MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware

MEASUREMENT ESSENTIALS
TABLE OF CONTENTS

Foreword

The 26 Chapters of Measurement Essentials are Units of Study. These units evolved during our weekly discussions of measurement on Sunday mornings from 1994 to 1998. Each Sunday we met at Ben's kitchen table and worked together on the explanation of a measurement topic which had arisen in our work.

Some chapters came together quickly. Others took weeks of reconsideration. As we worked we drew on whatever material seemed most useful. You will find excerpts from our 1979 book, BEST TEST DESIGN. This happens when we wanted to reconsider fundamental material.

You will find some chapters that return to a topic already discussed in an earlier chapter. This happens when we found we had more to say about a topic or wanted to explore a different perspective, but saw no particular fault with our first discussion.

You will also find our notation inconsistent. Our notational variations, however, need not trouble you. Whether the conjoint additive relation between person ability and item difficulty is represented by "B-D", "*B-D*", "b-d", or "*b-d*" the meaning remains entirely clear.

We enjoyed building these 26 Units of Study. We wrote them for ourselves, for our students and for you. We hope they will be useful to you.

1. THE IDEA OF MEASUREMENT

No discussion of scientific method is complete without an argument for the importance of fundamental measurement - measurement of the kind characterizing length and weight. Yet, few social scientists attempt to construct fundamental measures. This is not because social scientists disapprove of fundamental measurement. It is because they despair of obtaining it.

The conviction that fundamental measurement is unattainable in social science and education has such a grip that we fail to see that our despair is unnecessary. Fundamental measurement is not only obtainable in social science but, in an unaware and hence incomplete form, is widely relied on. Social scientists are already practicing a kind of fundamental measurement but without knowing it and hence without enjoying its benefits or building on its strengths.

The realization that fundamental measurements can be made in social science research is usually traced to Luce and Tukey (1964) who show that fundamental measurement can be constructed from an axiomatization of comparisons among responses to arbitrary pairs of objects of two specified kinds. But Thurstone's 1927 Law of Comparative Judgement (1928a, 1928b, 1929) contains results which are rough examples of fundamental measurement. Fundamental measurement also occurs in Bradley and Terry (1952) and in Rasch (1958, 1960/1980, 1966a, 1966b, 1967, 1977).

The fundamental measurement which follows from Rasch's "specific objectivity" is developed in Rasch 1960/1980, 1961, 1967 and 1977. Rasch's "specific objectivity" and R.A. Fisher's "estimation sufficiency" are two sides of the same implementation of inference. Andersen (1977) shows that the only measuring processes which support specific objectivity and hence fundamental measurement are those which have sufficient statistics for their parameters. It follows that sufficient statistics lead to and are necessary for fundamental measurement.

Several authors connect "additive conjoint" fundamental measurement with Rasch's work (Keats, 1967, 1971; Fischer 1968; Brogden, 1977). Perline, Wright and Wainer (1977) provide two empirical demonstrations of the equivalence of non-metric multidimensional scaling (Kruskal, 1964, 1965) and the Rasch process in realizing fundamental measurement. Wright and Stone (1979) show how to obtain fundamental measurement from mental tests. Wright and Masters (1982) give examples of its successful application to rating scales and partial credit scoring.

In spite of these publications advancing, explaining and illustrating the successful application of fundamental measurement in social science research, most contemporary psychometric tests and much practice are either unaware of the opportunity or mistake it for impractical.

MAINTAINING A UNIT

Thurstone says,

> The linear continuum which is implied in all measurement is always an abstraction....All measurement implies the recreation or restatement of the attribute measured to an abstract linear form.
>
> and
>
> There is a popular fallacy that a unit of measurement is a thing - such as a piece of yardstick. This is not so. A unit of measurement is always a process of some

> kind which can be repeated without modification in the different parts of the measurement continuum (Thurstone, 1931, p. 257).

Campbell (1920) specifies an addition operation as the hallmark of fundamental measurement. At bottom it is maintaining a unit that supports addition.

Rasch (1980, 171-172) shows that, if

$$P = exp(b\text{-}d)/G \tag{1.1}$$

where

$$G = [1 + exp\ (b\text{-}d)] \tag{1.2}$$

is the way person ability b and item difficulty d combine to govern the probability p of a successful outcome and, if Event AB is person A succeeding but person B failing on a particular item, while Event BA is person B succeeding but person A failing on the same item, then a distance between persons A and B on a scale defined by a set of items of a single kind can be estimated by

$$b_A - b_B = \log N_{AB} - \log N_{BA} \tag{1.3}$$

where N_{AB} is the number of times A succeeds but B fails and N_{BA} is the number of times B succeeds but A fails on any subset of these items.

This happens because, for any item difficulty d under Rasch's model,

$$P_{AB} = P_A(1 - P_B) = \exp(b_A - d) \,/\, G_A G_B \tag{1.4}$$

and

$$P_{BA} = P_B(1 - P_A) = \exp(b_B - d) \,/\, G_A G_B \tag{1.5}$$

so that d, G_A and G_B cancel out of $P_{AB} \,/\, P_{BA} = \exp(b_A - b_B)$ leaving

$$\log(P_{AB} \,/\, P_{BA}) = b_A - b_B \approx \log(N_{AB} \,/\, N_{BA}) \tag{1.6}$$

a unit of distance which holds regardless of the value of d. This result is equivalent to Case 5 of Thurstone's 1927 Law of Comparative Judgment and to Bradley and Terry of 1952 and conforms to Luce and Turkey of 1964.

Since d does not appear in this equation, estimates of the distance between A and B are modelled to be statistically equivalent whatever the item difficulty d.

Since the unit defined by the distance between A and B holds over the range of the continuum defined by whatever values d may take but is independent of the particular value of d, it follows that Rasch's model for specifying measures is exactly the unit-maintaining process which Thurstone (1931) requires.

Whether a particular batch of data can be disciplined to follow the Rasch process can only be discovered by applying the process to the data and examining the consequences. It is worth noticing, however, that whenever we have deemed it useful to count right answers or to add scale ratings, we have taken it for granted that the data concerned do, in fact, follow the Rasch process well enough to suit our purposes. This is so because counts and additions are exactly the sufficient statistics for the Rasch process and for no other. When we accept the counts as useful ,then, however innocent our adventure, we also accept the Rasch model as the mathematical explanation of what we are doing and also its only mathematical justification.

If we subscribe to Thurstone's requirement, then we want data that we can govern in this way. That means that fitting the Rasch process becomes more than a convenience. It becomes the essential criterion for data good enough to support the construction of fundamental measures. *The Rasch process becomes the criterion for valid data.*

VERIFYING FIT, IDENTIFYING BIAS

How well does data have to fit the Rasch process in order to obtain fundamental measurement? The only reasonable or useful answer is: "Well enough to serve the practical problem for which the measures are intended, that is, well enough to maintain an invariance sufficient to serve the needs at hand."

How can we document the degree of invariance the Rasch process obtains with a particular set of data? One method is to specify subsets of items in any way that is substantively interesting but also independent of the particular person scores we have already examined (N_{AB}, N_{BA}) and then to see whether the new counts resulting from these item subsets estimate statistically equivalent distances between the persons.

The extent to which the distance between persons A and B is invariant over challenging partitions of items is the extent to which the data succeeds in making use of the Rasch process to maintain a unit.

A more general way to examine and document fit is to compose for each response $x = 0$ or 1 the score residual:

$$y = x - Ex = x - P \qquad 1.7$$

$$\text{in which} \quad P = \exp(b-d)/[1+\exp(b-d)] \qquad 1.8$$

comes from the current estimates of person ability b and item difficulty d and the expected value Ex of observation x is

$$Ex = P \qquad 1.9$$

and then to accumulate these score residuals over the item subsets chosen to challenge fit.

If $(b_1 - b_0)$ is defined as the extent to which a subset of items implied by b_1 fails to maintain the unit constructed by the full set of items implied by b_0, then that subset sum of score residuals Σy estimates:

$$Ey \approx (b_1 - b_0) \sum (dy/db) \qquad 1.10$$

in which the summation $\sum$ is over the items in the designated subset.

When the data fit the Rasch process, then the differential of y with respect to b equals the score variance $P(1-P)$ so that

$$dy/db = dP/db = P(1-P) = q \qquad 1.11$$

$$Ey \approx (b_1 - b_0)\Sigma q \qquad 1.12$$

and

$$(b_1 - b_0) \approx \Sigma y / \Sigma q = g. \qquad 1.13$$

Thus the simple statistic $g = \Sigma y / \Sigma q$ estimates the logit discrepancy in scale invariance $(b_1 - b_0)$ due to the item subset specified, with g having expectation and variance

$$Eg = 0 \text{ and } Vg = 1 / \sum q \qquad 1.14$$

when the data fit this unit-maintaining Rasch process.

Subsets need not be limited to items. Groups of persons can be used to review the extent to which any item is vulnerable to bias for or against the type of persons grouped. In general, any combination of items and persons thought to interact in a way that might interfere with the unit-maintaining process can be used to define a subset for calculating g. The resulting value of g estimates the direction and logit magnitude of the putative disturbance to scale invariance. The stability of any particular value of g can be evaluated from the root of its model variance, $Vg = 1 / \Sigma q$:

$$SE_g = (\Sigma q)^{-1/2} \; . \qquad 1.15$$

CONSTRUCTING ADDITION

The way to build a linear scale is to construct an addition operation which answers the question: "If person A has more ability b_A than person B with ability b_B, then how much "ability" must be added to b_B to make the performance of B appear equal to the performance of A?" To be more specific, "What 'addition' to b_B will cause $P_B = P_A$?"

To answer this question we must realize that the only situation in which we can observe these probabilities of success is the one in which we expose the persons to items of the specified kind. This changes the question: "What change in the situation through which we find out about persons by testing them with items will give B the same probability of success as A ?" In other words:

"What 'addition' to b_B will cause $P_{Bj} = P_{Ai}$?"

To be more explicit, "What item j of difficulty d_j will make the performance of person B appear the same as the performance of person A on item i ?"

The Rasch process specifies that when $P_{Bj} = P_{Ai}$ then

$$b_B - d_j = b_A - d_i \; . \qquad 1.16$$

The 'addition' required to cause B to perform like A is then

$$b_B + (b_A - b_B) = b_A \; . \qquad 1.17$$

The way this 'addition' is accomplished is to give person B an item j with difficulty

$$d_i - d_j = b_A - b_B \; . \qquad 1.18$$

easier than item *i*, namely an item *j* with difficulty

$$d_j - d_i = b_A - b_B \text{ so that} \qquad 1.19$$

$$b_B + (b_A - b_B) = b_B + (d_i - d_j) = b_A \text{ and} \qquad 1.20$$

$$P_{Bj} = P_{Ai} \; . \qquad 1.21$$

The way the success of this 'addition' is evaluated is to see whether the performance of person *B* on items like *j* is observed to be statistically equivalent to the performance of person *A* on items like *i*. This, in fact, is the comparison actually checked in every detailed analysis of fit.

CURRENT PRACTICE

It has long been customary in social science research to construct scores by counting answers (scored by their ordinal position in a sequence of ordered response possibilities) and then to use these scores and monotonic transformations of them as measures. When the questions asked have only two answer categories, then we count right answers. When the questions have an ordered series of answer categories, then we count how many categories from 'least' to 'most' ('worst' to 'best', 'weakest' to 'strongest') have been surpassed. There is scarcely any quantitative data in social science research not already in this form or easily put so.

If there has been any progress in quantitative social science, then this kind of counting must have been useful. But this has implications. Counting in this way implies a measurement process, not any process, but a particular one. Counting implies a process which derives counting as the necessary and sufficient scoring procedure.

Now counting is exactly the unique sufficient statistic for estimating measures with the Rasch process. Since the Rasch process constructs simultaneous conjoint measures whenever data are valid for such a construction, we have, in our counting, been practicing the first steps of fundamental measurement all along. All we need do now is to take this implication of our actions seriously and to complete our data analyses by verifying the extent to which our data fit the Rasch process and so are valid for fundamental measuring. When our data can be organized to fit well enough to be useful, then we can use the results of counting to construct Thurstone linear scales and to make Luce and Tukey fundamental measures on them.

That we—in social science and education—have been content to use unweighted raw scores, just the count of right answers, as our 'good enough' statistic for ninety years, testifies to our latent conviction that the data with which we work can be usefully managed with a process no more complicated than the Rasch process. It is useful to keep in mind that, among all of the intriguing mathematical possibilities which might seem useful to transform right answer counts into measures, it is only the Rasch process which can maintain units that support addition and so produce results that qualify as fundamental measurement.

2. *OBJECTIVITY*

This chapter introduces the essentials of objectivity (also known as monotonicity, composite transitivity, conjoint additivity and fundamental measurement), and to deduce the measurement model that objectivity requires.

The progress of science depends on the invention, construction and maintenance of useful measures. Science lives on measurement. Measurement exists on objectivity. An everyday term for objectivity is generality. Objectivity is the expectation and, hence, requirement that the amount and meaning of a measure has been well enough separated from the measuring instrument and the occasion of measurement so that the measure can be used as a quantity without qualification as to which was the particular instrument or what was the specific occasion.

Although a measuring occasion is necessary for a measure to result, the utility of the measure, depends on the specifics of the occasion disappearing from consideration. It must be possible to take the occasion for granted and, for a time being, to forget about it. Were such a separation of meaning from the circumstances of its occasion not possible, not only science but also commerce, and even communication, would become impossible.

The essentials of measurement can be brought out by reviewing the characteristics of the archetypical variable, length. When you ask a person's height and it is reported 70 inches, you do not demand the yardstick or to know who made the measure, when or where. You expect, and hence require as a precondition for continuing communication concerning height, that 70 inches was obtained in the usual way. Even though you know what the circumstances necessary to produce the measure were necessarily fraught with unique particulars, you use the 70 inch quantification of length as though it were entirely independent of those circumstances of its construction. In other words, you take 70 inches to be objective. Were you unable to do that, the quantity 70 inches would become meaningless not only to you but also to anyone, for nobody would know to what, if any, enduring state it referred.

Measuring length is so familiar and commonplace that the way we do it seems obvious. We are tempted to think of length as an explicit, manifest variable that can be seen directly. But there are essential details which the measurement of length requires. Although these details are taken for granted, they cannot be neglected, if length is actually to be measured. In fact, that they can be taken for granted signifies that we have made a solid habit of not neglecting them.

In spite of its "looks," length is not, by itself, manifest. Nor, in fact, is there any variable at all which is manifest on its own. Variables are inventions and measurements are constructions. An agent of measurement, a ruler of some kind, is necessary to make length "visible." Length cannot be "seen" on its own, let alone measured, without the deployment of some kind of ruler. This requires the measurement of length to be a conjoint operation. The calibrated ruler and the thing to be measured must be brought into a disciplined conjunction. The ruler, through its calibration, recapitulates the founding definition of the variable "length." The ruler's calibrations are the criterion definition of this variable. The ruler, while necessarily concrete in its realization of "length," depends for its utility on the extent to which it implements an abstract fiction. It must not matter at all which particular concrete realization of a "ruler" is actually used to make the abstract measurement. It must only need be any "ruler" in good standing.

All measurements made by all calibrated rulers must be quantitatively comparable without any reference to the physical details or work histories of the particular rulers used or who used them.

SOCIAL SCIENCE MEASUREMENT

These ideas are not new to social science. To be generally useful, the individual measure must not depend on which particular test items are used.

> It should be possible to omit several test questions at different levels of the scale without affecting the individual score.
>
> It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point, being selected by the examiner, should not directly affect the individual score (Thurstone, 1926, p. 446).

Nor should the measuring function of a test, that is, the calibrations of the test items, depend on which particular persons are being measured.

The scale must transcend the group measured. One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited.

If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired.

> Within the range of objects for which the measuring instrument is intended, its function must be independent of the objects of measurement (Thurstone, 1928, p. 547).

Indeed, Thurstone's eloquent and detailed 1931 specification of the essentials of measurement meets and resolves most of the "big" measurement misgivings that social scientists continue to fret about.

Measurement is Necessarily One-Dimensional:

> One of the most frequent questions (concerning the possibility of social measurement) is that a score on an attitude scale, let us say the scale of attitude toward God, does not truly describe the person's attitude.
>
> There are so many complex factors involved in a person's attitude on any social issue that it cannot be adequately described by a simple number such as a score on some sort of test or scale. This is quite true, but it is also equally true of all measurement.
>
> The measurement of any object or entity describes only one attribute of the object

measured. This is a universal characteristic of all measurement. When the height of a table is measured, the whole table has not been described but only that attribute which has been measured.

Similarly, in the measurement of attitudes, only one characteristic of the attitude is described by a measurement of it.

Measurement is Necessarily Linear:

Only those characteristics can be described by measurement which can be thought of as linear magnitudes. In this context, linear magnitudes are weight, length, volume, temperature, amount of education, intelligence, and strength of feeling favorable to an object. Another way of saying the same thing is to note that the measurement of an object is, in effect, to allocate the object to a point on an abstract continuum. If the continuum is weight, then individuals may be allocated to an abstract continuum of weight, one direction represents small weight while the opposite direction represents large weight.

Measurement is Necessarily Abstract:

The linear continuum which is implied in all measurement is always an abstraction. For example, when several people are described as to their weight, each person is in effect allocated to a point on an abstract continuum of weight. All measurement implies the reduction or restatement of the attribute measured to an abstract linear form. There is a popular fallacy that a unit of measurement is a thing such as a piece of yardstick. This is not so. A unit of measurement is always a process of some kind which can be repeated without modification in the different parts of the measurement continuum (Thurstone. 1931, p.257).

But no ruler in its concrete embodiment of the abstract idea of length does its job without further specification. There are rules concerning how rulers must be employed to produce acceptable measures. The ruler and the object to be measured must be carefully aligned so that they lie parallel to one another. A starting point, or origin, and units to count must be installed. The line of sight along which the viewer reads the object against the ruler must be determined and maintained. The procedure by which coincidence is identified and interpolation accomplished must be specified. Without care for these rules, the results of ruler measurements become too disorderly to be useful.

AXIOMATIC MEASUREMENT THEORY

The axiomatic theory of measurement has made great strides in the past 30 years. There are detailed and scholarly discussions of these accomplishments in print. Unfortunately these discussions are too esoteric for most social scientists. It is hard for practitioners to see how to put axiomatic measurement theory to work.

The heart of axiomatic measurement theory, however, can be simply put. The crucial axiom which all measurement theorists agree is necessary for the construction of measurement is the one they call "monotonicity" or "conjoint additivity."

This axiom can be useful to social scientists because it marks out exactly the condition which both scientist and layman expect of numbers which are intended to serve as measures, namely generality or objectivity.

The joint ordering of conjoint additivity is also not new to social science. Monotonicity under the name of "conformity" and later "objectivity" appears in the practical work of Georg Rasch in 1953 and is defined, developed and implemented in detail in his seminal book of 1960 (Rasch, 1960/1980) and article of 1961 (Rasch, 1961).

> A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one (Rasch, 1960, p. 117).

Rasch "objectivity" is a stochastic conjoint additivity. Even earlier in 1944, Louis Guttman (1944, 1950) formulated what must be the best known, but least followed, requirement for social science measurement. Guttman deduced that a score could not be unequivocally on a "scale," unless the particular data from which the score was accumulated were completely specified by the value of the score.

> If a person endorses a more extreme statement, he should endorse all less extreme statements if the statements are to be considered a scale.
>
> We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person (Guttman, 1950, p. 62).

Guttman "scalability," a deterministic conjoint additivity, is impractical when applied deterministically. But its stochastic version is identical to Rasch's objectivity and entirely practical, as Rasch demonstrated in the 1950's and as has been shown so many times since for hundreds of tests and questionnaires (Wright and Bell, 1984).

What may not be quite as obvious is that the stochastic version of Guttman's requirement is equivalent to Ronald Fisher's seminal definition of a sufficient statistic (1958/1922). Fisher's "sufficient" statistic is the one and only statistic that exhausts the information modeled in the data with respect to the parameter to be estimated.

What this definition means is that a Fisher "sufficient" statistic is the statistic that provides the best stochastic reconstruction of the data. This is exactly Guttman's scalability criteria, expressed stochastically. The realization that Fisher "sufficiency" is a necessary concomitant of stochastic monotonicity may prove, in the end, to be the decisive reason for preferring sufficient statistics over all others.

The common sense of this, so often reiterated, foundation for measurement is plain enough. It would seem that no sane researcher could argue or act otherwise. Yet, and strangely, few social scientists require or even hope for conjoint additivity in the numbers they use as "measures."

The consequence of this innocent carelessness is a plethora of ill-defined and unstable pseudo-quantifications and a great deal of confusion and disappointment over ambiguous and irreproducible results.

This unhappy situation is completely unnecessary. A derivation of a practical stochastic measurement model from the requirement of monotonicity, conjoint additivity or objectivity is easy to follow and the resulting model for measurement is easy to apply.

Here is a simple derivation of the model necessary to meet Thurston's 1928 requirement that a scale be independent of the objects of measurement.

THURSTONE INVARIANCE

The construction of a scale depends on the relative calibrations of the items used to define the scale. These calibrations must be established in a way that can be made independent of which persons happen to provide the calibration data. We begin by asking what is required so that the comparison of any two items i and j will be independent of whatever persons are used to elicit evidence of the relative scale standing of these two items?

