item i, we will ascend to its abstract implication, the probability P_{ni} that person n will succeed on item i. This liberates us from the specifics of the particular and so enables generalization. In **Figure 7**, the vertical ordinate has become probability P_{ni} while the horizontal abscissa remains the measured difference $(B_n - D_i)$.

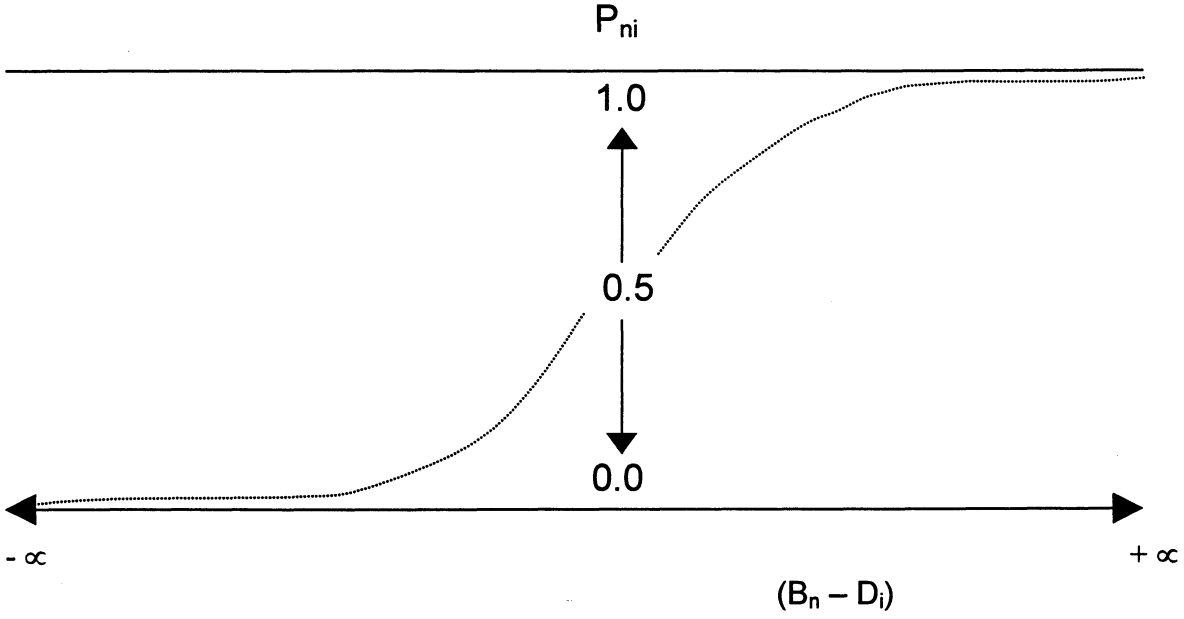

Figure 7: Probability Person *n* succeeds on Item *i*

What simple function of P_{ni} can reproduce the ogive in **Figure 7** and also separate parameters B_n and D_i?

In **Figure 7** the boundaries

$$0 < P_{ni} < 1 \qquad \text{differ from} \qquad -\infty < (B_n - D_i) < +\infty$$

But simple transformations of P_{ni} and $(B_n - D_i)$ produce new boundaries

$$0 < \frac{P_{ni}}{(1 - P_{ni})} < +\infty \quad \text{and} \quad 0 < \exp(B_n - D_i) < +\infty$$

These boundaries match.
Since this pair of expressions have the same range, we will define our measurement model by equating them. The resulting Rasch model is:

$$\frac{P_{ni}}{(1 - P_{ni})} \equiv \exp(B_n - D_i)$$

which solves for P_{ni} to produce

$$P_{ni} \equiv \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}$$

which in its linear form becomes

$$\log\left[\frac{P_{ni}}{1-P_{ni}}\right] \equiv B_n - D_i$$

P_{ni} is the probability that object *n* will succeed on agent *i*
$P_{ni}/(1-P_{ni})$ are the success odds of object *n* against agent *i*
$\log[P_{ni}/(1-P_{ni})]$ is the log-odds of success
B_n is the linear measure of object *n*
D_i is the linear calibration of agent *i*.

When object and agent have exactly the same measure, then the log-odds are 0, the odds are 1, so that $P_{ni} = (1-P_{ni})$ and the probability of success is 0.5. This occurs when an examinee encounters a test item of difficulty exactly matching the examinee's ability.

This model is formulated to produce measures in log-odds units, *logits*. Any convenient linear transformation of logits, however, may be employed. The *Lexile* is closely related to such a transformation.[2]

An essential ingredient of successful physical measurement is that it must not matter which well-constructed yardstick is employed in measuring length. Similarly, in reading, it must not matter which well-constructed set of reading items is used in measuring ability. To settle this question, we must prove that the Rasch model can separate parameters B_n and D_i.

Let the interactions between person *n* and item *i* be summarized by the log-odds term G_{ni}. Let the interactions between person *m* and item *i* be summarized by the log-odds term G_{mi}. These terms can be estimated from data without prior knowledge of person abilities B_n, B_m or item difficulty D_i. Now consider G_{ni} and G_{mi} as defined by the Rasch model.

If $G_{ni} = B_n - D_i$, then $G_{mi} = B_m - D_i$ *and* $G_{ni} - G_{mi} = B_n - B_m$

which contains no reference to D_i.

Thus, as "explained" by the Rasch model, comparisons of person *n* and person *m* become independent of choice of item. We see that comparisons of B's do, indeed, separate from D_i and so, therefore, from any other D's which might be used to estimate B's. We also see that comparisons of items $G_{nj} - G_{ni} = D_i - D_j$ do not depend on any B's and so do not depend on which persons are measured. This proves that the Rasch Model is **sufficient** for parameter separation.

When we initiate our model building with the requirement that parameters **B and D must be separable** and deduce what model follows, we arrive at the same Rasch model (Roskam & Jansen 1984, Wright & Linacre 1987, Wright 1989). The demonstration above establishes the

[2] Indeed Lexiles were developed because of the linear relation found between reading item calibrations and passage reading difficulties calculated from word frequencies and sentence lengths.

sufficiency of the Rasch model for separating parameter estimates. The deduction goes further and proves that the Rasch model is not only sufficient but also **necessary** for the construction of objective linear measures (See Appendix for a short version of this deduction.).

Estimating Calibrations and Measures

To construct a particular measurement system we organize our observations of person-item interactions into a data matrix of scored responses. Since we want enough replications of each person's performance potential to calculate a useful estimate of their continuing ability, we impose enough relevant items on each person to collect a convincing sample of their behavior. The same need for replications applies to items. We need enough replications of how items work with a variety of relevant persons to calculate a useful estimate of their continuing item difficulty.

Replications over items for each person and over persons for each item produce a data matrix $\{\{x_{ni}\ n = 1, N\ \ i = 1, L\}\}$ of N person rows and L item columns. But not all cells need be filled. There can be plenty of missing data, as there must be, for example, when testing is computer-adaptive and every person takes a unique set of items tailored to their particular behavior at the time of testing. All that is necessary to proceed with the analysis is a network of responses in the data matrix sufficiently overlapping to connect all items and all persons through some chain of links. **Figure 8** is an example of a typical CAT data matrix. (For further study of item networks and item banking see Wright & Stone 1979, Chapter 5).

Computer-Adaptive Test Responses	Items in Difficulty Order 1-----10-----20-------30-----40-----50
Persons in Ability Order	1 1100 10010 0 0111110 0 0 1 100001 10 1 110000 1 0 1 11011 1 0 0 0 0 01 110110 00 11 1110 00 00 100111100 0 1110110 0 0 0 1 1 1110 0 00 11 110001 0 0 11 01 10001 0 1 11 00011 00 111 01 10 0 00 010011 011 0 1 1 0100101 0 110110001 0111 1 01000

Figure 8: Computer-adaptive test responses, showing overlapping response strings despite missing data.

In the simplest case, developed here, each response $x_{ni} = 0,1$ of person n to item i is coded into one of two categories. A "1" marks a "correct" response pointing in the direction of "more" of our intended variable. An "0" marks any other, hence "incorrect", response.