Items *i* and *j* can be observed to differ only when they are answered differently. Realizing a comparison of *i* and *j*, then, requires counting how often *i* is answered 'yes' by persons when *j* is simultaneously answered 'no' and comparing this "$i > j$" count with the reciprocal "$j > i$" count of how often the reverse occurs among other persons.

The estimation of a quantitative comparison of items *i* and *j* from this pair of reciprocal counts requires a probability model for the occurrence of the counts which can implement an objective, i.e. sample-free, person-invariant, comparison of their probabilities.

The pair of probabilities can be represented by

$$\Pr[(i = yes),\ (j = no)]$$

and

$$\Pr[(i = no),\ (j = yes)]$$

and their comparison specified by the ratio,

$$\frac{\Pr[(i = yes),\ (j = no)]}{\Pr[(i = no),\ (j = yes)]} \qquad 2.1$$

Let $P_{ni} = f(n,i)$ be the, as yet undefined, probability that person n succeeds on item i.

What we seek is the particular function $f(n,i)$ which maintains Thurstone (1928) invariance and hence Rasch (1960/1980) objectivity.

To obtain invariance the comparison of probabilities in Equation 1 must stay the same regardless of which persons are involved. That is, Equation 1 must hold for any suitable persons n or m as in,

$$\frac{\Pr[(i = yes),(j = no)]}{\Pr[(i = no),(j = yes)]} = \frac{P_{ni}(1-P_{nj})}{(1-P_{ni})P_{nj}} \equiv \frac{P_{mi}(1-P_{mj})}{(1-P_{mi})P_{mj}} \qquad 2.2$$

for all n and m

where n is some person, m is any other person and the symbol $\equiv$ specifies that the comparison of item i with j is "defined" to remain the same whoever the persons, n or m.

To simplify our appreciation of the implications of Equation 2 for $P_{ni} = f(n,i)$, we can choose $j = o$ and $m = o$ as origins for the item and person scales so that the calibration of item i becomes its comparison with a reference item $j = o$ and the measure of person n becomes their comparison with a reference person $m = o$.

We can also align these scale origins so that the reference person has a fifty-fifty chance to succeed on the reference item. This makes

$$P_{mj} = P_{oo} = 1/2 \text{ and } (1-P_{oo})/P_{oo} = 1$$

When we insert $j = o$ and $m = o$ into Equation 2-2 and solve the middle and right side for the odds of person n succeeding on item i we get

$$\frac{P_{ni}}{(1-P_{ni})} = \frac{P_{no}}{(1-P_{no})}\frac{P_{oi}}{(1-P_{oi})} = g(n)*h(i) \qquad 2.3$$

$(P_{no})/(1-P_{no})$ 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 measurement scale defined by Equation 2-3 is a ratio scale. Zero corresponds to the measure for a person having no chance of success on any item and also to the calibration of an item on which there is no chance of success by any person.

The ratio scale defined by $P_{ni}/(1-P_{ni})$ can be transformed into an equal-interval linear difference scale by taking logorithms.

$$\log[P_{ni}/(1-P_{ni})] = \log[P_{no}/(1-P_{no})] + \log[P_{oi}/(1-P_{oi})] \qquad 2.4$$

$$= G(n) + H(i) \text{ for an interval scale}$$

$$= B_n - D_i \text{ for convenience}$$

or

$$P_{ni} \equiv \exp(B_n - D_i) / [1 + \exp(B_n - D_i) \qquad 2.5$$

where the item calibration D_i depends only on the attributes of item i, which we can call its difficulty, and the measure B_n depends only on the attribute of person n, which we can call his ability.

This model relating the ability of person *n* and the difficulty of item *i* to the performance of person *n* on item *i* is the objective model of measurement known as the Rasch model.

This deduction arrives at the only *f(n, i)* which can support the construction of Thurstone invariant or Rasch objective scales.

Equation 2-2 can be rewritten to address Thurstone's concomitant 1926 requirement that the individual measure not depend on which particular items are used so that it becomes "possible to omit several test questions at different levels of the scale without affecting the individual score." This requires that the comparison of any pair of persons *n* and *m* be invariant with respect to the particular items employed as in

$$\frac{\Pr[(i = yes),\ (j = no)]}{\Pr[(i = no),\ (j = yes)]} = \frac{P_{ni}(1 - P_{mi})}{(1 - P_{ni})P_{mi}} \equiv \frac{P_{nj}(1 - P_{mj})}{(1 - P_{nj})P_{mj}} \qquad 2.6$$

for all *i* and *j*

which is equivalent to Equation 2-2 and so leads to Equation 2-5.

3. THE IDEA OF A VARIABLE

When we make measures, we do so with the intention of being accurate enough for our practical purpose. We do not expect absolute precision. Our notion of "accurate" does not imply "perfect." Instead it implies "close enough to be useful." We record a person's height to some useful approximation like the nearest half-inch. This is sufficient for most practical purposes. More precision, such as to the nearest eighth or sixteenth of an inch, is rarely necessary and we would not ordinarily expect it to be given.

This example reminds us that while we want to be accurate there is always an implied, if not explicit, tolerance in our measures. Unless height measures require some particular accuracy, it is not necessary, and without scientific instrumentation, impossible, to make measures of height more accurate. However, we are not at all frustrated by our lack of absolute precision because "to the nearest half-inch" is practical and useful. We make measures which are good enough for the occasion, good enough to satisfy our practical requirements.

Measures are based on observations. Observations are essentially qualitative. To make measures we develop rules by which to control how these observations are best made. These rules include specifying the degree of accuracy that we want. When measuring height, for example, we ask people to remove their shoes, stand straight and not wiggle in order to standardize the observations. Then we observe which marks on our yardstick they exceed and which they fail to exceed. We find the marks closest to the top of their head. We pick the mark that looks closest and call their height the calibration of that nearest mark. These rules provide the level of accuracy we need in order to make useful measures of a person's height. We use the constructed functional unidimensionality of the yardstick to bring out and record the single dimensioned height of the multidimensional person.

Measuring "ability" is analogous to measuring "height." First we bring to the fore our idea of the variable we want to measure. Next we determine what observations it will be useful to consider as informative manifestations of that variable. Then we construct agents, write items, intended to elicit singular instances of this "made-to-be" unidimensional "ability" variable.

The idea of a variable can be visualized as a line that has direction. When we think about "length" we think about a line that is as long as necessary for our work. This idea is manifest in a one-foot ruler when we expect measures to be 1 to 12 inches, in a yardstick for 1 to 36 inches, on a surveyor's tape for longer distances and so on. In each of these instances the agent of measurement is a focused manifestation of our infinite linear image of the variable "length".

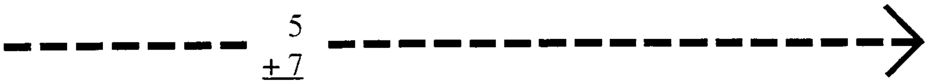
With these simple ideas of measurement in hand, let us turn to the problem of measuring an ability. Consider arithmetic ability and, more specifically, the computation skills needed for the whole number operations of addition, subtraction, multiplication and division.

We imagine a line of arithmetic items progressing from left to right with each successive item harder than the previous one. A few items will suffice for our example. Additional items are added to the line by designing them to fit between any two items that we have already placed upon the line and then verifying their location by observing student responses.

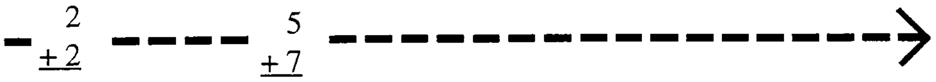
The idea of a line upon which to position arithmetic items provides us with a picture of the arithmetic variable and shows us how to proceed in the construction of tests to measure along that variable. We use our knowledge of arithmetic to position items along the line. Theoretical locations for the items can be hypothesized initially by teaching experts who have experience with students learning arithmetic. Later we can add and reposition items on the line as we observe how well students actually answer these items.

CONSTRUCTING THE LINE OF THE VARIABLE

We begin with a single item and position it on the line of the arithmetic variable:



Can an easier item be constructed? Yes, and so we will position it somewhere to the left. A harder item will be positioned to the right. Hence:



Now we have three items and the process of constructing new ones only requires hypothesizing their expected positions among the existing items and then estimating item positions empirically by collecting responses to them. The critical decision to make at this point is where each item belongs, in our best judgment, relative to the items already positioned on the line. There is also no reason why items cannot be repositioned according to better information from teachers about their difficulty relative to other items. Item construction thus proceeds in an orderly fashion guided by the idea of a line and the successive placement of items on the line according to our best expectations of their relative difficulty. These items now serve as the agents designed to evoke manifestations of our arithmetic variable.

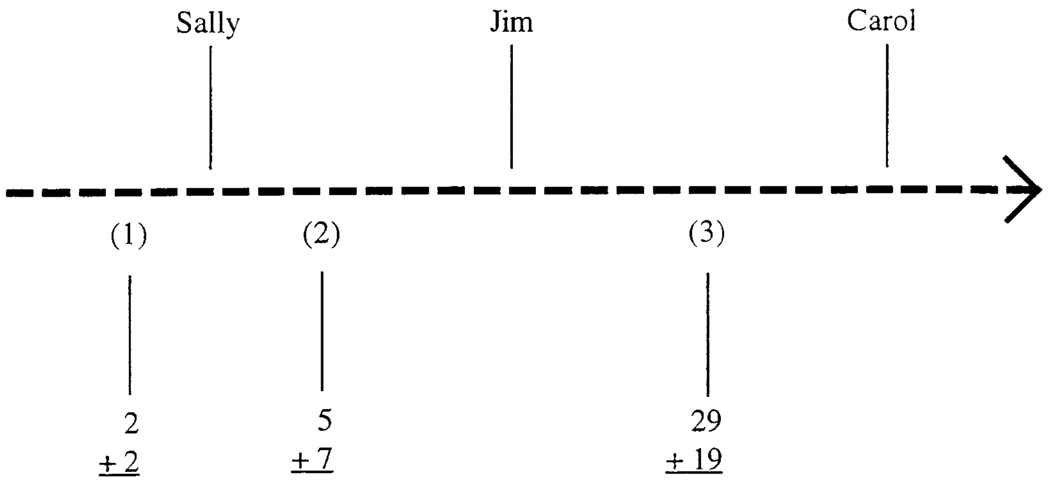
The line of our variable can be made as long as necessary to describe the variable. It can be divided into segments for ease in handling. It can be abbreviated according to our practical needs for administration in exactly the same way that we partition our idea of length into measuring tools - rulers, yardsticks, tapes of various extension - all to facilitate the measurement of various lengths.

The idea of a line helps us to determine item positions by considering each item relative to the items already positioned on the line. This determination can be done by comparing pairs of items with respect to their relative difficulties along the line. Each successive item position as we move to the right indicates "more" of the variable to be measured.

The idea of the variable becomes defined by the construction of the items which work to elicit indications of the variable. As we define the variable with more and more instances, using more and more items, our work of building the variable proceeds in a logical manner and our conceptualization of the variable becomes ever more clearly defined.

Once the variable is constructed by the line of items, we can proceed to position students on this same line. Their probable positions can be specified initially by our best guess as to their ability to correctly answer the items which define the variable. The line of our variable shows both the positions of items and the positions of students. Eventually the positions of students will become more explicit and more empirical as we observe what items they correctly answer.

Consider this picture:



Sally's position on the variable is indicated by an expected correct response to Item 1 but an expected incorrect responses to Items 2 and 3. Her differing responses to Items 1 and 2 locate her on the variable between two items that describe her ability in arithmetic computation. She can add 2 and 2 but not 5 and 7.

Jim's position is between Items 2 and 3 because we expect him to answer Items 1 and 2 correctly but not Item 3. In Jim's case we have somewhat less precision in determining his arithmetic ability because of the lack of items between Items 2 and 3. If we had additional items in this region, we could obtain a more accurate indication of Jim's position on the variable as defined by his responses to these additional items.

Carol solves all three problems. Her ability is "above" Item 3. But what her position is beyond Item 2 remains unknown. We cannot position her more precisely on the variable because we do not know whether her true position is only slightly above the position of Item 3 or far beyond it. If we had her responses to additional items on the variable above Item 3, Carol's position might be indicated more exactly.

Now we give a more specific example of how to construct a variable for arithmetic.

First, we choose 17 items and arrange them on a test form in what we expect to be their approximate order of difficulty.

Then we administer this test form to 270 students in Grades one to six in order to obtain actual data with which to calibrate these 17 items objectively.

Next we calibrate these items and determine person measures for this sample. (See Wright & Stone, 1979, for the details of how to do this. It is not hard to do.)

The calibrations of the 17 items are used to map the items in Table 3.1. The variable line goes down in difficulty from hard at the top to easy at the bottom. On the left side of Table 3.1 is the person count for this sample of 270 students at every raw score position, then comes the raw score, the measure implied by each raw score and the associated estimation error. Items are identified by their item number and text and positioned according to the difficulties calibrated from the observations gathered from our sample.

We have constructed an arithmetic computation variable and located items and students along it from our observations of how these students were able to answer these items.

Our development of this emerging variable defined by items and students provides an operational definition. The variable's limits are bounded only by the range of agents (items) and objects (students) that we can position along the line. We can make variables of interest as dense as we need. The tests which implement these variables can be sparse for rough screening or dense for more specific pinpointing.

Accuracy (i.e. reliability) of student position is given by the standard error associated with each measure. The unit of measurement used in this table is the logit expressed as a decimal centered on 0.0 for this set of 17 items. Observe that the standard errors are smallest (most precise) where items are most dense and we have the most information about the measure and largest (least precise) at the extremes where items are least dense and we have the least information.

Table 3.1 can be examined to determine where we have gaps between items, where there are too many items at a particular position and where more items are needed to extend the variable above and below the items already calibrated. (For an example of this kind of variable building, see Wright & Stone, 1979, pp. 83-93). The map of the variable is a picture of the extent to which we have accomplished the task of variable construction. The map also shows us what to do next.

Variable maps begin by showing item positions along the line of the variable as shown in Table 3.1. We can also add students along the line of the variable and index their positions on the map by name, grade, gender or other student characteristic. As we add to the map we enrich our picture of the variable and increase its utility.

The construction of an empirical variable map enhances the value of testing. A good variable map is self-explanatory because the visualization of the variable makes explicit what the variable represents. The interpretation of test results is facilitated because all items calibrated and all students measured are positioned together on the same variable - along with whatever additional information has been added to make the map more useful.

CRITERION REFERENCING

A variable map is automatically criterion-referenced by the relative positions of item content. The texts of the items in their positions along the variable describe in detail the explicit hierarchy of content and hence the construct implied by the variable. This item-by-item criterion referencing of the variable applies to any measure subsequently derived from any test composed of some items which have been calibrated on this variable. Thus, criterion referencing is complete and the evidence of content and construct validity is explicit.

NORM REFERENCING

Personal and demographic characteristics of any and all students tested can be added to the variable map at the measured positions of these students. This provides as extensive and versatile norm referencing as the use of tests based on items calibrated on this variable can provide. Thus, norm referencing is also as complete as possible with the data available.

Table 3.1

The Item Map of the Arithmetic Variable

STUDENT COUNT	RAW SCORE	MEASURE SCALE*	STANDARD ERROR	ITEM NUMBER	ITEM TEXT
		6.50		#17	$\begin{array}{r} 7\frac{1}{6} \\ \underline{-\frac{3}{4}} \end{array}$
		6.30			
17	16	6.10	1.35		
		5.90			
		5.70			
		5.50			
		5.30			
		5.10			
		4.90			
22	15	4.70	1.11	#16	$\begin{array}{r} \frac{1}{3} \\ \underline{+\frac{1}{3}} \end{array}$
		4.50			
		4.30			
		4.10		#15	$536\overline{)4762}$
		3.90			
24	14	3.70	0.99		
		3.50			
		3.30			
		3.10			
		2.90			
23	13	2.70	0.88		
		2.50		#12	$\begin{array}{r} 42 \\ \underline{\times 29} \end{array}$
		2.30			
22	12	2.10	0.81		
		1.90			
		1.70		#9, #14	$\begin{array}{r} 837 \\ \underline{\times 7} \end{array}$ $\begin{array}{r} 5204 \\ \underline{-530} \end{array}$
24	11	1.50	0.76		
		1.30		#11	$31\overline{)62}$
18	10	1.10	0.74		
		0.90			
		0.70		#13	$9\overline{)72}$
11	9	0.50	0.74		
		0.30		#10	$\begin{array}{r} 23 \\ \underline{\times 3} \end{array}$
		0.10		#8	$\begin{array}{r} 84 \\ \underline{-36} \end{array}$
17	8	-0.10	0.75		
		-0.30			
23	7	-0.50	0.79		
		-0.70		#6	$\begin{array}{r} 45 \\ 16 \\ \underline{+27} \end{array}$
		-0.90			

Table 3.1 (Continued)

The Item Map of the Arithmetic Variable

STUDENT COUNT	RAW SCORE	MEASURE SCALE*	STANDARD ERROR	ITEM NUMBER	ITEM TEXT
23	6	-1.10	0.85	#7	$\begin{array}{r}14\\-8\\\hline\end{array}$
		-1.30		#5	$\begin{array}{r}67\\+4\\\hline\end{array}$
		-1.50			
		-1.70			
15	5	-1.90	0.99		
		-2.10			
		-2.30			
		-2.50			
		-2.70			
		-2.90			
18	4	-3.10	1.15		
		-3.30			
		-3.50			
		-3.70			
		-3.90			
		-4.10			
9	3	-4.30	1.04		
		-4.50			
		-4.70		#2	$\begin{array}{r}6\\+7\\\hline\end{array}$
		-4.90			
3	2	-5.10	0.98	#1, #4	$\begin{array}{r}7\\+4\\\hline\end{array}$ $\begin{array}{r}8\\-3\\\hline\end{array}$
		-5.30			
		-5.50		#3	$\begin{array}{r}6\\-4\\\hline\end{array}$
		-5.70			
		-5.90			
1	1	-6.10	1.16		
270				17	

MEAN ABILITY OF PERSONS = 1.03 LOGITS
STANDARD DEVIATION = 2.56 LOGITS

* Should we wish a numbering system simpler than the decimal logits, a linear conversion can be made to positive whole numbers. See Wright & Stone, 1979, pp. 191-209.

4. DEDUCING THE MEASUREMENT MODEL

OBJECTIVITY

In this chapter we deduce the Rasch Model from Thurstone's requirement that item comparisons be sample free. Thurstone (1928) says, "The scale must transcend the group measured ... its function must be independent of the object of measurement." (p. 228). This ideal for measurement requires that the comparison of two items *i* and *j* be independent of whatever persons are used to elicit evidence of the scale difference between these two items.

Because of the symmetry in any person-by-item interaction, Thurstone's ideal also requires that the comparison of any pair of persons *n* and *m* be invariant with respect to the particular items employed. As Wright (1968) explains, "Object-free instrument calibration and instrument-free object measurement are the conditions which make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements....When we compare one item with another in order to calibrate a test, it should not matter whose responses to these items we use for the comparison. Our method for test calibration should give us the same results regardless of whom we try the test on. This is the only way we will ever be able to construct tests which have uniform meaning regardless of whom we choose to measure with them." (p. 87-88).

Rasch (1960, 1961, 1968, 1977) designated this measurement property *objectivity*. "In the beginning of the 60's I introduced a new - or rather a more definite version of an old - epistemological concept. I preserved the name of *objectivity* for it, but since the meaning of that word has undergone many changes since its Hellenic origin and is still, in everyday speech as well as in scientific discourse, used with many different contents, I added a restricting predicate: specific." (1977, p. 58).

Let us examine this measurement goal with a simple example.

COMPARING TWO ITEMS

We require the comparison of two items to be independent of which people help us to make that comparison. What are the possibilities?

1. Person 1 takes both items, *i* and *j* and answers them both correctly. In this case, we cannot compare these two items on the basis of these two responses because both responses are the same. We can see no difference between them.

2. Person 2 answers both items incorrectly. Again we cannot compare the two items for the same reason.

3. Person 3 answers item *i* incorrectly but item *j* correctly. Now we see a difference between the responses to items *i* and *j* and can infer that item *j* is probably easier than item *i* for Person 3.

4. Person 4 answers item *i* correctly but item *j* incorrectly. Again we have a difference, although the reverse of the previous case. Now we infer that item *i* is probably easier for Person 4 than item *j*.

Our inferences from these examples are based upon the reasonable, even necessary, requirement that. other factors being equal. an item solved correctly is easier than an item solved incorrectly by the same person.

Let more and more people take this simple test of two items. The only respondents that tell us about the difference between the two items are those who differ in their outcomes, i.e. those who answer one item correctly but the other item incorrectly.

As the number of persons who take these two items becomes indefinitely large, we want to be able to record the outcome without bothering with the exact number of persons who happen to be involved. To do this we change our recording from counts to percents.

Suppose, among persons getting one item correct but the other incorrect, we have 10% correct for item *i* but 90% correct for item *j*. We can use the ratio of these two percents to indicate the difference in difficulty between items *i* and *j*: i.e., item *j* is gotten correct (90%) / (10%) = 9 times more often than item *i*.

If a stable and hence useful relation between items *i* and *j* exists, then we must expect the ratio of their relative success rates to remain statistically equivalent irrespective of the people who respond to them. Should the ratio vary substantially between different groups of people, then the differences in ratio would have to be traced to extraneous factors differentiating the groups and thence to local interactions between item content and group characteristics.

When varying results of this kind occur, items cannot be calibrated objectively. Then we need to continue our investigation with contrasting groups of persons to uncover and bring under control the extraneous factors which cause the ratios to vary and thus prevent the establishment of a sample-free comparison of the items.