We use this $\{\{x_{ni}\}\}$ data matrix to estimate person ability measures $\{B_n - n = 1,N\}$ and item difficulty calibrations $\{D_i, i = 1,L\}$ which maximize the joint probabilities

$$P_{xni} \equiv \frac{\exp[x_{ni}(B_n - D_i)]}{1 + \exp(B_n - D_i)} \qquad x_{ni} = 0,1$$

of these particular data $\{\{x_{ni}\}\}$.

For this we find vectors $\{B_n\}$ and $\{D_i\}$ which maximize the product

$$\prod_{n=1}^{N} \prod_{i \in n}^{L_n} (P_{xni})$$

over all observed x_{ni}.

Simple equations which accomplish this by iteration are:

$$B'_n = B_n + \frac{\sum_{i \in n}^{L_n} (x_{ni} - P_{ni})}{\sum_{i \in n}^{L_n} P_{ni}(1 - P_{ni})} \quad \text{and} \quad D'_i = D_i - \frac{\sum_{n \in i}^{N_i} (x_{ni} - P_{ni})}{\sum_{n \in i}^{N_i} P_{ni}(1 - P_{ni})}$$

in which, since $x_{ni} = 0,1$, P_{xni} is simplified to $P_{1ni} = P_{ni}$ and $P_{0ni} = (1 - P_{ni})$.

Each time one of these equations is executed, it improves an estimate of B_n or D_i and a more accurate value for $P_{ni} = \exp(B_n - D_i)/[1 + \exp(B_n - D_i)]$ is calculated from the improved estimate. As these equations reiterate, the new values of the B_n's and D_i's move closer and closer to the values which maximize the probability of the data $\{\{x_{ni}\}\}$. When the sums of residuals $\Sigma(x_{ni} - P_{ni})$ for each person and each item are all less than 1/2 a score point, iterations have reached a level of accuracy at which the useful information in the integer data has been exhausted. The final values of B_n's and D_i's are those for which these data $\{\{x_{ni}\}\}$ are most probable.

These values are estimates. They will vary from sample to sample, even when samples are identical random replications of exactly the same process. In recognition of this fact of science,

the model uses a binomial distribution to "explain" the observed variations in data. This binomial is governed by probabilities P_{ni} which are defined by differences (B_n - D_i) between linear person and item parameters B_n and D_i. This binomial distribution defines the error of estimation which the model expects from independent replications. Each response x_{ni} = 0,1 with modelled probability P_{ni} for x_{ni} = 1 contains an amount of measurement information equal to $Q_{ni} = P_{ni}(1-P_{ni})$. The total information obtained by any set of $\{x_{ni}\}$ is $\sum[P_{ni}(1 - P_{ni})] = \sum Q_{ni}$. The inverse square roots of these total informations

$$S_{B_n} = \frac{1}{\sqrt{\sum_{i \in n}^{L_n} P_{ni}(1-P_{ni})}} \quad \text{and} \quad S_{D_i} = \frac{1}{\sqrt{\sum_{n \in i}^{N_i} P_{ni}(1-P_{ni})}}$$

are the model estimation errors of measures and calibrations (Wright & Panchapakesan 1969, Wright & Douglas 1977, Wright & Stone 1979, Wright & Master 1982, Wright & Linacre 1994).

Useful Estimation from Summary Statistics

Analysis of response-level data permits meticulous quality control and so the highest accuracy in the resultant measures. Nevertheless, measure estimates based on summary statistics alone are accurate enough for most practical purposes. This permits the construction of measures from tests for which response level data is not available (Stenner, Wright and Linacre, 1994).

Suppose that the conventional item statistics and examinee raw score summary statistics are all that remain of the data from a previous testing. These statistics provide the mean of the N examinee raw scores, their standard deviation and, for each item i, the rate of success of the sample of examinees, a "p-value", P_i, for $i = 1$ to L items.

In order to equate examinee performances on this earlier test with those on other tests, or to add these earlier items to an item bank, conversion from the raw score metric to a linear metric is needed. This can be achieved with a simple, usefully accurate technique, provided that it is reasonable to think of the examinees as randomly selected from a normal distribution.

Here's how to construct logit measures from raw score statistics:

1. Check the raw score statistics for consistency:

$$\text{Is the examinee mean raw score} \approx \sum_{i=1}^{L} P_i \ ?$$

Are there typographical errors?

Even if p-values were obtained from one sample and mean examinee score from another, they may still be close enough for this computation.

2. Compute a raw score-to-ability conversion factor, C_b:

$$C_b = \frac{1}{\sum_{i=1}^{L} P_i(1 - P_i)}$$

3. Compute a logit examinee ability variance from SD, the examinee raw score standard deviation:

$$V_b = C^2_b * SD^2$$

4. Obtain an item calibration expansion factor, X_f, to adjust item difficulties for examinee ability variance:

$$X_f = \sqrt{1 + \frac{Vb}{2.9}}$$

5. Compute a logit difficulty calibration for each item, d_i:

$$d_i = X_f * \log\left[\frac{1 - P_i}{P_i}\right]$$

6. The standard error, SE_i, of calibration d_i for item *i* is:

$$SE_i \approx X_f * \sqrt{\frac{1}{N * P_i(1 - P_i)}}$$

7. Compute an initial logit ability estimate b^o_r corresponding to each raw score r from 1 to L-1:

$$b^o_r = X_f * \log\left[\frac{r}{L - r}\right]$$

8. Compute a final ability estimate b_r corresponding to each raw score r from 1 to L-1 by iterating the equation:

$$b_r \Leftarrow b^{o}_{r} + \frac{r - \sum_{i=1}^{L} P_{ri}}{L * 0.25}$$

where

$$P_{ri} = \frac{1}{1 + e^{(di - br)}}$$

replacing b_r^{o} by b_r and recomputing, until

$$\left| r - \sum_{i=1}^{L} P_{ri} \right| \leq 0.25$$

9. For examinees with extreme scores of 0, compute b_0 using steps 7 and 8 with r = 0.25.

10. For examinees with extreme scores of L, compute b_L using steps 7 and 8 with r = L - 0.25.

11. The standard error, SE_r, of person ability measure b_r, corresponding to raw score r, including r = 0.25 and r = L - .25, is:

$$SE_r \approx \sqrt{\frac{1}{\sum_{i=1}^{L} P_{ri}(1 - P_{ri})}}$$

Quality Control

Why Evaluate the Quality of Measures?

When it is announced that "Jane scored 5 out of 10 on the Reading test", we would correctly deduce that Jane succeeded on half the items, but then we would infer that that half was the easier half. But what if this is not what happened?

Consider a 10 item test with the items listed in order of increasing difficulty. Were we to encounter a pattern like the following

Easy: 0 0 0 1 0 1 1 0 1 1 : Hard Score: 5

we would be puzzled and wonder how could this person answer the two hardest questions correctly and yet get the easiest three questions incorrect. Was this person taking the easy items too casually, effectively "sleeping" on the easy portion of the test? Or perhaps this person was unfamiliar with the response format and "fumbled" through the first few items?

On the other hand, were we to encounter the response pattern

Easy: 1 0 1 0 0 0 0 1 1 1 : Hard Score: 5

our surprise would be as great, but now we might be inclined to explain the irregularity as the result of lucky "guessing" on the three hardest items.

Both the probabilistic nature of the Rasch model and our everyday experience with typical response patterns lead us to expect patterns which have a central region of mixed correct and incorrect response. A pattern like

Easy: 1 1 1 0 1 0 1 0 0 0 : Hard Score: 5

fits our expectation of success on the easy items, failure on the hard items, and a transition zone of mixed success and failure on those items located around the person's current ability level. This is the normal pattern.

Consequently, we may consider a pattern like

Easy: 1 1 1 1 1 0 0 0 0 0 : Hard Score: 5

as "too good to be true." Though this pattern is often thought to be ideal, the sudden switch from success to failure prompts us to suspect that some other factor, apart from the person's ability, is

in evidence. For instance, this unexpectedly regular pattern is sometimes produced by persons who work slowly and carefully, refusing to proceed to the next item until they have done everything possible to answer the present item correctly. This response style is identified as "plodding" .