If, however, in many additional groups taking these two items, we observe a series of ratios close to the ratio of the first group, i.e. about 9 to 1 (with minor variations like 9.5 to 1 and 8.5 to 1) we may decide to interpret these ratios as statistically equivalent and to conclude that we have observed a consistency on which to build an objective calibration of items and hence an operational definition of a stable variable.

We may conclude that it will be useful to think of a "fixed" difference in difficulty between these two items, one that is independent of the differences among the groups of people that produced ratios near 9 to 1 and hence to characterize this difference between these two items in a general way by this ratio (or, to express it explicitly as a "difference", by the logarithm of this ratio).

Observing a comparison of *i* and *j* requires counting how often item *i* is answered "correct" by persons who also answer "incorrect" to item *j* and comparing, by means of their ratio, this "$i > j$" count with the reciprocal "$j > i$" count of how often the reverse occurs.

Estimating a difficulty ratio between items *i* and *j* from this pair of reciprocal counts requires a probability model for the occurrence of the counts which can implement a sample-free, person-invariant, comparison of the items.

The observed percents (i.e. relative frequencies) can be extrapolated conceptually to probabilities of correct (*P*) and incorrect (1 - *P*) responses to items *i* and *j* and we can use this abstract probability *P* to model what is likely to happen when any person tries items *i* and *j*.

Let the probabilities for the two outcomes to the pair of items be:

$$\Pr[(i = yes),\ (j = no)] = P_{ij} \text{ for } "i > j"$$

and

$$\Pr[(i = no),\ (j = yes)] = P_{ji} \text{ for } "j > i"$$

and let the specification of the comparison of the items be the ratio:

$$\frac{\Pr\left[(i = yes),\ (j = no)\right]}{\Pr\left[(i = no),\ (j = yes)\right]} = \frac{P_{ij}}{P_{ji}} \qquad 4.1$$

Let P_{ni} $f(n,i)$ represent the, as yet unknown but now to be deduced, probability that person *n* succeeds on item *i*. Then the comparison of *Equation 4.1* becomes:

$$\frac{P_{ij}}{P_{ji}} = \frac{P_{ni}(1 - P_{nj})}{(1 - P_{ni})P_{nj}} \qquad 4.2$$

The particular function $P_{ni} = f(n,i)$ which we seek is one which maintains Thurstone's invariance or Rasch's objectivity, one which enables *Equation 4.2* to be a person-free comparison of items *i* and *j* - a comparison independent of who person *n* happens to be.

To obtain Thurstone's invariance or Rasch's objectivity, the comparison of probabilities in *Equation 4.2* must stay the same regardless of which persons are involved. *Equation 4.2* must, therefore, hold for *any pair* of suitable persons, such as persons *n* and *m*:

$$\frac{P_{ni}(1 - P_{nj})}{(1 - P_{ni})P_{nj}} \equiv \frac{P_{mi}(1 - P_{mj})}{(1 - P_{mi})P_{mj}} \qquad 4.3$$

Equation 4.3 can be used to specify the odds that person *n* answers item *i* correctly as:
where the triple equal sign "≡" means "this equation is required by definition."

To obtain a general invariance, and hence a useful "objectivity," this equation must hold for *all suitable persons n* and *m* and, by the way, also for all suitable items *i* and *j*.

The triple equals sign "≡" signifies that this equality relation is "the definition" of an "objective" comparison, i.e. the definition of sample-free item calibration and also a test-free person measurement, i.e. Rasch's "objectivity," or Thurstone's "invariance."

We intend to deduce the specification of $P_{ni} = f(n,i)$ from *Equation 4.3*. Since we are entirely free to choose the particular other person m and the particular other item j in any way that is convenient and since the definition of every scale requires the specification of an origin to anchor that scale, it is particularly convenient to choose $m = o$ to be any person with ability right at the origin of the scale and also $j = o$ to be any item with difficulty at the same origin. This choice completes the anchoring of the scale by specifying that when any person takes any item with a difficulty which exactly matches their ability, then their probability of success on that item will be exactly $P = 1/2$.

Inserting $j = o$ and $m = o$ into *Equation 4.3* and solving the middle and right side of the equation for the odds of person n succeeding on item i produces:

$$\frac{P_{ni}}{(1-P_{ni})} = \left[\frac{P_{nj}}{(1-P_{nj})}\right] * \left[\frac{P_{mi}}{(1-P_{mi})}\right] * \left[\frac{(1-P_{mj})}{P_{mj}}\right]$$

$$= \left[\frac{P_{no}}{(1-P_{no})}\right] * \left[\frac{P_{oi}}{(1-P_{oi})}\right] * \left[\frac{(1-P_{oo})}{P_{oo}}\right]$$

$$= g(n) * h(i) * C = g(n) * h(i) \qquad 4.4$$

because $g(n) \equiv \dfrac{P_{no}}{1-P_{no}}$ is a function of n and the choice of orgin, but not a function of i,

$h(i) \equiv \dfrac{P_{oi}}{1-P_{oi}}$ is a function of i and the same choice of origin, but not a function of n,

and $C \equiv \left[(1-P_{oo}) / P_{oo}\right] \equiv 1$ because we chose to relate persons and items so that

$P_{oo} \equiv (1-P_{oo}) \equiv 1/2$.

Equation 4.4 specifies that the odds of person n succeeding on item i must be entirely determined by the product of a single valued function characterizing person n and another single-valued function characterizes item i and by nothing else. This defines a ratio scale in *g(n)* and *h(i)*. To express the relation between person n and item i on an interval, or difference scale, in B_n and D_i, we take the logarithm of *Equation 4.4*:

$$\log\left[\frac{P_{ni}}{(1-P_{ni})}\right] = \log\ [g(n)*h(i)] = \log\ g(n) + \log\ h(i)$$

$$= G(n) + H(i) = B_n - D_i$$

$$\text{where } B_n \equiv \log\ g(n) \text{ and } -D_i \equiv \log\ h(i)\ . \qquad 4.5$$

Equation 4.3 can also be used to address Thurstone's concomitant 1926 requirement that the individual measure not depend on which particular items are used so that it becomes "possible to omit several test questions at different levels of the scale without affecting the individual score" (p. 446). (By "score" Thurstone denotes a generic test-free "measure" rather than a necessarily test dependent raw score.) This requires that the comparison of any pair of persons *n* and *m* be invariant with respect to the particular items employed for all *i* and *j*. This requirement also leads to *Equation 4.3* and thence to *Equation 4.5.*

Another way to write *Equation 4.5* is to solve for *P* so that:

$$P_{ni} \equiv \exp(B_n - D_i) / [1 + \exp(B_n - D_i)] \qquad 4.6$$

This is the equation known as the "Rasch Model" because Rasch was the first person to use this equation to construct measurements.

Most important, this specification of P_{ni} is unique in that it is both *sufficient* and *necessary* for measurement to occur. It is the one and only $P_{ni} = f(n,i)$ which can support the construction of invariant scales meeting Thurstone's criteria, or any other measurement criteria, for objectivity in measurement.

PARAMETER SEPARATION

The Rasch model can be used to seek a useful joint ordering of items and persons. The form in which its parameters occur, $(B_n - D_i)$, linear and without interactions, permits likelihood equations in which the relation between data and person ability parameters can be entirely contained in one estimation equation and the relation between data and item difficulty parameters entirely in another. This happens because the algebraic separation of parameters specified by the Rasch model enables derivation of conditional estimation equations for either set of parameters such that the equations for estimating item difficulties do not involve the person ability parameters and the equations for estimating person abilities do not involve the item difficulty parameters.

SEPARATING ITEM COMPARISONS FROM PERSONS

Equation 4.6 can be used to specify the odds that person *n* answers item *i* correctly as:

$$[P_{ni} / (1 - P_{ni})] = \exp(B_n - D_i)\ . \qquad 4.7$$

The logarithm of *Equation 4.7* is:

$$\log\left[P_{ni} / (1 - P_{ni})\right] = B_n - D_i \ . \qquad 4.8$$

in log-odds units or "logits."

The comparable log-odds for any other item *j* and the same person *n* is:

$$\log\left[P_{nj} / (1 - P_{nj})\right] = B_n - D_j \ . \qquad 4.9$$

Items *i* and *j* can be compared without interference from B_n or any other B_m by subtracting *Equation 4.8* from *Equation 4.9*. This yields:

$$(B_n - D_j) - (B_n - D_i) =$$

$$(D_i - D_j) = \log\left\{\left[P_{nj}(1 - P_{ni})\right] / \left[P_{ni}(1 - P_{nj})\right]\right\} \qquad 4.10$$

Equation 4.10 does not involve B_n at all - exactly what Thurstone called for in 1928.

The comparison of item *i* and item *j* in *Equation 4.10* depends on the participation of relevant persons, but not on any particular persons. P_{ni} and P_{nj} are both dependent on the ability of person *n*. But the parameter separation which is unique to the Rasch model allows us to combine them in *Equation 4.10* so that B_n cancels out leaving the comparison $(D_i - D_j)$ of items *i* and *j* completely untroubled by person effects.

SEPARATING PERSON COMPARISONS FROM ITEMS

For any other person *m* and item *i* the log-odds is:

$$\log\left[P_{mi} / (1 - P_{mj})\right] = B_m - D_i \ . \qquad 4.11$$

Now persons *n* and *m* can be compared by subtracting *Equation 4.11* from *Equation 4.8*:

$$\left(B_n - D_i\right) - \left(B_m - D_i\right) =$$

$$(B_n - B_m) = \log\left\{\left[P_{ni}(1 - P_{mi})\right] / \left[P_{mi}(1 - P_{ni})\right]\right\} \qquad 4.12$$

Equation 4.12 does not involve the item parameter D_i at all - exactly what Thurstone called for in 1926.

The comparison of person n and person m in *Equation 4.12* depends on the use of relevant items, but not on any particular items. P_{ni} and P_{mi} are both dependent on the difficulty of item i. But the parameter separation which is unique to the Rasch model allows us to combine them in *Equation 4.12* so that D_i cancels out leaving the comparison $(B_n - B_m)$ of persons n and m completely untroubled by item effects.

The possibility of estimation equations for B_n which are free from the individual effects of particular D_i is referred to as "test-free person measurement." The possibility of estimation equations for D_i which are free from individual effects of particular B_n is referred to as "sample-free item calibration" (Wright, 1968).

For explanations and examples of Rasch measurement applied see Wright & Stone (1979) and Wright & Masters (1982). For easy Rasch analysis on a PC, see Wright & Linacre (1991).

5. TURNING SCORES INTO MEASURES

MEASURES ARE ALWAYS ANALYZED AS THOUGH THEY WERE INTERVAL

What every scientist and layman means by a "measure" is a number with which arithmetic can be done, a number which can be added and subtracted and differences from which can be multiplied and divided with results that maintain their numerical meaning. The original observations in any science are never measures in this sense. They cannot be measures because a measure implies and requires the previous construction and maintenance of an abstract quantitative system which has been shown in practice to be useful for measuring.

BUT ORIGINAL DATA IS ALWAYS ORDINAL

All data originate as ordinal, if not nominal, observations. All we can observe directly is the presence or absence of a well-defined quality. All we can count directly are numbers of classified occurrences.

All classifications are *qualitative*. Some classifications can be ordered and so become more than nominal. Other classifications, like sex, are usually not ordered, although there may be perspectives from which an ordering becomes useful such as more "male" or more "female." This does not mean that nominal data cannot have explanatory power. It does mean that nominal data are not measurement in the accepted sense of the word.

Quantitative science begins with identifying conditions and events, qualities, which, when observed, are deemed worth counting. The resulting counts are sometimes called "raw scores" to distinguish them from "weighted" or "scaled" scores. But usually they are just called "scores." As such, they are no more than counts of particular concrete events that have been observed. They are essential for the construction of measures. But they are not yet measures because they do not have the numerical properties necessary to support arithmetic.[1]

Counting is the beginning of quantification. Measurement is constructed from well-defined sets of counts. The most elementary level is to count the occurrence of a defined event. But more information can be obtained if the conditions that identify countable events can be organized into ordered categories which increase in status along an intended underlying variable. It then becomes possible to count, not just the occurrence of an event, but the number of steps up the ordered set of successive categories which is implied by the particular category observed.

When the three response categories of a rating scale are labeled: "none," "plenty," "all," the inarguable order of these labels enables their use as steps from less to more. The observation of a "none" can be counted as 0 steps up this scale. The observation of a "plenty" can be counted as 1 step up the scale.

[1] Since "scores" are so often mistaken for "measures" and then misused statistically as though they were measures, we will take the trouble to refer to "scores" as "counts" so that their empirical basis and consequent failure to be "measures" will remain explicit.

up the scale and of an "all" as 2. But this counting has nothing to do with any measures or numerical weights which might be "assigned" to the categories. "Plenty" might have been labeled "20" or "40" by the test author. But an "assertion" of such a numerical category label would not alter the fact that on this rating scale "plenty" is only observable as 1 step up from "none."

But counting steps up a set of successive categories, up a rating scale, says nothing as yet about the distances between the ordered categories. Nor is it a requirement that all items on a test employ the same category labels. It would make no difference to the step counting if, for some other item, the categories were labeled, "none," "almost none" and "all." Even though the relative meanings and implied amounts corresponding to the alternative labels are obviously different, their order is the same and so the observable step counts can only be the same. Whenever category labels share the same ordering, no matter how the labels themselves may differ in implied amounts, progress through them can only be observed as a series of single steps. The possible quantitative differences in the qualitative labels can only be discovered later by modeling the differently labeled categories separately and then using relevant data to estimate their relative difficulties.

CONFUSING COUNTS WITH MEASURES

Counts of events are on a primitive ratio scale. They have an origin at "none" and the raw unit of "one more of these kinds of things." But the events actually counted are unique rather than idealized replicates, specific rather than general, concrete rather than abstract and thus varying rather than uniform in the way they represent whatever latent variables they may be intended to imply. Sometimes the next "one more thing" implies a small increment as in the seemingly short step from "none" to "almost none." Sometimes the next "one more thing" implies a big increment as in the seemingly long step from "none" to "plenty." The relative sizes of these steps cannot be obtained directly, but must be constructed from analyses of relevant data produced by observing how these steps are used in practice.

Since all we can do in practice is to count one more step, any particular raw count is insensitive to the possibly differing implications of the steps counted. To get at reproducible empirical magnitudes for the step sizes, we must construct an abstract measuring system based on relevant parameterizations of coordinated sets of observed counts.

This construction requires a measurement analysis of the ordinal observations which comprise the initial data in every science. Even counts of time-honored units like grams, centimeters and seconds, so useful as measures in many contexts, may not function as measures in others (Thurstone, 1927). Counting the "milliseconds" it takes a student to react to a stimulus does not necessarily provide a linear measure of "student responsiveness." To construct a linear measure of "student responsiveness" based on time elapsed we must count the milliseconds taken by a relevant sample of students of varying responsiveness to react to a range of relevant stimuli. Then we must analyze these counting data to discover whether a linear measure of "student responsiveness" can be constructed from them and, if so, what its relation to "milliseconds" may be. This relationship will probably be monotonic. But it need not be linear.

FROM OBSERVATION TO MEASUREMENT

Thorndike (1926) stressed the necessity of a step from counting to measuring in 1904. Thurstone (1928) spent the 1920's developing partial solutions. Then in 1953 Rasch (1980) invented a model which, upon investigation, has turned out to be necessary as well as sufficient for the construction of measures in any science. Rasch realized that a measure must retain its abstract quantitative status regardless of the qualitative context in which it occurs. This means an item is only useful for measuring persons among whom it approximates a single fixed difficulty, and a person is only useful for calibrating items among which the person approximates a single fixed ability.

Rasch also realized that the outcomes of interactions between persons and items could never be fully pre-determined but must always involve an unpredictable element. This lead him to a probabilistic form of Guttman's (1944) requirement that the more able the person, the *more likely* a success on any item. The more difficult the item, the *less likely* a success for any person. The unique measurement model necessary for converting counts into measures follows by deduction from this requirement.

CHOOSING AN ORIGIN

"Measurement" implies a count of "standard" (hence necessarily abstract) units from a "standard" starting point. The most familiar picture of this is a distance between points on a line. There is, however, no measurement requirement to find "the" point of minimum intensity or to extrapolate a "zero mobility." It is only necessary to anchor the scale by choosing a convenient starting point or origin. Usually there are useful frames of reference for which particular choices are particularly convenient.

The seemingly non-arbitrary origin of a ratio scale is theoretical rather than practical - conceptual rather than empirical. Logarithms convert any ratio scale into an interval scale and exponentiation converts any interval scale into a ratio scale. The interval scale's origin becomes the unit of the ratio scale and the interval scale's minus infinity becomes the ratio scale's origin. The main difference between the two is arithmetical preference. Do you prefer to calculate comparisons as ratios or differences? Most of the usual statistical techniques are focused on differences rather than ratios.

The practical convenience of measuring length from an arbitrary origin, like the end of a yardstick, far outweighs the abstract benefit of measuring from some "absolute" origin, such as the center of the earth or sun. Once an interval scale is constructed from relevant counts, we can always answer ratio questions such as "Is the amount learned in first grade twice the amount learned in second grade?".

WHY RAW SCORES SEEM TO WORK AS MEASURES

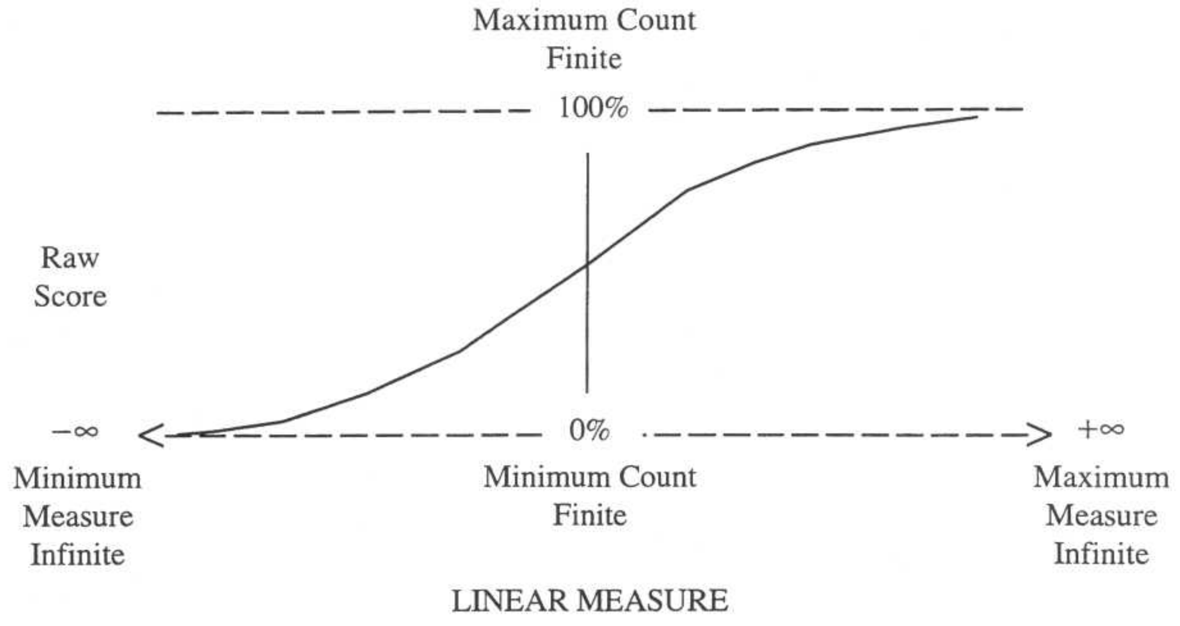
In view of the fundamental quantitative differences between counts and measures, why do statistical analyses of raw score counts and Likert rating scale labels mistaken for measures sometimes "seem to work?"

When data is complete and all data are used, then the relationship between concrete raw scores and the abstract measures they may imply becomes monotonic. This makes covariation analyses of raw scores and the measures they may imply appear similar.

Even for complete data, however, the relationship between raw scores and measures is ogival because the finite interval between the minimum observable score and the maximum observable score must extend to an infinite interval of implied measures (See Figure 5.1). Toward the center of this ogive, however, the relationship between raw score and measure, for complete data, is approximately linear. When statistical analysis of raw scores obtained from complete data is focused on this central region, conclusions will be similar to those based on genuine measures.

Figure 5.1

The relationship between scores and measures.



The monotonic curve (function) transforming observable counts (raw scores) into implied measures is approximately linear near 50% but must curve at the extremes to bring the finite boundaries of the observable counts into coincidence with the infinite extremes of the implied measures.

But the monotonicity between score and measure holds: only when data are complete, only when every subject responds to every item, only when no responses are disqualified. This means no missing data and no tailoring item difficulties to person abilities. Further, the approximate linearity between central scores and their corresponding measures deteriorates increasingly as the scores approach their observable extremes.

UNIDIMENSIONALITY

An occasional apprehension raised against the Rasch measurement model is that it "requires" unidimensionality. This objection is puzzling because "unidimensionality" is an intrinsic meaning of the term "measurement." The necessity of Rasch's model as the only method for constructing measures from observations is due to its deduction from the measurement requirement of unidimensionality. It is the undimensionability of "measurement" which requires Rasch, not the other way around.

In practice, unidimensionality is conceptual rather than factual, qualitative rather than quantitative, an idea and intuition rather than an experience. No actual test can be perfectly unidimensional. Indeed no empirical situation can completely satisfy the requirements for measurement which imply the Rasch model. But this essential "reality" is encountered and managed by every science. Physicists' corrections for the unavoidable multidimensionality they must encounter are an integral and essential part of their experimental technique.

If an educational test containing a mixture of arithmetic and reading items is used to make a single admission or graduation decision, then the examination board, however inadvertently, has decided to use the test as though it were unidimensional. This is quite beside any qualitative or quantitative arguments which might claim or demonstrate multidimensionality. The board's practice does not make arithmetic and reading identical or exchangeable anywhere but in their pass/fail decisions. Their "unidimensional" behavior, however, does prove that they have decided to make arithmetic and reading exchangeable in their decisions and hence unidimensional in their tests.

Unless each item is treated as a test in itself, every test score for which right answers are counted is a compromise between the essential ideal of unidimensionality and the inescapable qualitative "reality" uniqueness of the items used. These "multidimensionalities" are the unavoidable exigencies of practice.

Before observations can be used to support quantitative research, they must be examined to see how well they fit together and cooperate to define the intended underlying variable. Rasch measurement provides theory and technique to accomplish this. But the extent to which any particular set of observations can serve measurement is empirical. No total score can be accepted before verifying that its meaning is enough in accord with the meanings of the individual scores of its item components to lead to a measure useful for the purpose at hand. Assistance in doing this is provided by fit statistics which report the degree to which any actual observations approximate the assumptions necessary for constructing measurement, and hence quantify the numerical validity of the data.

The process of test evaluation can never be finished. Every time items are used to collect new information from new persons to estimate new measures, we must verify again that the unidimensionality requirements of the measuring system have been well enough approximated by these new data to maintain the intended quantitative utility of the measures produced. Whether a particular set of data can be used to initiate or to continue a measuring system is always empirical and must always be verified.

This empirical question can be addressed by:

1) analyzing the relevant data according to a relevantly parameterized unidimensional measurement model—a model implementing the essential requirements of measurement—a Rasch model.

2) discovering how well and in what parts these data conform to the intention to measure and,

3) examining carefully those parts of the data which do not conform and hence cannot be used for measuring to learn from them how to improve our observations, how to obtain better data, and so, how to better achieve our intention to measure.

Only after interval (linear) measures have been successfully constructed, does it become reasonable to proceed with statistical analysis in order to determine the predictive validity of measures or to compare measures produced by different tests to see if they are measures of the same thing, like inches and centimeters, or different things, like inches and ounces.

6. HOW VARIABLES ARE CONSTRUCTED

The definition of measurement found in textbooks is "the assignment of numbers to objects." Having settled that, authors retreat to other matters. Instead, we will investigate this pregnant proposition.

FOCUSING INTENTION

To make a measure it is necessary to focus attention on a single matter of interest. If we want to measure audience attendance, we focus upon "persons in the auditorium." We do not focus on other matters which might be important in other contexts. Focusing requires avoiding distraction. Other matters, important as they may be, are nevertheless, determined to be unimportant at this time. To count attendance is to give no attention to anything else. We act as if nothing else existed.

Measurement requires a singular intent which, once selected, immediately requires a withdrawal of attention from all other matters. Naturally, the choice of this singular attention is critical. In our simple example it is easy to determine our focus of attention. In the pursuit of science, however, it is seldom simple to determine what we should attend to. But "what to attend" is the decision that must be made in order not to be overwhelmed by voluminous, unsorted perceptions.

COUNTING AND MEASURING

To make measurement beyond counting the persons in the auditorium, we need a model. Measurement is reached by way of analogy. Some analog is required to enable our task. We measure time by moving clock hands. We measure length by concatenating unit lengths into rulers. All measurement utilizes analogy. The natural numbers (positive integers) are always the starting point for the model we apply to our problem. We assign each "person" a number as we note their presence one-by-one. We "count" by reciting the arithmetic counting model as we focus upon each next person. Having observed the last person, we note the numeral associated with that last individual and designate the count of attendees by that numeral. If someone else enters the auditorium, we count one more and use that next successive numeral to signify the new total. The persons in the auditorium are the objects manifest in our experience. The numerals taken from the natural number system are the unit values from the model we applied to accomplish the counting.

Counting is so familiar that we take for granted how it is done. Counting objects is so universal that it is frequently assumed that the "count" is automatically a "measure." This assumption is so common that it usually goes unquestioned. But counts are not necessarily measures. Measurement depends upon whether the mathematical model on which it must be based can provide the desired outcome. We saw that the natural numbers served to count. But are they necessarily measures?

Using the natural numbers in our example brings with it the observation that distances between adjacent natural numbers are taken to be equivalent, and whole, not fractionated. The number "space" between any two adjacent numbers 2 and 3, 19 and 20 or 164 and 165 is equal in our counting to a same "one more person."

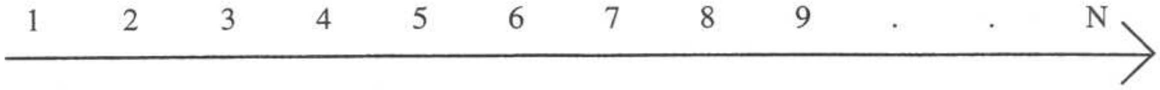
When measuring length with a ruler, we also use the equal steps of the real numbers as our model. But fractional parts are permitted. Any unit or fractional part of a unit along the real number line is equal in magnitude to any other like unit or fractional part anywhere along the line.

The natural numbers used in counting are logically consistent. But does the application of these numbers to every experience automatically endow the objects of that experience with similar numerical properties? This question often goes unanswered because it is unrecognized. Since numbers are logically coherent, the objects to which they are applied are often assumed to possess similar logical coherent numerical characteristics. But this association can not be assumed. It must be constructed, demonstrated and maintained. Unfortunately, this demonstration is rarely done and objects are usually naively assumed to possess the same numerical characteristics as those of the number system that is used to count them.

Since measurement is made by analogy, the model chosen for measurement cannot itself substantiate the numerical qualities assigned to the objects. We have to demonstrate that the relationship of the objects to the model is consistent with our use and understanding of the numerical result and demonstrate this connection by means of the analogy.

While counting is the model for measurement it is seldom measurement itself. We begin by assigning numerals to objects. But we must proceed in a manner which is systematic and reproducible and which allows us to validate continually the correspondence between the objects addressed and their counts and between the counts obtained and the measures they may imply.

The count of attendees can be represented directly on the number line.



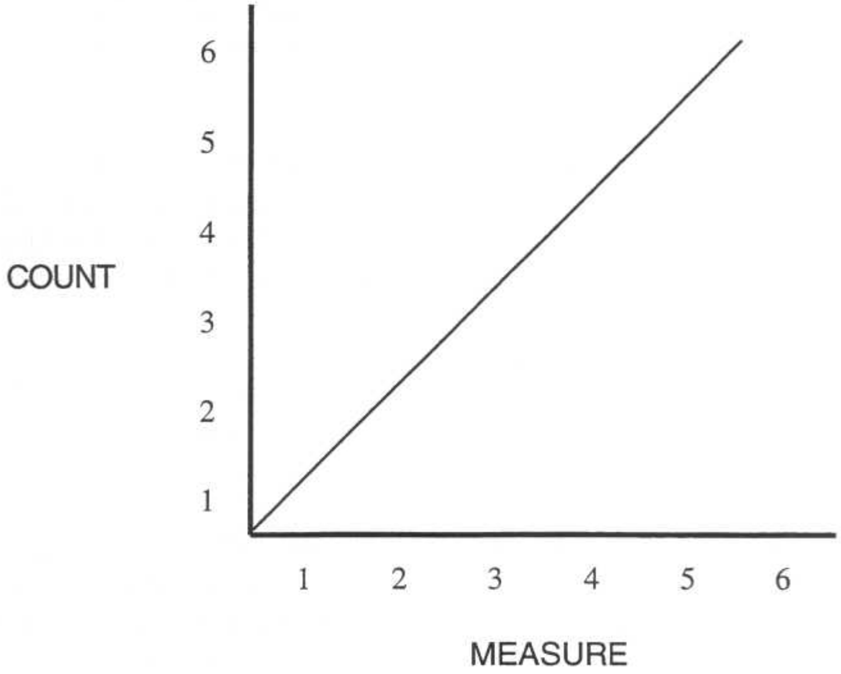
The number line gives us a model which appears to solve our attendance problem. The model of the number line (our variable of interest) specifies an equal unit distance between the numerals. The successive count of persons parallels movement along the line by equal counts. An additional person adds another equal count to the line. The number line made from our counts is associated with the "length" registered on a "number of persons" variable.

1. Successful measurement requires focus upon an attribute of intent to the exclusion of all other (often equally interesting, but not at this moment relevant) attributes. We cannot progress until we have made such a choice. The choice is both intent to attend and also intent to disregard.

2. A coherent measurement analog is necessary to accomplish the measurement task. The one we used in our example was the natural number model of arithmetic progression. Application of this model to our problem gave us a useful answer. We needed an analog to solve our measurement problem: establishing a one-to-one correspondence between counting individuals, seated in the auditorium and numerals on a number line. The idea of a number line gives us a picture of the variable of intent in a way that facilitates understanding, application and interpretation.

3. Determining the count of people in the auditorium answered our question, "How many are here?" Was it a measure? Yes, but only when demonstrated by analogy to be associated in a meaningful way to our number line - the variable. Use of the number line as the model leads this time to using the count to measure the "amount" of persons expressed in a linear form. Consequently, the variable of attendance can be visualized as a line with our total attendance indicated by a position or "amount" on the line. Carrying around such "lines" of amount is quite unnecessary to say nothing of inconvenient. We revert back to counts to express this length in a more compact form. But our understanding of the variable form is the line. The natural number line is the full expression of the measure of persons, not the count. We use the count as shorthand for the measure.

   Our use of the number line and its associated unit characteristics provides, by analogy, a clear correspondence between the count and measure. This relationship can be seen as:



   where a straight line at 45 degrees with slope of "one" associates counts to measures. But even though the observed counts end up as the same as the measures in this example, the counts are not in themselves measures even in this case. They became so here through their analogy with the number line.

4. Validity of this approach can be demonstrated by replication. The individuals counted can be seen as alternate equal instances of a single individual - the experiential unit of interest. We make our count and measure by enumeration. The measurement process is unaffected by where the count begins i.e. from the back, front or side of the auditorium. Precision depends only on accuracy in counting and not on any other aspect of the process. The count and the resulting measure are unaffected by the other matters of gender, race, height, weight.

We now proceed to more sophisticated problems, keeping in mind the steps encountered in our simple example. Although our example was simple, the tasks were sophisticated. We count with such ease and so routinely that we are no longer aware of the steps we take nor of the confusion that can result between determining how to make counts and how to make measures. The most difficult task to recognize is how to determine measures from counts. This is because counts are what we always begin with to determine amounts. Counts can be associated with measures only by establishing a useful relationship. We need to generalize this by deliberating upon the measurement process, specifically upon constructing a variable.

CONSTRUCTING A VARIABLE

A "variable" is a line with direction - an arrow. The direction implies "more" of the variable. An "amount" is a distance along the line. More is "more" distance along the line. Length is the example that comes most easily to mind and eye. More inches, feet, yards, is easily visualized as more distance along the ruler being applied to the object. Interpretation of such a variable is straightforward and easily demonstrated visually. This keeps the process forthright and observable. There is no mystery, no algebra. Calculations may be complex but the outcome can be seen. The line conceptualizing the variable makes the measurement task visible.

The "variable" must establish and maintain a single line of inquiry. That line of inquiry has to be operationally determined and reproducible at will. To be useful its determination must also be insightful and rich. Conceptualization as a one unidimensional line is necessary, if our thinking is to progress usefully. When this is not done, confusion and frustration result. Then there is no simple way to determine what the variable should be.

Our intent specifies what we will address and not address. Our choice not to address matters is selective and not unalterable. We can change our mind. What we cannot do is to pursue two things at once and remain clear about what we are accomplishing.

A measure of "attendance" need not be a measure of speaker "attraction." It may be an interesting problem to discover whether attendance is a measure of "agreement" with the speaker. But attendance itself, does not directly answer that question. We confuse ourselves when we pose unclear questions. The task is to separate these different issues: attendance, attraction, agreement into clear dimensions and to do the best we can to measure each of them, not confuse our thinking by allowing simultaneous consideration of contrasting matters.

UNIDIMENSIONALITY

Human behavior is clearly complicated and so might be thought of as multidimensional. But its scientific investigation and understanding cannot begin multidimensionally. The investigation must be systematically built up in successive stages. Each measurement stage requires that a single unidimensional variable be established. When several dimensions have been successfully developed, the study of their association by statistical analysis such as multiple regression, may bring out relationships. When unclear, inadequately identified "dimensions" are subjected to statistical analysis before establishing their measures, however, then interpretation is obscure.

The variable is an idea, an intention, something we want to realize, a construction made from collecting and selecting observations. While human experience is complex, when making measures we decide to isolate one major ingredient and not to become overwhelmed by the endless and enumerable possibilities that are impractical to examine simultaneously. Successful measurement, like good science, is always practical. The well-defined measures are the useful ones. Indeed, they often become so useful that their origins are forgotten and, because of their familiar utility, we take them for granted.

If we did not have the natural number model for counting, we could not enumerate and so measure the audience. We could only get lost in the problem of how to enumerate. A well formulated variable will demonstrate its utility time and again and its recurrent utility will be the demonstration of its validity.

Variables are constructed out of similarities noticed, the "replications" we infer in experience. Although nothing actually recurs exactly, we take it that a "thing" of our defining does reappear again and again although in different guises. It is our discernment of what to specify and what to ignore that distinguishes a useful variable from a useless one.

DIALOGUE OF INVENTION AND DISCOVERY

Variable construction is a dialogue between invention and discovery. We observe out of experience. We abstract an aspect to capture an essence and invent a dimension for this essence, wording it carefully to avoid becoming overwhelmed by the multiplicity of the provoking experience. We construct this singular idealization knowing that it is only a model representation and not a real thing. Good variable construction is a mix of discerning observation, creative intention and careful disregard.

The idea of hot/cold is useful in enumerable applications. We measure temperature by analogy using lengths of liquids in uniform tubes.

We experience wind directly. We make a vane to "experience" wind for us and to show its direction. We attach cups to rotor arms to capture air, connect the rotating arms to a dial and make, by analogy, a measure of wind velocity along the arc of the dial.

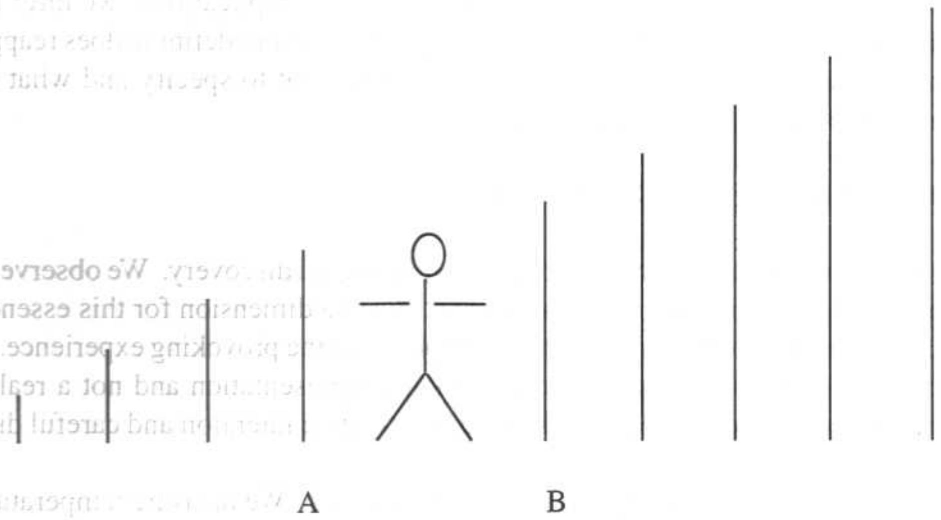
We can combine temperature and wind speed with the heat loss of evaporation to produce a new dimension - windchill. This construction we find is immediately experienced. Low temperature days without wind are experienced as warmer than windy days recording a higher temperature.

Useful variable definition is clear thinking. But it also requires intuitive leaps. We must go beyond previous experience to capture the possibility of new experiences of which we have awareness, but cannot "see" with clarity. Variable definition combines intuition and reason into a constructive fiction that embodies an essence of experience liberated from the infinite complexity of total experience.

BUILDING A VARIABLE

Imagine a person and a collection of sticks of varying lengths. Our problem is to measure the height of the person from the lengths of the sticks. The logical way to proceed is to stand each stick next to the person and keep track of which sticks exceed the person's "height" and which do not. We aim to "capture" the person in the midst of the available sticks. Should no stick exceed the person's height (or all exceed his height), we will need more sticks longer (or shorter) and, cannot capture this person among the available sticks.

When we have sticks above and below the person, we can locate the person between two of the sticks. When sticks themselves are compared one to another just as we compared them to the person and we arrange the sticks in their order of length we can locate the person between a particular pair of adjacent sticks. One stick of the pair is shorter than the person (as are all the successively shorter sticks) and one stick is taller than the person (as are all the successively taller sticks). We have positioned the person among the available sticks and ordered the sticks with, in this case, our person exceeding four shorter sticks but exceeded by five longer ones. The length of this person can be abstracted as somewhere between sticks A and B. That is, the "measure" of this person's "height" is defined by these two sticks.



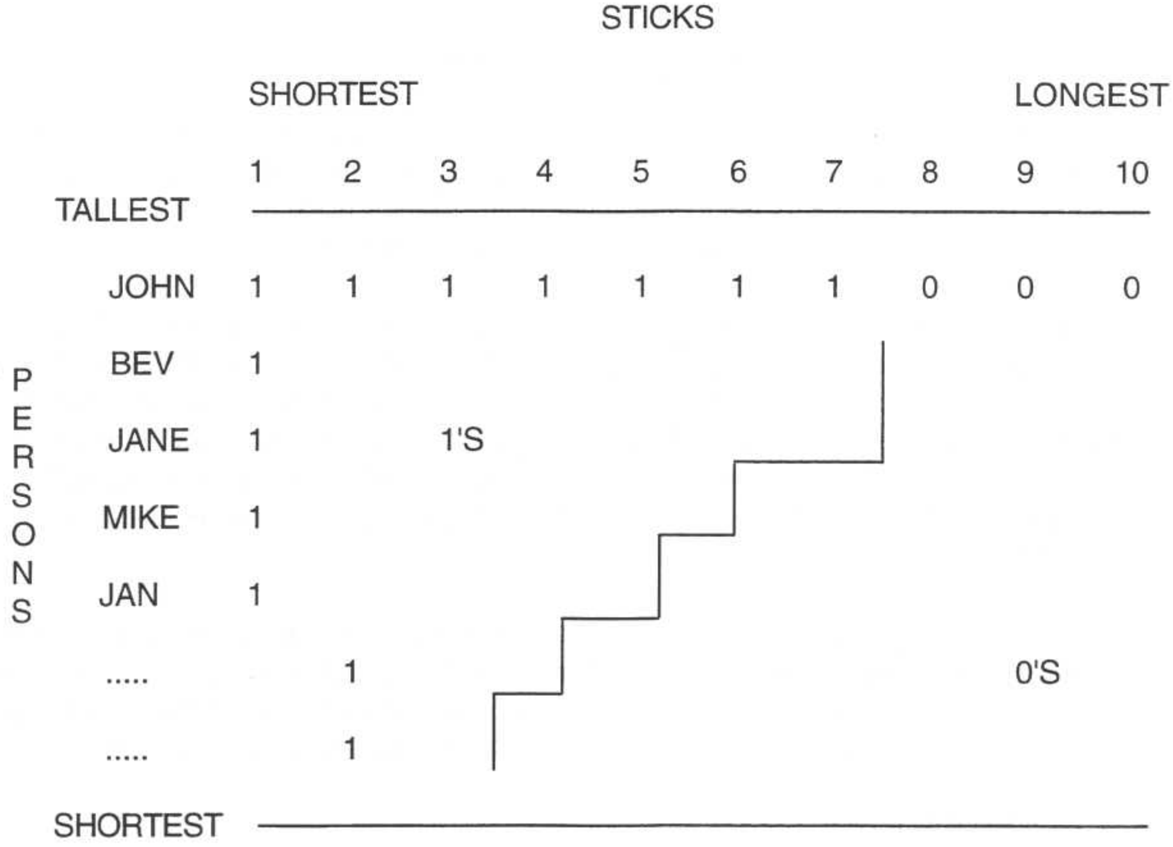
We have constructed a variable of sticks by arranging them according to length and then applying the sticks to the measurement of a person. Given more persons we can compare them to one another or more easily to each of the sticks and so arrange all the persons among the sticks, above those sticks they exceed and below those sticks they fail to reach.

To simplify the machinery and keep the arrangement of sticks and persons orderly, we can dispense with actually placing all persons next to all sticks and instead designate each stick by an ordinal number and locate each person halfway between the numbers of the pair of sticks that bracket that person.

By keeping a record of the persons who exceed (or fail to exceed) each stick number we can produce a picture of the relation between persons and sticks. This strategy of "sticks exceeded" is the basis for our variable "height." Each person "tested" is compared to each stick with a tabulation made for each stick according to the number of persons exceeding it. In this manner we "calibrate" our sticks and use them to measure persons according to the criteria of which sticks a person exceeds.

The development of any variable proceeds from the strategy illustrated in this example. No matter what the intent of our variable, the procedure for calibrating the agents (sticks) and measuring the objects (persons) always begins this way.