Detailed examination of each response string in this way is impractical, but quality control is necessary because the only measures that are really useful are those produced by a response pattern similar to the normal one. Consequently we use a statistical summary to direct our attention to major departures in the data from our expectations.

Statistical Quality Control

The degree to which any particular sample of observations will cooperate to imply useful measures is unknown at first, but becomes increasingly clear as the measurement process continues. Incongruities encountered can prompt the elimination of some observations as irrelevant, the reconceptualization of others as misunderstood, and the collection of further observations now discovered to be potentially germane.

We construct data from present experience to test our understanding of the past and to help us predict the future. Real data do not come from the "perfect world" idealized by models. Real data are necessarily complex, influenced by myriad circumstances. We do not expect data to be any better than unperfect manifestations of our ideals. But they are our only sample of a relevant "reality" in which we have a particular interest. Indeed, they are our only new information about that interest.

We expect data **not to fit** our measurement model. Indeed, it is data imperfection on which we depend to learn more useful ways to understand the past, to construct better data, to clarify the generalities, measures and calibrations, which data may imply and, finally, to increase our knowledge of what to think and do next.

We are particularly interested in where and when particular data x_{ni} do not come close to their modelled expectation P_{ni}. To pursue this interest we study the relation between "score" residuals from expectation y_{ni} and their expected variations Q_{ni}

$$y_{ni} = x_{ni} - P_{ni} \quad \text{and} \quad Q_{ni} = P_{ni}(1 - P_{ni})$$

We use y_{ni} and Q_{ni} to calculate standardized residuals

$$z_{ni} = y_{ni} / \sqrt{Q_{ni}} \quad \text{with improbabilities } P_{z_{ni}} = 1/(1 + z^2_{ni})$$

When we work with many persons or many items, we need a systematic way to focus attention on the most salient sources of improbable residuals. One way to do this is to summarize residuals for persons over the items they use and residuals for items over the persons who use them. Though many statistical summary statistics have been proposed, two have proved particularly useful. These are the OUTFIT and INFIT mean-square statistics.

These mean-square (MNSQ) residual fit statistics are, for persons $n = 1,N$,

$$\text{MNSQ OUTFIT: } u_n = \sum_{i \in n}^{L_n} z^2_{ni} / L_n \qquad \text{MNSQ INFIT: } v_n = \sum_{i \in n}^{L_n} y^2_{ni} / \sum_{i \in n}^{L_n} Q_{ni}$$

and, for items $i = 1,L$,

$$\text{MNSQ OUTFIT: } u_i = \sum_{n \in i}^{N_i} z^2_{ni} / N_i \qquad \text{MNSQ INFIT: } v_i = \sum_{n \in i}^{N_i} y^2_{ni} / \sum_{n \in i}^{N_i} Q_{ni}$$

These statistics focus on two types of misfit that threaten the validity of measure estimates. Large OUTFIT values indicate the presence of unexpected responses by persons whose ability levels are far from the difficulty levels of the items in question. Thus "lucky guesses", unexpected successes by low ability persons on high difficulty items, and "carelessness", unexpected failure by high ability persons on low difficulty items, are flagged by large OUTFIT values. Thus OUTFIT detects the most blatant departures in the data from modelled expectation.

The pattern of responses in the transition zone is probed by INFIT. This pattern is expected to go from mainly success to mainly failure. Disturbances in this pattern, zones of local success or failure, also raise questions about the precise meaning of a measure. Such disturbances can be caused by special knowledge, alternative curricula, and changes in item type. Though INFIT detects more subtle effects than OUTFIT, INFIT is often more important to the understanding of the underlying variable, i.e., what is being measured.

The OUTFIT and INFIT mean-square summaries are ratios of observed variation in squared residuals $(x_{ni}- p_{ni})^2$ to their model expectations $Q_{ni} = P_{ni}(1-P_{ni})$. The mean-squares measure misfit magnitude on a ratio scale. They can be linearized for linear comparison and visualization by taking logs. They can be standardized to a mean of "0" and a variance of "1" by a cube root transformation (Wright in Rasch 1980 p. 194, Wright & Masters 1982 p.101). Interpretation of standardized fit statistics, however, must take into account the impact of test length and sample size.