RECORD OF PERSON X STICK COMPARISONS



GUTTMAN AND RASCH RESPONSE PATTERNS

The above matrix of 1's and 0's when it contains no 0's in the pattern of 1's nor any 1's in the pattern of 0's is a "Guttman Scale." Such a pattern is what Guttman called the necessity for a score to have meaning. The pattern is of perfect order. The difficulty is in constructing real tests such that a perfect order is achieved. Rather than seeking a perfect "deterministic" pattern in the data and failing to construct it, Rasch measurement expects an imperfect "stochastic" order and proceeds to evaluate it in terms of perfectly ordered probabilities. In Rasch's model it is the probabilities that form the perfect Guttman pattern and not the responses.

Thus when persons $n = 1, N$ are ordered by the magnitudes of their measures B_n and items $i = 1, L$ are ordered by the magnitudes of their calibrations D_i, the conjoint order of the Rasch probabilities that person *Pni* will succeed against item *i*.

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

forms a perfect Guttman pattern.

SCORES AND MEASURES

Why is it that scores are counts, but not necessarily measures?

When counts are accumulated they are made up of concrete, tangible objects. The objects counted are interpreted to be exchangeable, as though they were identical for the purpose of counting. But in their concreteness they are manifestly not identical nor in any exact way exchangeable. Indeed, upon close examination the objects counted can always be seen to be different, unique.

Counting is a daring fiction. Counting is also a necessary fiction. It is the essential fiction of "againness," of deciding to pretend we are encountering a "same thing" again and again, even though we know that no two encounters are ever the same. It is the fiction of "recognition," of being able to interpret a future in terms of past experience. Without the fiction of "againness," "time" is meaningless, indeed thought is impossible. Since counts build on fiction, the choice and nature of the fiction is a critical step on the way to the counts becoming useful. The invention and governance of what to count is the Observation Model.

Once we acknowledge that any concrete objects counted have been fictitious in their counting, we can see that their actual concrete differences are clumsy and thus think it better to count perfect ideas rather than imperfect objects, that is to count ideally equal, exchangeable units, since we are making it up in any case. Counting "exchangeable things" is scoring. Counting "ideal units" is measuring.

DIFFERENCES AND RATIOS

If real counts are on the way to ideal measures, how do we go the rest of the way? For that we must ask what kind of comparisons we intend to construct and why. The comparisons suited to visual analysis are distances, differences, linear comparisons. This suggests that any comparison we wish to visualize for comprehension must be made in terms of differences and not, for example, ratios. Plots of ratios confuse our visual ability to analyze distances, and so deny access to those pictures worth a thousand numbers.

When considering counts, however, we realize that the comparisons which have meaning depend on the size of the counts. When we are concerned with small differences between large counts we find ourselves interested in balancing their differences. We equate these large counts by small additions and subtractions to reach equity, to enable fair trade, to do business. Then we are interested in differences of counts.

But when we encounter large differences between counts, it is not differences which emerge as the usefully invariant comparisons. Now it is the ratio of the counts that matters. What does it take to double our yield, triple our gain? Comparisons of wide ranging counts, like income, are more useful as ratios than as differences.

The next step, then, is to turn ratios of counts into differences of measures. This is where "logs" and their inverse "exponents" enter and why loglog (and semilog) plots became (and still are) important in the visual analysis of numbers.

Counts are concrete, history. But our use of counts is seldom to reconstruct the past. We are most interested in counts to make inferences about the future. Analysis of data is seldom just for historical description. It is usually for prediction. To predict we must accept scores as crude indicators of what might become and so manage them accordingly, in terms of what they probably mean.

The utility of counting brings to our attention the convenience of equal units. Why not construct measures of magnitude in equal units? To do so accepts the sampling status of the counts and brings us to the next step of finding an estimate from the sample of what the counts might usefully imply.

Since we are estimating an abstract idea of a magnitude from a count of concrete experiences, we encounter a need for a concomitant second estimate, this time of the expected uncertainty or error of the estimated magnitude, of the improbability of our inferences. This brings up "standard errors."

We also must figure out an answer to the question as to whether the collection of experiences or observations which have been counted on to imply estimates of magnitude and error are in their own details consistent with the two estimates they have led to? This brings up a third statistic, "fit."

When we use the estimates of magnitude to go back and predict the counts from which they arose, to calculate what observations would be expected were they entirely caused by the magnitudes we have estimated, we can discover for each observation how well it fits into the measures for which we intend to use these data.

The study of "fit", particularly the identification of outstanding misfit, is our chief source of new information about the world of possible experience, our chief opportunity for discovery. The observation model by which we define what to count and the measurement model by which we construct estimates of ideal magnitudes from the crude concrete counting are the inventions of measurement. The misfits that then appear are the discoveries of measuring.

The growth of science, indeed of mind, arises out of an evolving dialogue between invention and discovery - between the reassurance that we know what we are doing because our inventions work and the provocation that we must not know everything about what we are looking for because we are surprised by what we find. Constructing variables engenders an interaction of experience and idea, a dialogue between invention and discovery, that is the life force of science and mind.

7. FIT ANALYSIS

The Rasch model specifies the relation that must dominate what happens when a person takes an item for the resulting responses to be useful for measurement. A complete analysis must include an evaluation of how well the data fit this essential specification. If a person answers the hard items on a test correctly but misses several easy items, we are surprised by the resulting implausible pattern of incorrect responses. While we could examine individual records item-by-item to determine this kind of invalidity, in practice we want to put such evaluations on a systematic and manageable basis. We want to be specific but also objective in our reaction to implausible and hence invalid observations.

Even when a particular application tends to fit the measurement model, we cannot predict in advance how well new items or old ones will continue to work in every new situation to which they might be applied. We cannot know in advance how all persons will always respond. Therefore, if we are serious in our intention to measure, we must examine every application to see how well each new set of responses corresponds to our measurement intentions. We must evaluate not only the plausibility of the sample of persons' responses, but also the plausibility of each persons' responses to their set of items. To do this we must examine the response of each person to each item to determine whether that response is consistent with the general pattern of responses observed.

We begin fit analysis by examining the data resulting from the administration of a test of L items to a sample of N persons producing an N x L matrix of responses with every row consisting of the responses of each person n to the L items and every column consisting of the responses of the N persons to each item i. When the responses are dichotomous, the resulting matrix will consist of correct ($X_{ni} = 1$) and incorrect ($X_{ni} = 0$) responses.

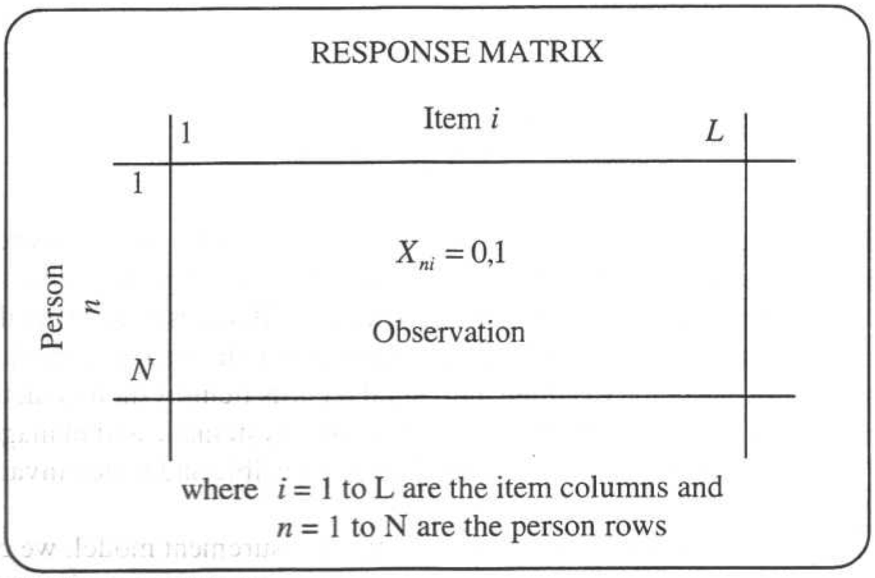
The construction of useful measures and calibrations does not require that these data be complete. The particular items addressed by each person can vary as long as there is a sufficient network of overlaps to connect the entire matrix. For simplicity here, however, we will carry out the explanation as though data were complete.

From the N x L data matrix of $X_{ni} = 0$ or 1 we count the item scores S_i and person scores R_n used to estimate the abilities of persons B_n and the difficulties of items D_i. Procedures for this are explained in Best Test Design (Wright and Stone, 1979, pp. 28-65).

This chapter explains the analysis of fit (Wright and Stone, 1979, pp. 66-82 and 165-181).

RESIDUAL FROM EXPECTATION

To evaluate fit we compare the observed person and item responses X_{ni} to the expected values P_{ni} that are determined for them by the measurement model. The expected value of the dichotomous observation X_{ni} is $P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$.


RESPONSE MATRIX
Item i
1
L
1
Person
n
N
$X_{ni} = 0,1$
Observation
where $i = 1$ to L are the item columns and
$n = 1$ to N are the person rows

A consequence of the Rasch model is that the person right answer count, a total score for an individual, contains all of the information needed to measure that person and the item right answer count, a total score for an item, contains all the information needed to estimate the difficulty of that item. That is to say, that right answer counts are sufficient statistics for estimating person measures and item calibrations.

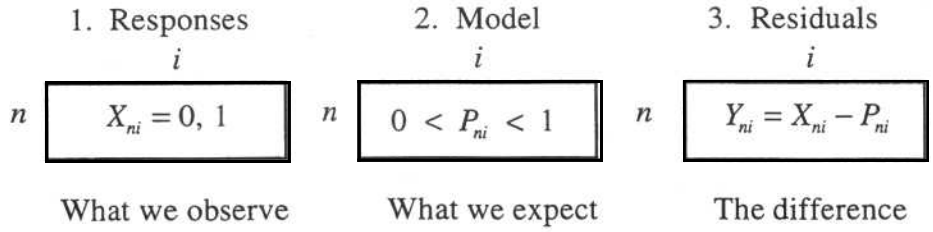
RASCH MODEL EXPECTATIONS

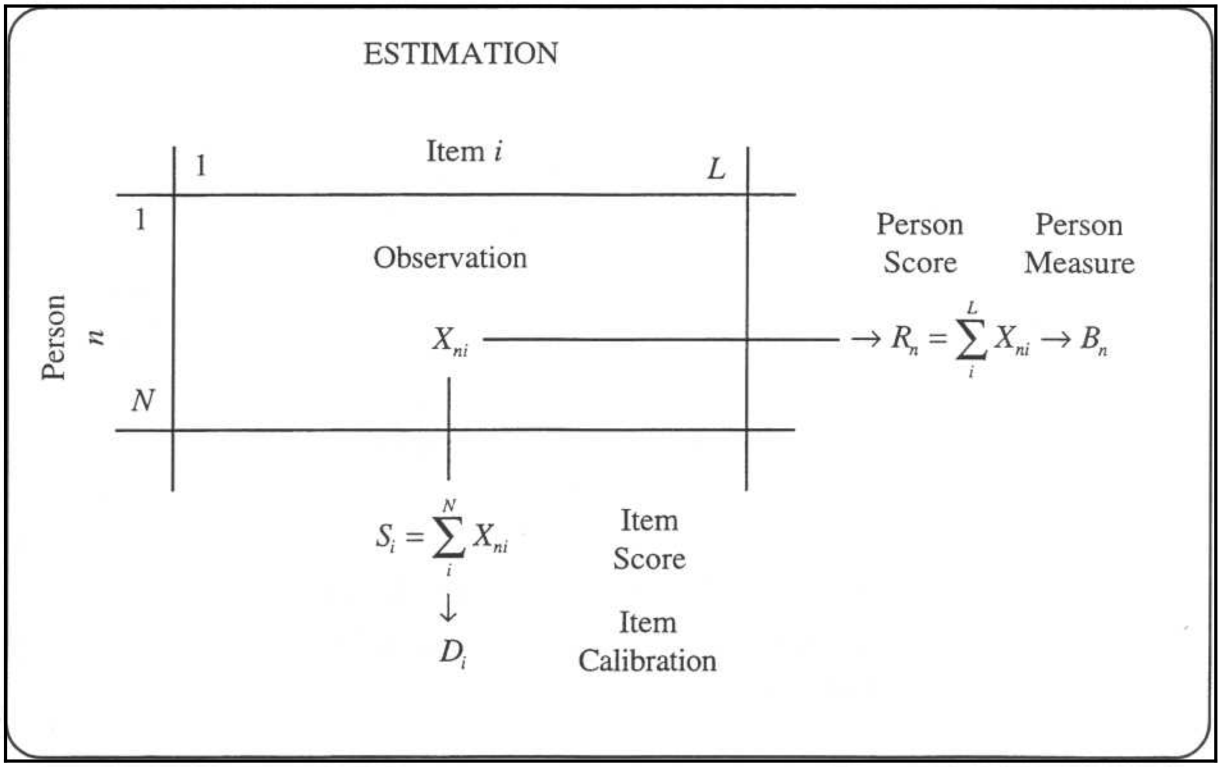
The Rasch model is derived from the requirement that person measures and item calibrations be separately estimable. This requires that (1) a more able person always have a greater probability of success on any item than a less able person, and (2) any person always be more likely to do better on an easier item than on a harder one. Fit analysis evaluates the extent to which particular data serve this fundamental requirement for measurement. Fit analysis shows us the extent to which any data can be used to construct measures. Each data analysis must include an evaluation of how well those particular data fit the expectations which measurement requires.

When an observed pattern of responses shows significant deviation from measurement expectations, we can use the particulars of the measurement model together with the person and item estimates to calculate a statistical index of unexpectedness for any particular response or any subset of responses including all of the responses to a particular item or all of the responses made by a particular person.

DETERMINING FIT

The procedure for analysis of fit involves the three steps:





R_n is the sum $\sum_i^L$ of the person responses X_{ni} over item $i = 1, L$.

B_n is the person measure estimated from R_n.

S_i is the sum $\sum_n^N$ of the item responses X_{ni} over persons $n = 1, N$.

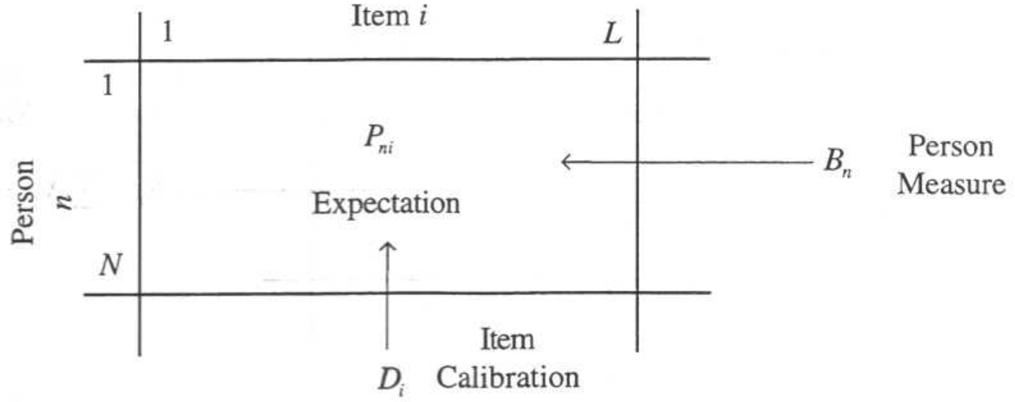
D_i is the item difficulty estimated from S_i.

EXPECTATIONS

To observe response plausibility, validity or fit we calculate the difference $(B_n - D_i)$ between the estimates of person ability B_n and item difficulty D_i for each person n and item i. When this difference is positive it means that the item should be easy for the person. The more positive this difference, the easier the item is expected to be and hence the greater our expectation that the person will succeed on that item and $X_{ni} = 1$.

When the difference is negative, however, the item should be difficult for the person. The more negative the difference $(B_n - D_i)$ becomes, the more difficult the item should be for the person and hence the greater our expectation that the person will fail on that item and $X_{ni} = 0$.

PREDICTION



$P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$ and B_n and D_i are the person measure and item calibration for person *n* and item *i*.

Chi-square and mean square goodness-of-fit statistics can be constructed from the residual difference $Y_{ni} = X_{ni} - P_{ni}$ between the observed X_{ni} and its expectation P_{ni}. This residual quantifies the fit between each person *n* and each item *i*.

The model estimates the expected value or probability of dichotomous response $X_{ni} = 1$ as:

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

where B_n = the estimated ability measure of person *n*

D_i = the estimated difficulty calibration of item *i*

and P_{ni} = the probability that $X_{ni} = 1$.

RESIDUALS

The probability P_{ni} is an estimate of the expected value of instances of the response $X_{ni} = 0,\ 1$. The expected binomial variance of these instances of X_{ni} around P_{ni} is estimated by $Q_{ni} = P_{ni}(1 - P_{ni})$. These expectations P_{ni} and variances Q_{ni} can be combined to form a standardized residual Z_{ni} for each X_{ni}: $Z_{ni} = (X_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})] = (X_{ni} - P_{ni}) / Q_{ni}^{1/2}$.

We estimate this standardized residual Z_{ni} by subtracting from the observed X_{ni} its estimated expected value P_{ni} and dividing the difference by its expected standard deviation $Q_{ni}^{1/2} = [P_{ni}(1 - P_{ni})]^{1/2}$.

This standardized residual Z_{ni} has a logistic distribution with an expected mean of 0 and a variance of 1. The reference values of 0 and 1 help us to evaluate the extent to which the standardized residuals deviate from their model expectations.

Examination of residuals shows us whether we can proceed to use the items to make valid measures or whether further work on the items are required in order to bring the testing elements into line with the intended plan. Examination of person residuals indicates the extent to which persons have responded to the test in the expected manner. Since X_{ni} takes only the two values of "0" and "1", the two values for the standardized residuals can be expressed in terms of the estimates B_n and D_i and the observed response X_{ni} .

Thus, $Z_{ni}^2 = [\exp(B_n - D_i)]$ can be used to indicate the unexpectedness of an incorrect response X_{ni} 0 to a relatively easy item, while $Z_{ni}^2 = [\exp(D_i - B_n)]$ can be used to indicate the unexpectedness of a correct response $X_{ni} = 1$ to a relatively hard item. These two expressions can be combined into one as $Z_{ni}^2 = \exp[(2X_{ni} - 1)(D_i - B_n)]$.

The values of this Z_{ni}^2 can be ascertained for each X_{ni} of 0 or 1 and then accumulated over items to evaluate the plausibility of any person measure, or over persons to evaluate the plausibility of any item calibration.

To evaluate any unexpected response X_{ni} we quantify its unexpectedness from the difference between the ability measure of that person B_n and the difficulty calibration for that item D_i. For example, an unexpected incorrect response of $X_{ni} = 0$ associated with a person ability $B_n = -1.2$ and an item difficulty of $D_i = -3.9$ produces a difference $(B_n - D_i)$ of $[(-1.2) - (-3.9)] = +2.7$ and a squared standard residual of $Z_{ni}^2 = \exp(2.7) = 14.9$.

We associate unexpected incorrect answers $X_{ni} = 0$ with $(B_n - D_i)$ and unexpected correct answers $X_{ni} = 1$ with $(D_i - B_n)$ because when the response is incorrect, and $X_{ni} = 0$, then the index of unexpectedness is $Z_{ni}^2 = \exp(B_n - D_i)$, but when the response is correct, $X_{ni} = 1$, then the index is $Z_{ni}^2 = \exp(D_i - B_n)$.

Unexpectedness is always marked by a positive difference, either $(B_n - D_i)$ or $(D_i - B_n)$. The values for Z_{ni}^2 can be looked up in Table 7.1 which gives either values of $Z_{ni}^2 = \exp(B_n - D_i)$ for unexpected incorrect answers $X_{ni} = 0$ or values of $Z_{ni}^2 = \exp(D_i - B_n)$ for unexpected correct answers $X_{ni} = 1$.

Thus, the entry C_x in Column 1 of Table 7.1 is either $C_0 = (B_n - D_i)$ when $X_{ni} = 0$ and the response is incorrect or $C_1 = (D_i - B_n)$ when $X_{ni} = 1$ and the response is correct.

Column 3 of Table 7.1 gives the improbability of the observed response $P_{ni} = 1/(1 + Z_{ni}^2)$.

Table 7.1

Evaluating Unexpectedness

1. Difference Between Person Ability and Item Difficulty C_x	2. Squared Standardized Residual $Z^2 = \exp\ C_x$	3. Improbability of the Response $P = 1/(1+Z^2)$
-0.6, 0.4	1	.50
0.5, 0.9	2	.33
1.0, 1.2	3	.25
1.3, 1.5	4	.20
1.6, 1.7	5	.17
1.8, 1.8	6	.14
1.9, 2.0	7	.12
2.1	8	.11
2.2	9	.10
2.3	10	.09
2.4	11	.08
2.5	12	.08
2.6	13	.07
2.7	15	.06
2.8	16	.06
2.9	18	.05
3.0	20	.05
3.1	22	.04
3.2	25	.04
3.3	27	.04
3.4	30	.03
3.5	33	.03
3.6	37	.03
3.7	40	.02
3.8	45	.02
3.9	49	.02
4.0	55	.02
4.1	60	.02
4.2	67	.02
4.3	74	.01
4.4	81	.01
4.5	90	.01
4.6	99	.01

This improbability P_{ni} provides a significance level for the null hypothesis of fit for any particular response. With our example of $(B_n - D_i) = 2.7$ we have a significance level of $P_{ni} = .06$ against the null hypothesis that the response of the person to this item is consistent with the model.

When the Z_{ni}^2 are accumulated over items for a person or over persons for an item, simulations have shown that the resulting sums can be usefully evaluated by chi-square distributions with L - 1 degrees of freedom for a person and N - 1 degrees of freedom for an item.

These fit statistics are called "outfits" because they are heavily influenced by outlying, off-target, unexpected responses. A useful alternative is to weigh residuals by the information they contain so that the fit statistics are information weighted or "infits" and hence focus on inlying, on-target, unexpected responses. The calculations for each type of fit statistic are outlined in the summary section.

SUMMARY

The following formulas summarize the calculations for the analysis of dichotomous fit.

Observed Response: $X_{ni} = 0,\ 1$

Expected Response: $P_{ni} = \exp(b_n - d_i) / \left[1 + \exp(b_n - d_i)\right]$

Response Variance: $Q_{ni} = P_{ni}(1 - P_{ni})$

Score Residual: $Y_{ni} = X_{ni} - P_{ni}$

Standardized Residual: $Z_{ni} = Y_{ni} / Q_{ni}^{1/2}$

Fit Mean Square:

Outfit:

$$U_n = \sum_i^L Z_{ni}^2 / L$$

$$U_i = \sum_n^N Z_{ni}^2 / N$$

Infit:

$$V_n = \sum_i^L Y_{ni}^2 / \sum_i^L Q_{ni}$$

$$V_i = \sum_n^N Y_{ni}^2 / \sum_n^N Q_{ni}$$

Fit Standard Errors:

Outfit: $SE_u = \left[\Sigma(1/Q - 4)\right]^{1/2} / \Sigma 1$

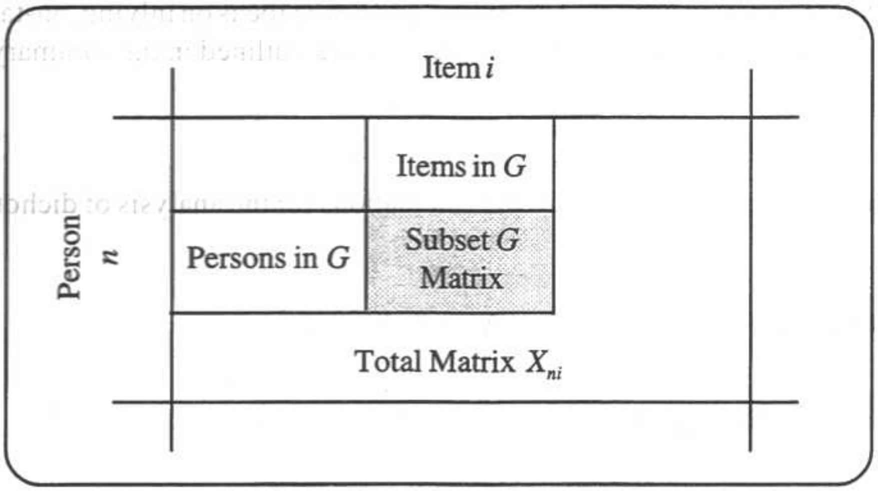
Infit: $SE_v = \left(\sum Q - 4\sum Q^2\right)^{1/2} / \sum Q$

Fit Standardization:

$$T_u = \left(U^{1/3} - 1\right)\left(3 / SE_u\right) + \left(SE_u / 3\right)$$

$$T_v = \left(V^{1/3} - 1\right)\left(3 / SE_v\right) + \left(SE_v / 3\right)$$

Fit analysis can also be done for subsets of person-item responses taken from the total matrix of responses. In this manner we can evaluate the responses of any person or subset of persons to any item or subset of items or evaluate any item or subset of items by any person or subset of persons:



For the analysis of any subset (G) of the data matrix of X_{ni} use the following formulas:

Logit Bias: $G = \sum^{G} Y_{ni} / \sum^{G} Q_{ni}$

Standard Error: $SE_G = \left(\sum^{G} Q_{ni}\right)^{-1/2} = 1 / \left(\sum^{G} Q_{ni}\right)^{1/2}$

Mean Score Residual in G: $Y_G = \sum^{G} Y_{ni} / \sum^{G} 1$

Infit Noise in G: $V_G = \sum^{G} \left(Y_{ni} - Y_G\right)^2 / \sum^{G} Q_{ni}$

Standard Error: $SEV_G = \left(\sum^{G} Q_{ni} - 4\sum^{G} Q_{ni}^2\right)^{1/2} / \sum^{G} Q_{ni}$

where $\sum^{G}$ means summed over *n* and *i* in *G*.

The analysis of fit evaluates how well our data cooperate with the construction of measurement. Analysis of fit gives us a tool to monitor the responses of persons and items. We can evaluate any set of items or persons to determine where misfit occurs. Fit analysis provides the quality control technique required to supervise and validate test items and person responses. When fit is within our guidelines, we have the control required to feel confident about item calibration and person measurement. When misfit is discovered we can locate its occasions and begin further study of the items or persons involved.

The analysis of fit is never completed because continued use of the instrument requires that we constantly monitor item and person responses to maintain quality control.

8. IDENTIFYING ITEM BIAS

The past twenty years has witnessed increasing concern about test bias. This has produced a substantial amount of literature. A few of these articles actually deal with the critical issues in test bias, but most of what has been published is ill-suited to actual practice.

Psychometricians have tried to deal with the technical issues of test bias from many perspectives. This chapter looks at item bias from the point of view of Rasch measurement and shows how item bias can be detected and dealt with in test practice. The techniques we describe are straightforward and easy to apply. They work with most measurement applications. Were these techniques to be routinely used, whatever item bias actually existed would be clearly identified and could be easily monitored and controlled.

There has been a fundamental error in thinking about bias which has lead to confusion over what bias is and, hence, how to detect it. This error occurs whenever the detection of any "difference" at all in test scores is immediately assumed to signify bias. The error typically occurs when contrasting samples are compared and found to be different in their measures. Examples of this confusion are differences in mean test scores between demographic types like males and females or blacks and whites. When such a difference is identified, the accusation is made that "bias has been found to exist." But "differences" of this kind do not signify bias.

The fallacy in such accusations can be illustrated by simple examples which show that "differences" in measures are not proof of measurement bias. Suppose we weigh two groups: professional football linemen and professional jockeys. When we compare the mean weight of these two groups a great difference in pounds will be found and indeed is expected. Would we then infer that this observed difference in weight indicates that the scale used for weighing these equally professional athletes was biased against jockeys? Similarly, if we compared the height, weight or general skills of 18 year olds with those of 8 year olds, would any differences found in favor of 18 year olds be taken to indicate bias in the measuring instruments? If the average height of 8 year olds is less than that of 18 year olds, is the ruler biased? Of all the numerous practical illustrations of this type that we could cite, none would cause us to conclude that the observed differences were indicative of biased measuring tools. Hence we must realize that differences in measures do not necessarily signify bias. We must look further into the question of bias for its *necessary* indicators.

The phenomena that is actually indicative of bias is significant and persistent *interaction* between some *but not all* persons and some *but not all* items. When a measuring process encounters unexpected differential effects within the replications necessary to estimate a measure, this *unmodeled interaction* is an indication of possible bias. *Differential interaction* between some items and some persons produces results which cannot be predicted within the intended frame of reference. Interaction confounds the intended interpretation of test scores. Interaction confuses interpretation because we can no longer base our measures upon the replications of the variable implied by the measuring instrument, but must, instead, take into account a second, poorly defined variable which differentially affects the manifest relations between some persons and some items.

Sometimes these interactions are substantial enough to spoil the resulting test scores and sometimes they are not. Suppose we give a sixth grade student an easy arithmetic word problem to read and solve. If he fails to give a correct answer, is it due to a reading problem, or to difficulty with arithmetic or to both or to something else? To identify an answer as incorrect without reviewing the probability of it's being incorrect and, when the answer is improbable, diagnosing the reasons for this unexpected incorrect answer, is incomplete. No count of right or wrong answers can, in itself, yield information about the reason for an improbable error. An improbable error (or success) implies the possibility of an interaction between person and item with respect to some secondary variable also active in the testing situation. When such confusion occurs, how can we detect it? What can we do about it? Here is how to proceed.

We want to find out if any items in a particular test are biased, say, against girls (or boys). Here are the steps to follow:

1. Examine the items carefully for sex-linked content and then classify them according to "theory" as a) those expected to favor boys, b) those expected to favor girls and c) those expected to be neutral. This is an important first step. If we really have no idea what we are looking for, we will surely have difficulty finding it. Worse, we will be seduced into mistaking accidental and transient irrelevancies for enduring effects.

2. A sample of girls and a sample of boys must take the test, if they have not already done so.

3. A separate calibration of the test in question is done for each sample - one for boys, another for girls. (Test calibration is explained and demonstrated step-by step in Wright and Stone, 1979, pages 28-62.)

4. The calibrated item difficulties from the separate analysis of each sample (a boy item calibration and a girl item calibration for each item) are centered and plotted against each other.

5. An identity line is drawn through the origin of this plot with slope one.

6. Statistical control lines are constructed around this identity line to guide interpretation and the plot is examined to see whether any items fall outside the control lines and hence are statistically identified as possibly biased. (See Wright and Stone, 1979, pp. 94-95 and Wright and Masters, 1982, pp. 115-117.)

We will illustrate these steps by examples designed to give the reader visual experience with the configurations that usually occur.

Figure 8.1 is a plot of two such item calibrations. The items expected to favor boys are indicated by triangles. The items expected to favor girls are indicated by circles. In Figure 1, the item plots center around the identity line. The items expected to be biased are not separated from each other. All items are within the 95% control lines. There is no indication of item bias in this plot which brings together the separate item calibrations for boys and girls. We must conclude that these data provide no reasons to suspect item bias with respect to sex.

Figure 8.2 is a different plot of item difficulties for boys and girls. In Figure 2 we can see two distinct item streams. One large item stream containing items favoring boys and also girls runs slightly above the (dotted) identity line. A second smaller stream of items favoring girls runs well below the dotted identity line.

To clarify what has occurred we draw a second (solid) identity line (also with slope one) through the middle of the larger stream of mixed items. Now we add control lines at two standard errors out around the solid identity line. This helps us to see the statistical separation of the two item streams. A difference is clearly indicated. There is an interaction between item content and sex which makes scores on the original mixture of items ambiguous. However, the majority of items in the larger stream might be used to provide unbiased measures on a "new" variable defined now by the particular items in the larger stream.

In Figure 8.3 we have another situation. Now we have three streams of items. One stream of items is above the identity line and favors boys as expected, another stream of items is below it and favors girls also as expected. Finally, a third stream of mixed items follows the identity line. Each stream of items is clearly distinct from the other. The question before us is: Which item stream defines the variable that we intend? The answer cannot come from the statistics. We must review the prior intention which motivated the composition of these items in order to make a sensible decision. We must decide which of the three streams of items contains the content which best, by our definition, defines the variable we intend. Once we have made this decision, the other items will become, by our definition, deviant from the frame of reference of this intention and hence "biased."

Figure 8.1 demonstrates what we will see when two samples produce no evidence of bias because all items plot along the expected identity line.

Figure 8.2 shows a larger stream of items slightly above the original identity line and a smaller stream of items below it. The simplest conclusion is that the smaller stream of items is biased with respect to the variable defined by the larger stream of items along the identity line.

In Figure 8.3, the situation is more complicated. We must decide which two of the three streams of items are deviant. We must decide which item stream marks out our intended variable. Does our intended variable remain with the original identity line or does it follow one of the offset streams of items? The example in Figure 3 causes us to realize that sometimes we will be forced to go beyond our statistics to outside criteria in order to establish a basis for judgment. Statistical analysis can show us what we have observed, but we must go beyond the data to make a criterion decision.

Our next example is from real data. It is a practical situation involving public school achievement test scores. Figure 8.4 is a plot of item calibrations made from two classrooms. One class is at Grade 2 and the other class is at Grade 3. Both classes took the same arithmetic computation test. The plot of item calibrations for the two samples, Grade 2 vs. Grade 3, shows two items clearly differing from the overall cluster of items.

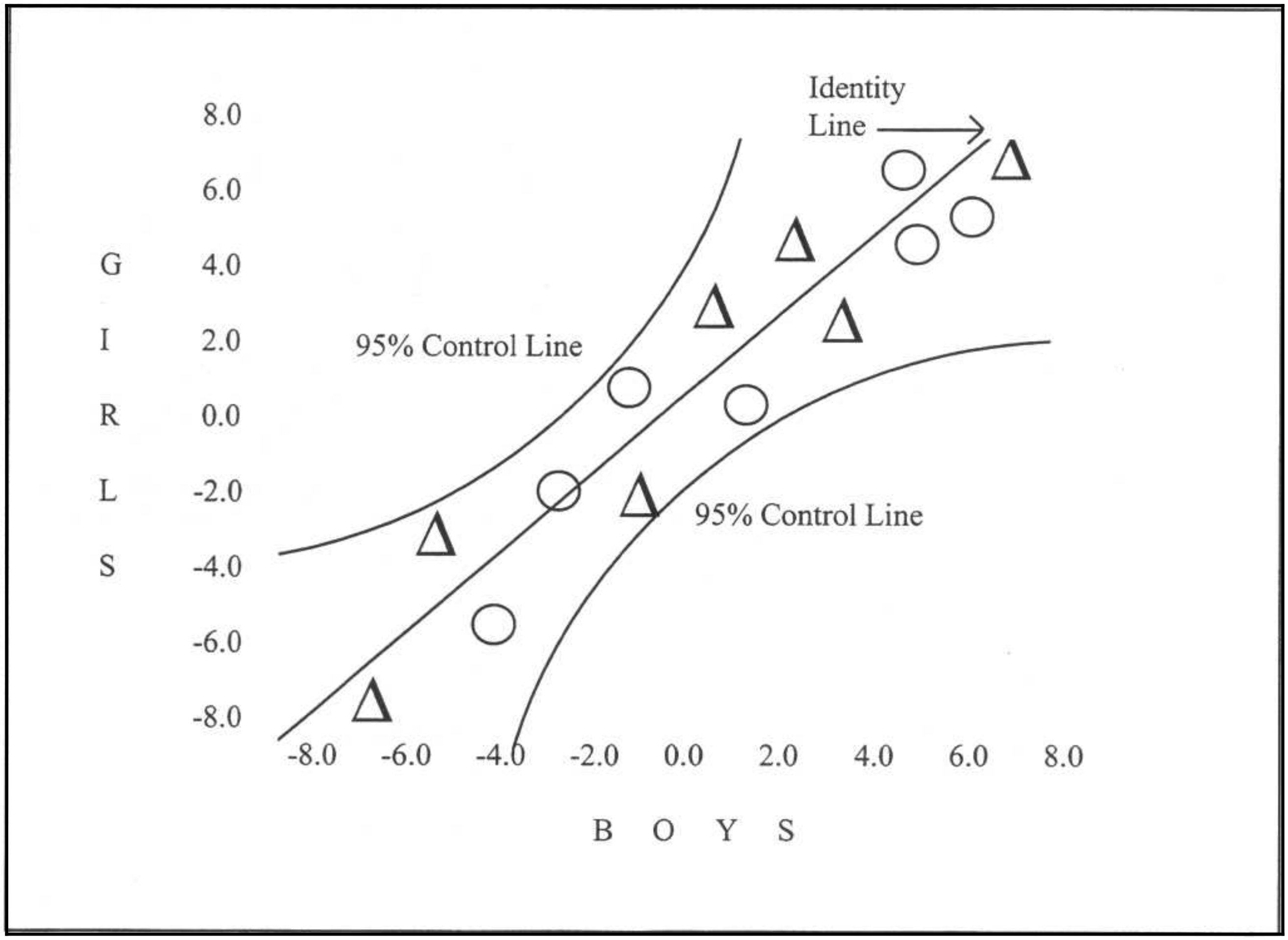
For these data we have some important external criteria, namely the content of the items. The computation skills required for most of these items are addition and subtraction of whole numbers without regrouping. The two deviant items, however, have common characteristics. They both involve subtraction with regrouping. This arithmetic operation marks a difference between the two grades. Within the frame of reference of whole number addition and subtraction without regrouping, there is an interaction between these two subtraction with regrouping items and grade level which makes these two items "biased."

These two items are biased against those second graders who have not yet learned how to do regrouping. The few second graders who were successful on these two items are ahead of their peers. For the others, these two items are almost impossible.

These examples make clear that a practical strategy is required to determine whether and how "bias" is evident. We have used Rasch measurement to illustrate how this can be done. It is especially important to be clear about our intentions prior to analysis in order to use the intended meaning of the items to help us understand the results of our analysis. Identification of bias is possible only when procedures like the one described have been applied. External criteria are needed to interpret results. But the criteria selected must be unequivocal in their application to the problem or they cannot be useful.

Figure 8.1

No evidence of item bias.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

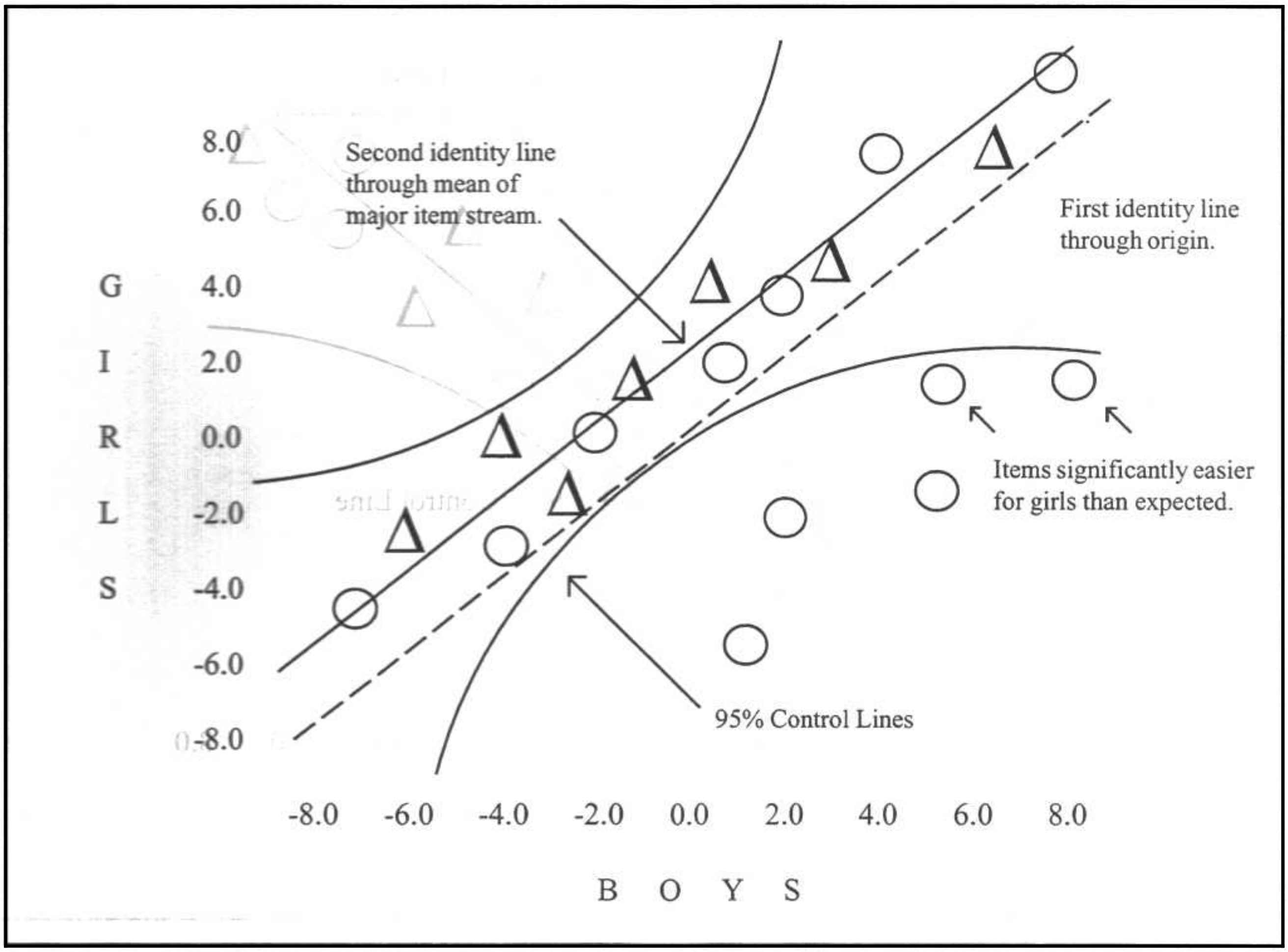
3. Circles = items preclassified "girl favoring."

4. Triangles = items preclassified "boy favoring."

5. For mathmetical specification of control lines see Wright and Stone, 1979, pp. 94-95 or Wright and Masters, 1982, pp. 115-117.