These person and item misfit summaries point to the particular persons and items which manifest behavior which is unusual among the majority of persons and items. It is the particulars of the misfitting responses of these person and items which teach us how to improve our data construction and how to clarify our operational definition of the variable which we intend to materialize and use (Smith 1986,1988, Linacre 1990).

Appendix: Deriving the Rasch Model from Objectivity

Objectivity is the requirement that the measures produced by a measurement model be sample-free for the agents (test items) and test-free for the objects (people).

A.1 Comparison of performances

The idea of "comparison" is essential to the concept of measurement. A measurement is the quantification of a specifically defined comparison. Consequently, it is necessary that we are explicit about the kind of comparison for which we intend to obtain measures.

In examining the performances of a person on a test, we can expect that the greater the length of the test, the greater will be the numerical differences between their counts of right answers and the counts of wrong answers. But for a test consisting of homogeneous items, we do expect that the ratio of the count of right answers to that of wrong answers will remain approximately constant whatever the length of the test. If the test were to be doubled in length, the ratio between the successes and failures should remain about the same. Consequently, a constant ratio is the type of comparison which we need in order to construct measures from right answers.

Objective measures from paired observations.

This derivation is based on the hypothetical administration of numerous replications of the same item to two people in order to produce the contingency table:

		Person n	
		right	wrong
Person m	right	R_nR_m	W_nR_m
	wrong	R_nW_m	W_mW_m

where R_nR_m counts the times that *n* and *m* both answer "right", W_nR_m counts the times that *n* answers "wrong" but *m* answers "right" and so on.

In those instances where *n* and *m* both answer "right" or both answer "wrong", we detect no difference in their performance. Consequently the only informative performance contrast is a comparison between $\mathbf{R_nW_m}$ and $\mathbf{W_nR_m}$. The ratio of these terms, then, is the comparison we want.

Consider $\mathbf{R_nW_m/W_nR_m}$. This ratio compares the success/failure frequencies of two people *n* and *m* on the item in question. In the limit, a comparison of frequencies becomes a comparison of probabilities multiplied by the number of replications. Here, however, the replications are the same for both cells so their count cancels.

Thus:

$$\frac{R_n W_m}{W_n R_m} \approx \frac{P_{ni}(1 \sim P_{ni})}{(1 - P_{ni})P_{mi}}$$

where *i* indicates the particular item being replicated, P_{ni} indicates the success probability of person *n* on this item *i* and $(1-P_{ni})$ is the corresponding failure probability.

A.2 Use of Objectivity

What happens when we require this comparison to maintain objectivity? Then the comparison of the performance of persons *n* and *m* must not depend on which particular item we use to compare them. When we use another item j, we must obtain the same result.
Expressing this necessity algebraically:

$$\frac{P_{ni}(1 - P_{mi})}{(1 - P_{ni})P_{mi}} \equiv \frac{P_{nj}(1 - P_{mj})}{(1 - P_{nj})P_{mj}} \qquad \textit{for all } i,j$$

Solving for $P_{ni} / (1 - P_{ni})$

$$\frac{P_{ni}}{(1 - P_{ni})} \equiv \frac{P_{nj}}{(1 - P_{nj})} \times \frac{P_{mi}}{(1 - P_{mj})} \times \frac{(1 - P_{mj})}{P_{mj}}$$

To maintain objectivity, the interaction of person *n* and item *i* must also not depend on which other person *m* or which other item *j* are used for comparison in the measuring process. Consequently we can choose the measure of any convenient person "o" to define the origin of the linear scale for person measures and the calibration of any convenient item "o" to define the origin of the linear scale of item calibrations. Thus,

$$\frac{P_{ni}}{(1 - P_{ni})} \equiv \frac{P_{no}}{(1 - P_{no})} \times \frac{P_{oi}}{(1 - P_{oi})} \times \frac{(1 - P_{oo})}{P_{oo}} \equiv f(n) \times g(i) \times \textit{constant}$$

When we bring the origins for person measures and item calibrations into conjunction by choosing the reference item and person such that $P_{00} = 0.5$, then the constant term becomes 1.