Figure 8.2

Five items biased in favor of girls.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

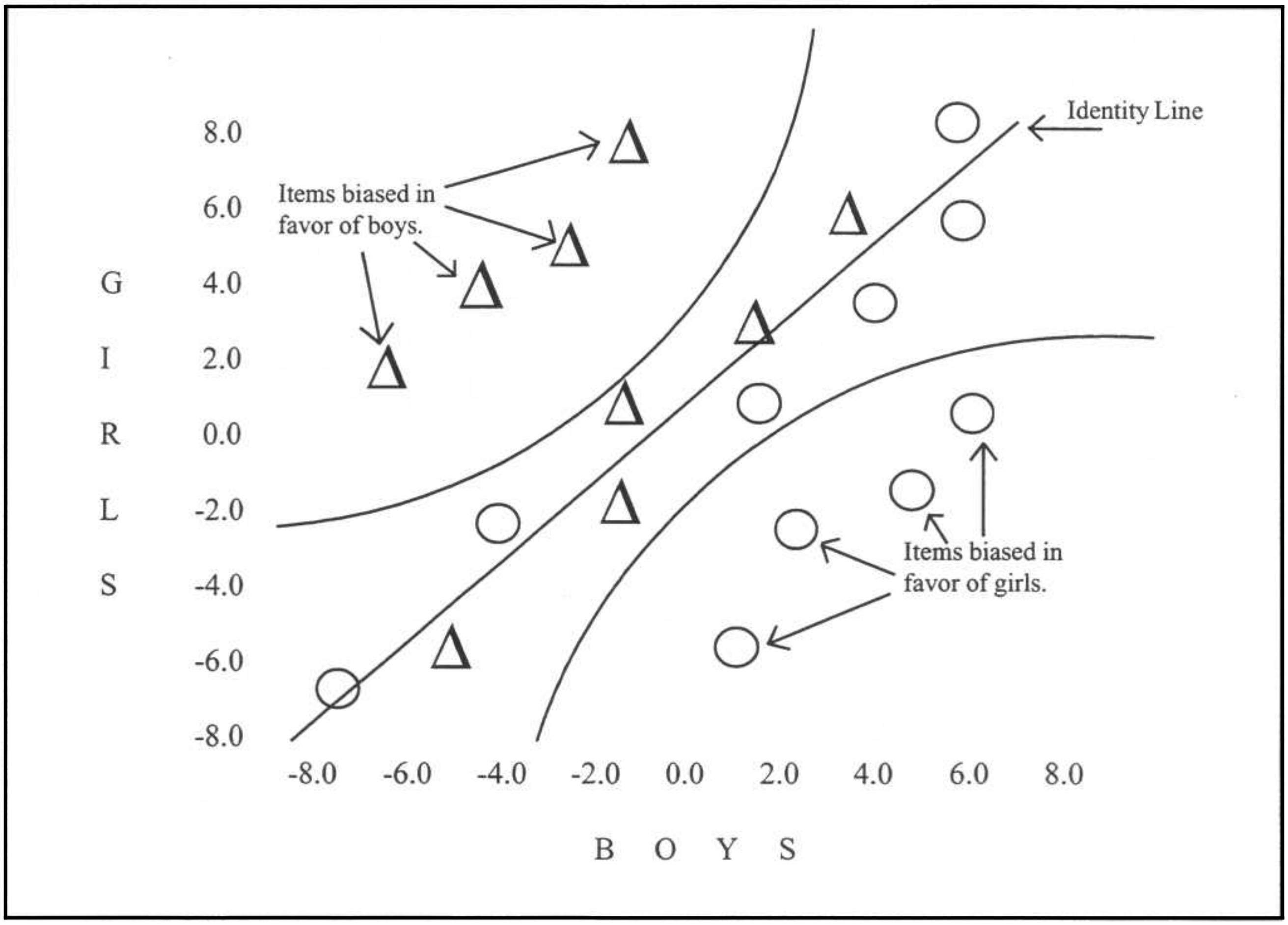
3. Circles = items preclassified "girl favoring."

4. Triangles = items preclassified "boy favoring."

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Maters, 1982, pp. 115-117.

Figure 8.3

Must decide which item stream defines the intended variable.



1. Vertical axis is item difficulty for girls.

2. Horizontal axis is item difficulty for boys.

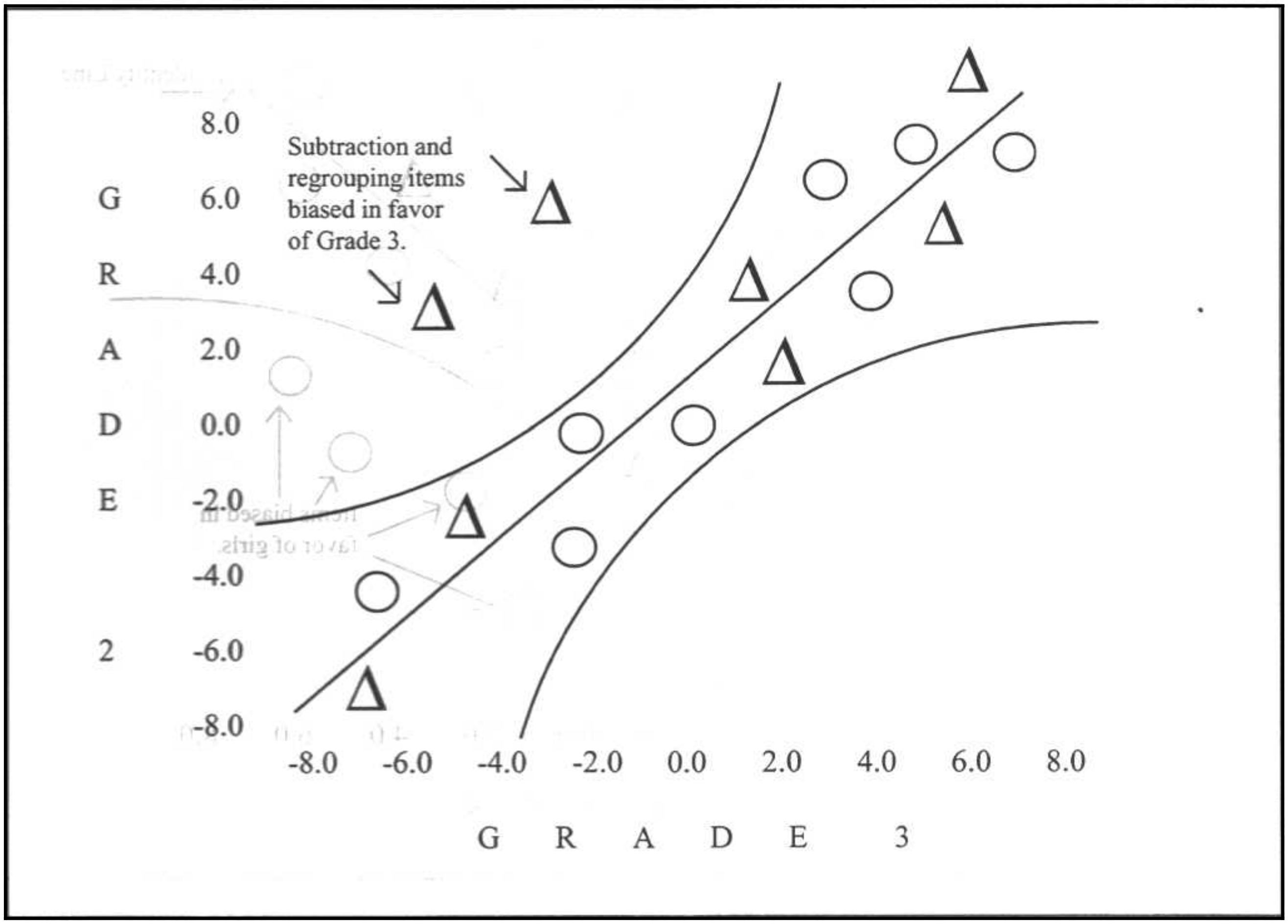
3. Circles = items preclassified “girl favoring.”

4. Triangles = items preclassified "boy favoring.”

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Masters, 1982, pp. 115-117.

Figure 8.4

Item calibratrions from grade 2 and grade 3 on an arithmetic achievement test.



1. Vertical axis is item difficulty for Grade 2.

2. Horizontal axis is item difficulty for Grade 3.

3. Circles = Grade 2 items.

4. Triangles = Grade 3 items.

5. For mathematical specification of control lines see Wright and Stone, 1979, pp. 94-95; Wright and Masters, 1982, pp. 115-117.

9. CONTROL LINES FOR ITEM PLOTS

When tests intended to measure on a particular variable are used with different groups of persons or to measure persons under different conditions, it is necessary to determine the degree of stability the tests maintain over these occasions. The quantitative comparisons sought depend on the tests retaining the same quantitative definition of the variable throughout the occasions to be compared. In order to determine this, a method is required to evaluate the invariance of the common test item calibrations from group to group or time to time.

In order to evaluate the invariance of these calibrations we need to compare item calibrations to see whether quantitative comparisons of the measures obtained from these occasions are possible. To do this we need to compare the centered calibrations for the items common to the two occasions.

In this chapter we explain how to make such comparisons (1) by plotting the centered item calibration estimates from two different occasions against one another, (2) by analyzing the standardized differences of the item calibrations between the two occasions and (3) by evaluating the correlation between the pairs of estimates over the set of common items.

In order to be explicit, we follow our explanations with an example to help the reader work through each step in the process. In the previous chapter, Identifying Item Bias, we showed how to evaluate item bias through the use of item plots. That chapter concentrated on explaining the concepts involved and using the figures to illustrate the concepts. In this chapter we explain the techniques by which such plots are constructed and evaluated.

PLAN OF ACTION

1. Estimate the item calibrations for each of the two occasions and identify the set of items common to both occasions. These alternative calibrations may come from two different samples of persons or from the same sample of persons tested at two different times. Estimate the item calibrations with their respective estimation standard errors and fit statistics. Thus for each calibration occasion and for each item *i* we calculate the item difficulty estimate d_i, its standard error s_i, and the fit of the calibrating data to these estimates, v_i.

2. Center each set of common item calibrations on the same origin (using perhaps the mean difficulty of the common items in the most recent or most important test) so that their comparison becomes independent of any translation effects between the centers of the two calibrated tests.

   (If there is a translation, then that amount would have to be accounted for before person measures from the two occasions could be compared. See Wright and Stone, 1979, pp. 96-98 and 112-117. The best way to proceed, however, is to carry out a third calibration of all of the data from both previous calibrations pooled into one combined data matrix. Usually this combined data matrix, in which every item on either test defines a column of possible responses and every individual test administration in either sample defines a row, has some empty cells where that item was not administered to that person. The "missing" data is easily managed in a calibration program like BIGSTEPS, (Wright, 1996)).

3. Plot these paired and centered item calibrations d_{1i} and d_{2i} against one another for each common item. A common variable is demonstrated when the plotted item points, which should estimate a single common difficulty for each item, fit an identity line, e.g. fall within one or two standard errors of their identity line.

4. Construct statistical control lines around the identity line by computing standard units of error along lines which are perpendicular to the identity line and passing through the item points. (The error control lines can be constructed for one or two error units producing 68% and 95% quality control.)

   These control lines can be used to evaluate, at a glance, the overall stability of the item calibrations shown on the plot. If more item calibrations fall outside the control lines than are expected by the control choices of 68% or 95%, we are led to doubt the stability of the calibrations in this study and to investigate the particular items causing the visible lack of invariance. Even when only a few items fall outside the control lines, we examine the particulars of these items carefully to determine why this has occurred and what we might do to control these particular conditions which threaten the validity of measurements made with these items.

5. Calculate the standardized difference between the alternate estimates of the single common item difficulty:

   $$z_{12i} = (d_{1i} - d_{2i}) / (s_{1i}^2 + s_{2i}^2)^{1/2} \ .$$

   This statistic has an expectation of zero and a variance of one when item stability holds. The pattern of these differences can be studied by plotting z_{12i} against $d_{.i} = (d_{1i} + d_{2i}) / 2$.

6. Correlate d_{1i} with d_{2i} over the $i = 1, L$ common items. This correlation r_{12} has a maximum value governed by the standard errors s_{1i} and s_{2i} and also the variance of the $d_{.i}$. This maximum correlation is:

   $$R_{\max} = 1 - (SE^2 / SD^2) = 1 - [(L-1)/L] * [\sum_{i}^{L} (s_{1i}^2 + s_{2i}^2) / \sum_{i}^{L} (d_{1i} + d_{2i})^2]$$

   $$\text{when } d_{1.} = d_{2.} = 0$$

   $$SE^2 = \sum_{i}^{L} (s_{1i}^2 + s_{2i}^2) / 4L$$

   $$SD^2 \sum^{L} (d_{1i} + d_{2i})^2 / 4(L-1)$$

Fisher's log transformation for linearizing correlations can be used to compare the observed correlation r_{12} with the maximum correlation $R_{\max}$ in order to test the hypothesis of item calibration stability.

$$t = \left[\frac{(L-3)^{1/2}}{2} \right] \log \left[\frac{(1+r_{12})(1-R_{\max})}{(1-r_{12})(1+R_{\max})} \right]$$

This statistic has expectation zero and variance one when item stability holds. It tests for the overall fit of these L items to the identity line which defines invariance.

AN EXAMPLE

These steps are illustrated in the following tables and figures. There is a first test form of 14 items calibrated on a sample of 34 persons. Then the variable was expanded by the development of 10 additional items making a second test form of 14 + 10 = 24 items which is given to a sample of 101 persons. The original 14 items remain common to both forms of the test. We evaluate the stability of the 14 items between these two test forms to determine whether the two item calibrations are statistically equivalent and so can be combined to define measures on a single common variable.

If this contention is supported by our analysis, then we can compare and pool the measures of the original 34 persons with the measures of the later 101 persons producing a sample of 135 persons measured on the same variable.

If, however, this contention is not supported, then we cannot compare or pool the original 34 measures with the subsequent 101 measures because we have found them to be measured on different variables. Then we are forced to review how these items are functioning in order to discover why the items are not working the way we intended.

1. Table 9.1 gives the item calibrations for each test form. The old and new item names for Forms 1 and 2 are given in Columns 1 and 5 with the old item calibrations for Form 1 listed in Column 2 and the new item calibrations for Form 2 listed in Column 6. The new item names for Form 2, given in Column 5, are shown with their old Form 1 item names in parentheses. These new item calibrations for the 14 original items are given again in Column 7.

   Observe that the center (mean) of the 14 Form 1 old item calibrations is at 0.0 (Column 2) and the center (mean) of the 24 Form 2 new item calibrations is also at 0.0 (Column 6). These zeros, however, are not equivalent, since the old zero defines the center of the old 14 items while the new zero defines the center of the new 24 items. In fact, the center (mean) of the new Form 2 calibrations for the 14 original items is now 0.4 on the new scale of Form 2 (Column 7). Because of this difference the calibrations of the original 14 items must be shifted by 0.4 (Column 3). This shift puts them on the same scale as the new 24 items and produces the adjusted values given in Column 4 which are the values that will be used to compare item stability between Forms 9.1 and 9.2.

2. The adjusted Form 1 (Column 4) and Form 2 (Column 7) calibrations of these 14 items are plotted in Figure 9.1. The plot shows that these items fall along the identity line rather well,

Table 9.1

Comparing the Calibrations of 14
Items Common to Two Test Forms

FIRST TEST FORM				SECOND TEST FORM		
(1) Old Item Name	(2) Old Item Calibration	(3) Shift Value	(4)* Adjusted Calibration	(5) New Item Name	(6) New Item Calibration	(7)** New Calibration (Original 14 Items)
				1	-6.0	
				2	-5.6	
1	-4.2	0.4	-3.8	3 (1)	-3.8	-3.8
2	-3.6	0.4	-3.2	4 (2)	-2.3	-2.3
3	-3.2	0.4	-2.8	5 (3)	-2.5	-2.5
				6	-4.0	
4	-3.6	0.4	-3.2	7 (4)	-2.3	-2.3
5	-2.2	0.4	-1.8	8 (5)	-1.8	-1.8
6	-3.2	0.4	-2.8	9 (6)	-1.8	-1.8
7	-1.5	0.4	-1.1	10 (7)	-0.8	-0.8
				11	0.1	
				12	-0.6	
				13	-0.3	
				14	-1.3	
				15	-0.5	
8	0.8	0.4	1.2	16 (8)	2.2	2.2
9	2.1	0.4	2.5	17 (9)	1.6	1.6
10	1.9	0.4	2.3	18 (10)	2.2	2.2
11	3.2	0.4	3.6	19 (11)	3.1	3.1
12	4.6	0.4	5.0	20 (12)	3.6	3.6
13	4.6	0.4	5.0	21 (13)	3.6	3.6
14	4.6	0.4	5.0	22 (14)	4.7	4.7
				23	6.5	
				24	6.0	
Column Mean	0.0		0.4		0.0	0.4
SD	3.4		3.4		3.4	2.8

* (4) = (2) + (3)

The comparison will be made between (4) and (7).

** (6) = (7)

but, as yet, we have no way to evaluate how much these item plots could deviate from the exact identity line before we would be forced to decide that the differences are too much. To accomplish this evaluation, we construct quality control lines. These lines guide our study of the plot to help us to make useful decisions.

3. Figure 9.2 lays out a simple way to construct these control lines. The standard unit of difference error parallel to either axis for item i is:

   $$S_{12i} = (s_{1i}^2 + s_{2i}^2)^{1/2}$$

   The notes appending Figure 9.2 give the details for determining the coordinates (X and Y) for a machine plot of the control lines. See Table 9.3 for application to our data. Entering these values in a plotting program can produce smoothed quality control lines.

   Table 9.2 shows how to do a simple hand plot of the control lines. This is used with our sample data and shown in Figure 9.3.

   A unit of error equivalent to S_{12i} but perpendicular to the 45 degree identity line is:

   $$T_{12i} = \left[\left(s_{1i}^2 + s_{2i}^2\right)/2\right]^{1/2} = S_{12i}/\sqrt{2}$$

   One of these T error units perpendicular to the identity line, through the $(d_{1i},\ d_{2i})$ item plot and extended in each direction from the identity line yields a pair of 68% control lines. Two of these T error units perpendicular to the identity line yields a pair of 95% control lines.

   Table 9.2 gives the standard errors s_{1i} and s_{2i} (Columns 6 and 7) for the 14 common items connecting Forms 1 and 2.

   We calculate T_{12i} for each of the 14 items and plot these locations in Figure 9.3 at two standard error units above and below the identity line. These points can be connected and smoothed to provide the quality control lines needed to evaluate the item plots.

4. Figure 9.3 shows that the plots of the 14 items of Forms 1 and 2 are all well within two standard errors of the identity line. It also shows that the hand and constructed methods of drawing in control lines lead to identical results. We conclude that these 14 items fall along the identity line, given their standard errors. Our variable extension is successful according to this sample data.

5. We can also evaluate the standardized item calibration differences between the Form 1 and Form 2 item calibration estimates for these 14 items by using:

   $$Z_{21i} = (d_{2i} - d_{1i}) / (s_{1i}^2 + s_{2i}^2)^{1/2}.$$

   These standardized differences are expected to have a mean of zero and a variance of one. The standardized differences of the 14 items are given in Column 9 of Table 9.2. Trends can be evaluated by plotting these Z_{21i} against $d.i$ for each item.

   Figure 9.4 is this plot. We observe that all of the remaining items are well within 1.0. All

Figure 9.1

Plot of common item calibrations: Form 1 versus Form 2.