The measurement system defined so far has the properties of a ratio scale. A zero is the measure of a person having no chance of success on **any** item with a non-zero calibration. Zero is also the

calibration of an item on which there is **no** chance of failure by **any** person with a non-zero measure.

In this ratio frame of reference $(P_{no})/(1-P_{no})$ has a value between 0 and infinity depending **only** on person *n*, and $(P_{oi})/(1-P_{oi})$ has a value between 0 and infinity depending **only** on item *i*.

The ratio scale defined by $(P_{ni}/(1-P_{ni}))$ can be transformed into an equal-interval linear scale by taking logarithms, so that

$$\log(P_{no}/(1 - P_{no}) = B_n$$

$$\log (P_{oi}/(1-P_{oi})) = -D_i$$

and

$$\log(P_{ni}/(1 - P_{ni})) = B_n - D_i$$

or

$$P_{ni} = \exp(B_n - D_i) / (1 + \exp(B_n - D_i)$$

Item calibration D_i is dependent only on an attribute of item *i*, which we may call "item difficulty"', and person measure B_n is dependent only on an attribute of person *n*, which we may call "person ability". The choice of $P_{oo} = 0.5$ produces a constant of value 1 with logarithm 0.

This Rasch model relating the ability of person *n* and the difficulty of person *i* to performances of person *n* on item *i* is the consequence of our deduction from the requirement of objectivity. Thus it is not only sufficient but also necessary for objective measurement.

REFERENCES

Brogden, H. E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. Psychometrika. 42, 631-634.

Campbell, N. R. (1920). Physics: The Elements. London: Cambridge University Press.

Eisenhart, C. (1963). Realistic evaluation of the precision and accuracy of instrument calibration systems. Journal of Research of the National Bureau of Standards 67C(2) 161-187.

Linacre J. M. (1990) Where does misfit begin? Rasch Measurement Transactions, 3(4) p. 79-80.

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

Perline, R., Wright, B. D. & Wainer, H. (1979). The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 3, 237-256.

Rasch, G. (1960/1980/1993). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: MESA Press.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. Henry (Eds.), Readings in Mathematical Social Science (pp. 89-107). Chicago: Science Research Associates.

Rasch, G. (1966b). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Roskam, E. E. & Jansen, P. G.W. (1984). A new derivation of the Rasch model. In E. Degreef & J. Van Buggenhaut (Eds.), Trends in Mathematical Psychology (293-307).

Smith, R. M. (1986). Person fit in the Rasch Model. Educational and Psychological Measurement, 46, 359-372.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. Educational and Psychological Measurement, 48, 657-667.

Stenner, A. J. (1996). The Objective Measurement of Reading Comprehension. Durham NC, MetaMetrics, Inc.

Stenner, A. J. (1997). Objectivity, units of measurement and zeroes. Rasch Measurement Transactions, 11:2, Autumn , 560-561.

Wright, B. D. (1983). Afterword. In Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: MESA Press.

Wright, B. D. (1989). Deducing the Rasch model from Thurstone's requirement that item comparisons be sample-free. Rasch Measurement Transactions, 3(1) p.49-50.

Wright, B. D. & Douglas, G. A. (1977). Best procedures for sample-free item analysis. Applied Psychological Measurement, 1, 281-294.
Wright, B. D. & Linacre, J. M. (1987). Rasch model derived from objectivity. Rasch Measurement Transactions, 1(1) p.2-3. Fall.

Wright, B. D. & Linacre, J. M. (1994). BIGSTEPS Rasch analysis computer program. Chicago: MESA Press

Wright, B. D. & Linacre, J. M. (1989) Observations are always ordinal: Measures, however, must be interval. Archives of Physical Medicine and Rehabilitation, 70, 857-860.

Wright, B. D. & Masters, G. N. (1982). Rating Scale Analysis. Chicago: MESA Press.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: MESA Press.