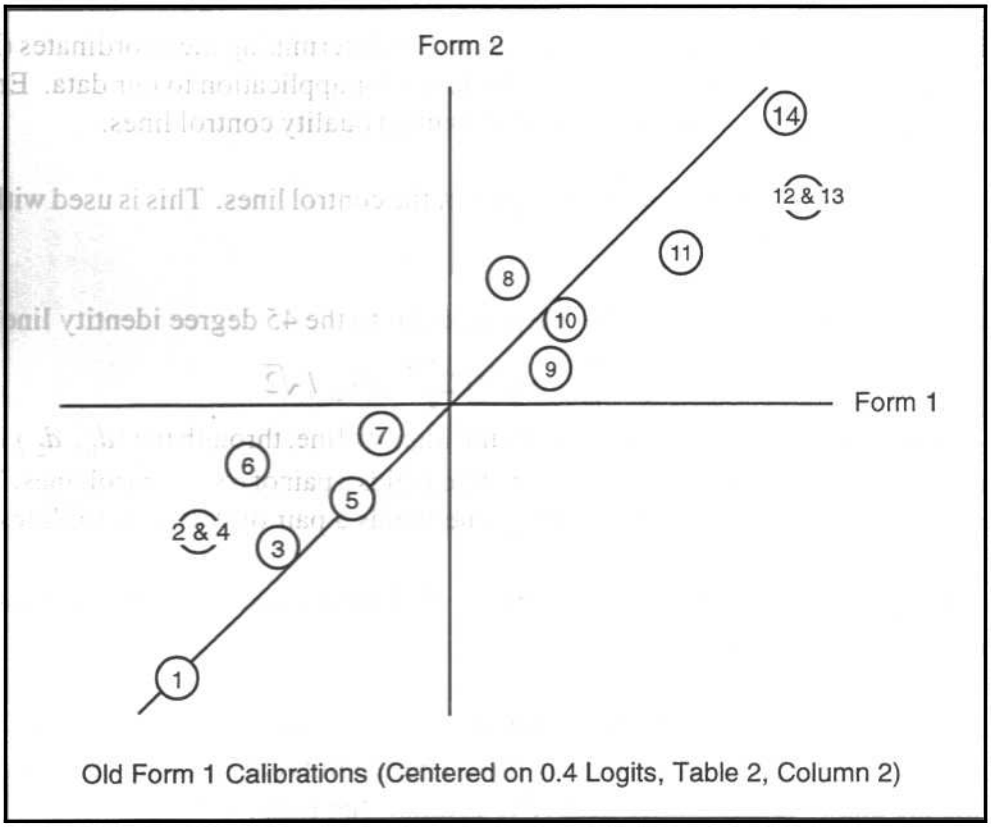
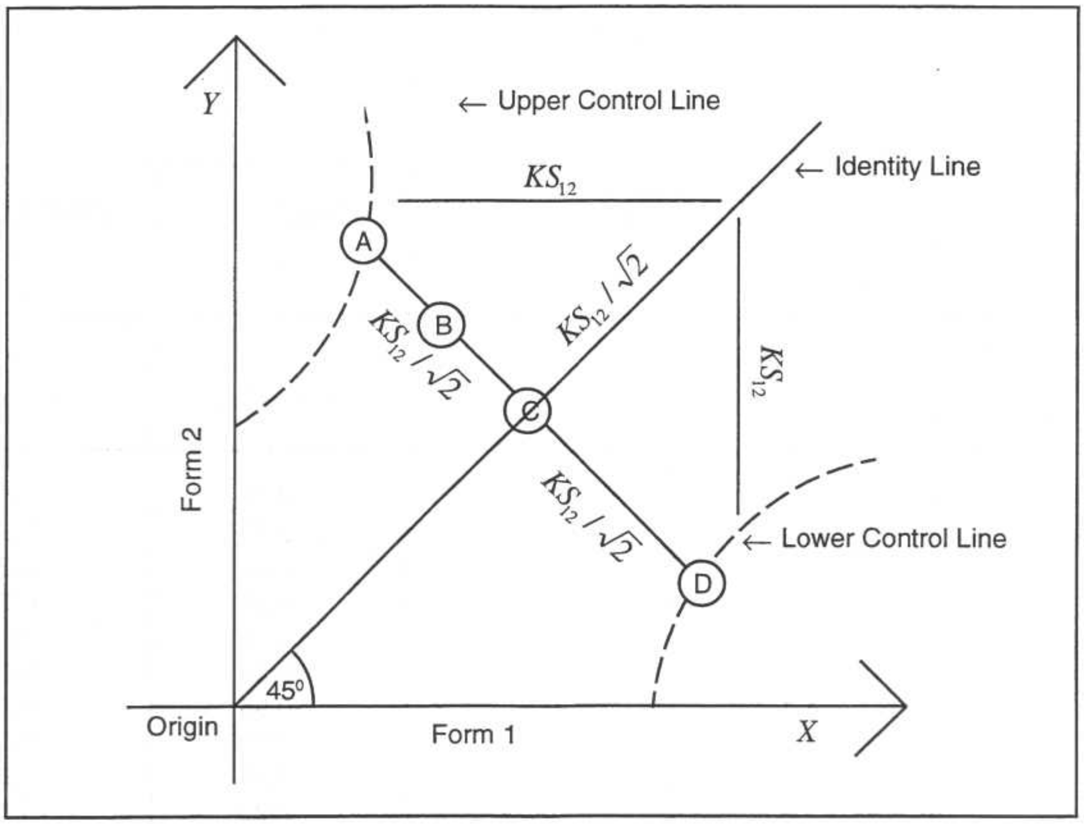


Figure 9.2

How to construct control lines.



Upper Control Line:

Position A: $X = d - KS_{12}/2 = (d_1 + d_2 - KS)/2;\ Y = d + KS_{12}/2 = (d_1 + d_2 + KS_{12})/2$

Item Plot:

Position B: $X = d_1;\ Y = d_2$

Identity Line:

Position C: $X = (d_1 + d_2)/2 = d;\ Y = (d_1 + d_2)/2 = d$

Lower Control Line;

Position D: $X = d + KS_{12}/2 = (d_1 + d_2 + KS_{12})/2;\ Y = d - KS_{12}/2 = (d_1 + d_2 - KS_{12})/2$

See Table 9.3 and Figure 9.3 for an example.

$K =$ number of standard error units chosen to set the confidence level control of the lines; e.g., $K = 1$ produces 68% confidence and $K = 2$ produces 95% confidence.

$S_{12} = \sqrt{S_1^2 + S_2^2}$ the standard error of the difference $(d_1 - d_2)$

S_l = the standard error of d_1
S_2 = the standard error of d_2
d = $(d_1 + d_2)/2$

Table 9.2.

Item Calibrations, Standard Errors
and Standardized Differences *Z*

	CALIBRATION		AVERAGE $d_{.i}$	DIFFERENCE
(1) Old Item Name	(2)* d_{1i}	(3)** d_{2i}	(4) $(d_{1i} + d_{2i})/2$	(5) $(d_{1i} - d_{2i})$
1	-3.8	-3.8	-3.80	0.0
2	-3.2	-2.3	-2.75	-0.9
3	-2.8	-2.5	-2.65	-0.3
4	-3.2	-2.3	-2.75	-0.9
5	-1.8	-1.8	-1.80	0.0
6	-2.8	-1.8	-2.30	-1.0
7	-1.1	-0.8	-0.95	-0.3
8	1.2	2.2	1.70	-1.0
9	2.5	1.6	2.05	0.9
10	2.3	2.2	2.25	0.1
11	3.6	3.1	3.35	0.5
12	5.0	3.6	4.30	1.4
13	5.0	3.6	4.30	1.4
14	5.0	4.7	4.85	0.3
MEAN*** S.D.	0.4 3.4	0.4 2.8		

* Column 4 from Table 1

** Column 7 from Table 1

*** Items have been centered at the common mean for Form 2 of 0.4. This separates the analysis of the calibration differences $(d_{1i} - d_{2i})$ from any overall difference in test form difficulty.

Table 9.2. (Continued).

	STANDARD ERROR		STANDARD ERROR OF DIFFERENCE	STANDARDIZED DIFFERENCE	ERROR UNIT
Old Item Name	(6) S_{1i}	(7) S_{2i}	(8) S_{12i}	(9) Z_{21i}	(10) $T_{12i} = S_{12i}/\sqrt{2}$
1	0.8	1.0	1.28	0.00	0.91
2	0.7	0.7	0.99	0.91	0.70
3	0.7	0.7	0.99	0.30	0.70
4	0.7	0.7	0.99	0.91	0.70
5	0.5	0.6	0.78	0.00	0.74
6	0.7	0.6	0.92	0.88	0.81
7	0.5	0.6	0.78	0.29	0.74
8	0.4	0.8	0.89	0.92	0.77
9	0.5	0.6	0.78	-0.86	0.74
10	0.5	0.8	0.94	-0.09	0.81
11	0.7	0.8	1.06	-0.41	0.86
12	1.1	1.0	1.49	-0.97	1.03
13	1.1	1.0	1.49	-0.97	1.03
14	1.1	1.2	1.63	-0.20	1.05
			MEAN	+0.02	
			S.D.	0.71	

$S_{12i} = (S_{1i}^2 + S_{2i}^2)^{1/2}$

$Z_{21i} = (d_{2i} - d_{1i})/S_{12i}$

Column (9) = (5)/(8)

The Quick Hand Method for Adding Control Lines

To draw 95% control lines by hand use the approximation (Wright and Stone, 1979, p. 95) for an error allowance perpendicular to the identity line:

$$2T_{12i} = [(S_{1i}^2 + S_{2i}^2)/2]^{1/2} \approx (S_{1i} + S_{2i}).$$

Mark off a piece of graph paper to match the plotting axes and then slide this special ruler along the identity line marking off the perpendicular distances $(S_{1i} + S_{2i})$ in each direction away from the identity line as each item point (d_{1i}, d_{2i}) is encountered. This is done in Figure 9.3 where the results are marked as small circles.

of the values are within the 68% control lines.

The correlation over $i = 1, 14$ of the calibrations d_{1i} and d_{2i} can also be determined. A limit for this coefficient is $R_{\max}$. In our example $R_{\max} = 0.98$ and the correlation for the observed item calibrations is also 0.98. Since $R_{\max} = 0.98$ is the same as $r_{12} = 0.98$, we see that the correspondence between the item calibration estimates computed from the Form 1 and Form 2 samples is as good as can be expected. This correlation, when evaluated for its statistical deviation from the intended equating of Form 1 and Form 2 using Fisher's log transformation, produces a $T \approx 0.0$. We retain the hypothesis of no statistical difference between these 14 pairs of item calibrations and hence of the stability of these items and the variable they define over the two occurrences. As a result we can pool and compare the 34 and 101 person measures.

Our example has illustrated the steps for evaluating the stability of item calibrations. In our example we confirmed the invariance of our item calibrations. If confirmation were not achieved, we could not undertake any quantitative comparisons of the measures from the two occasions and it would be necessary to determine why particular items failed to support our intention to equate Form 1 and Form 2 and to compare the measures they produced. Changes might be made in these items or new items constructed and the equating process repeated with a new sample. Even when changes do not appear necessary it is prudent to monitor item calibration stability continually as new samples occur in order to verify that conditions have not changed.

Figure 9.3

Plot of item calibrations: Form 1 versus Form 2 with 95% control lines using hand method and constructed method of Table 9.3.

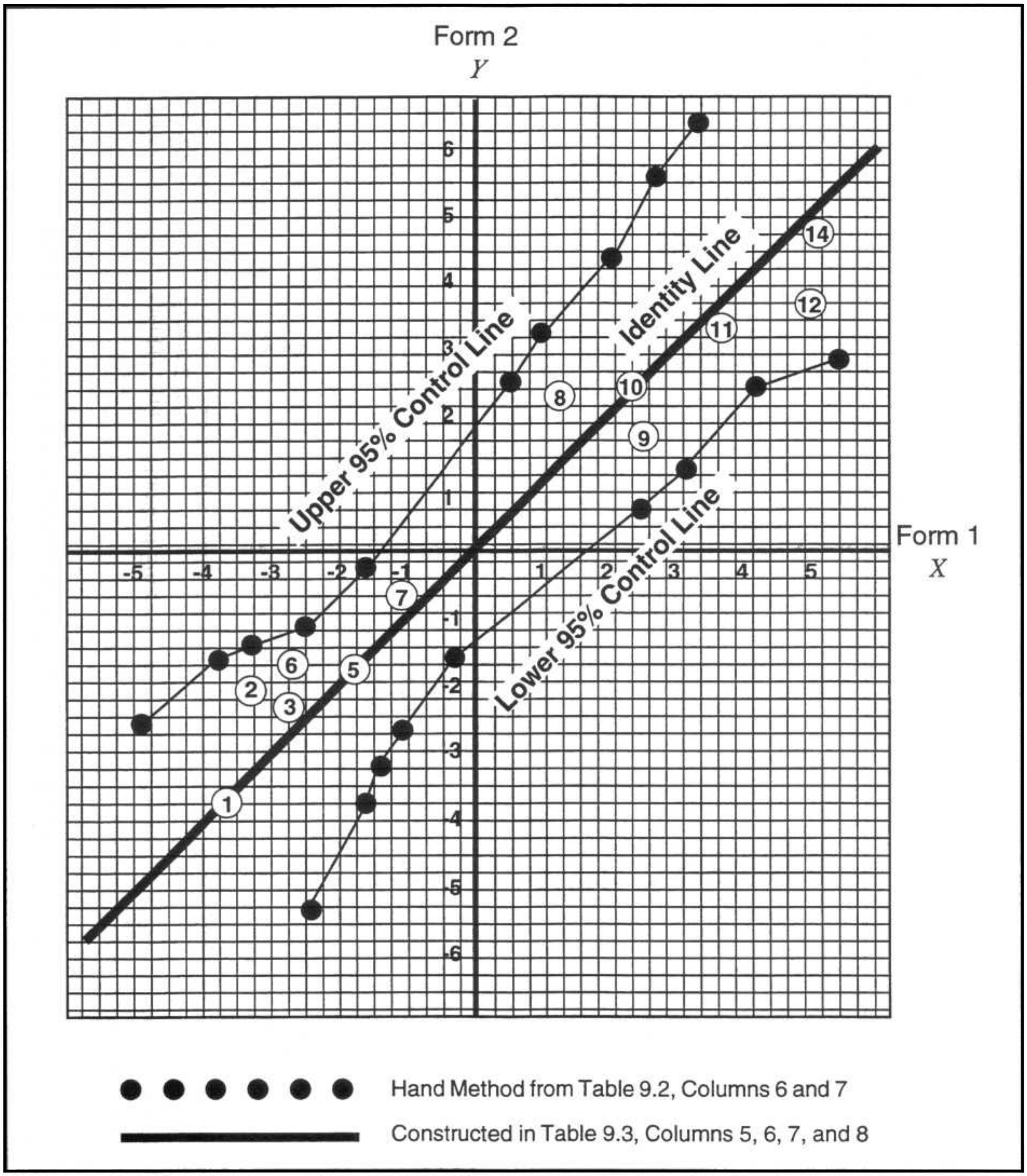


Table 9.3

Example Data for Constructing 95% (K=2) Control Lines

Old Item Name	Item Plot *Figure 2 (B)*		Identity Line *Figure 2 (C)*	Standard Error	95% Upper Control Line *Figure 2 (A)*		95% Lower Control Line *Figure 2 (D)*	
	d_{1i} *(X)* (1)	d_{2i} *(Y)* (2)	d_i *(X,Y)* (3)	S_{12i} (4)	$d-S_{12i}$ *(X)* (5)	$d+S_{12i}$ *(Y)* (6)	$d+S_{12i}$ *(X)* (7)	$d-S_{12i}$ *(Y)* (8)
1	-3.8	-3.8	-3.80	1.28	-5.08	-2.52	-2.52	-5.08
2	-3.2	-2.3	-2.75	0.99	-3.74	-1.76	-1.76	-3.74
3	-2.8	-2.5	-2.65	0.99	-3.64	-1.66	-1.66	-3.64
4	-3.2	-2.3	-2.75	0.99	-3.74	-1.76	-1.76	-3.74
5	-1.8	-1.8	-1.80	0.78	-2.58	-1.02	-1.02	-2.58
6	-2.8	-1.8	-2.30	0.92	-3.22	-1.38	-1.38	-3.22
7	-1.1	-0.8	-0.95	0.78	-1.73	-0.17	-0.17	-1.73
8	1.2	2.2	1.70	0.89	0.81	2.59	2.59	0.81
9	2.5	1.6	2.05	0.78	1.27	2.83	2.83	1.27
10	2.3	2.2	2.25	0.94	1.31	3.19	3.19	1.31
11	3.6	3.1	3.35	1.06	2.29	4.41	4.41	2.29
12	5.0	3.6	4.30	1.49	2.81	5.79	5.79	2.81
13	5.0	3.6	4.30	1.49	2.81	5.79	5.79	2.81
14	5.0	4.7	4.85	1.63	3.22	6.48	6.48	3.22

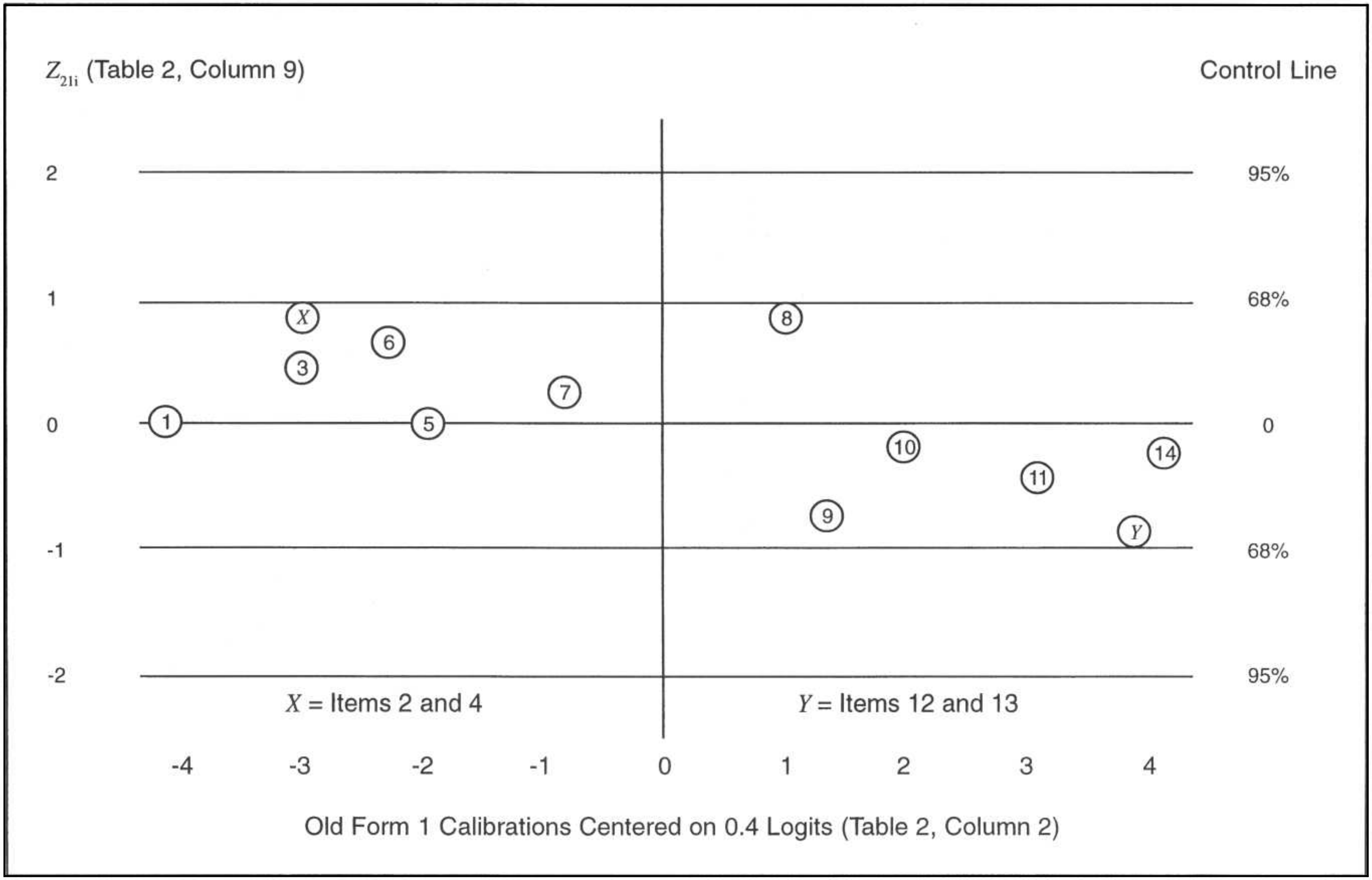
Columns 1 and 2 are from Table 2, Columns 2 and 3.

$d_I = (d_{1i} + d_{2i})/2$ (Table 2, Column 4)

$S_{12i} = (S_{1i}^2 + S_{2i}^2)^{1/2}$ (Table 2, Column 8)

Figure 9.4

Plot of item calibrations vs. standardized difference with 68% and 95% control lines.



10. INFORMATION FUNCTION AND MISFIT SENSITIVITY

In this chapter we discuss the "information" of a test and its items. "Information" is directly related to misfit sensitivity so we discuss both "fit" and "information."

The way "information" enters into determining the value of an observation is through its bearing on the precision of measurement. Measurement precision depends on the number of items in the response record and on the relevance of each item to the particular person. On-target items make for an efficient test, off-target items do not.

Since measurement precision depends on the number of items in the response record and on the relevance of each item to the particular person, the evaluation of each item's contribution to knowledge of the person can be calculated specifically. Information is the inverse square of the standard error of measurement. The information (*I*) in a test score or in a measure derived from a score is $I = 1/SE_m^2$ which is one over the square of the standard error of that score or measure. The smaller the standard error, the larger the information (*I*). When SE_m is in logits, information is in inverse square logits. Replications enter information through the numerator. For the standard error replications enter through the denominator.

For dichotomies where [*p*(1-*p*)] (proportion correct times proportion incorrect) is equal to Information (*I*), then the square root of 1 over [*p*(1-*p*)] is the standard error: $SE_m = 1\sqrt{1/I}$ or $= I^{-1/2}$.

Tests can be compared for their information in order to see which test provides the most information. The consequences of lengthening or shortening a test can be anticipated by observing the resultant gain or loss of information that accrues.

When item and person are close to one another, i.e. on target, then the item contributes more to the measure of the person than when the item and person are far apart. The greater the "distance" (the difference between the person's ability and the item's difficulty), the greater the number of items needed to obtain a measure of comparable precision.

Table 10.1 helps make such determinations. Column 1 is the absolute logit difference $|B-D|$ between person ability and item difficulty (*B-D*). Column 2 is the squared standardized residual $z^2 = \exp(|B-D|)$. *Exp (B-D)* is the unexpectedness of an incorrect response to a relatively easy item while *exp (D-B)* is the unexpectedness of a correct response to a relatively hard item. Each z^2 marks the "unexpectedness" of a response.

The values for every instance of unexpectedness can be ascertained and accumulated over items to evaluate the response pattern plausibility of any person measure, or summed over persons to evaluate the sample pattern plausibility of any item calibration. The mark of unexpectedness is a positive difference from (*B-D*) or from (*D-B*). Corresponding values for z^2 can be looked up in Table 10.1, which gives values of $z_o^2 = exp\ (B\text{-}D)$ for unexpected *incorrect* answers or values of $z_1^2 = \exp(D-B)$ for unexpected *correct* answers.

Table 10.1

Information in Terms of Relative Efficiency and Misfit Detection

Difference Between Person Ability and Item Difficulty $\|B-D\|$	Squared Standardized Residual $z^2 = \exp(\|B-D\|)$	Misfit Detection as Response Improbability $p = 1/(1+z^2)$	Relative Efficiency of the Observation $I = 400p(1-p)$	Number of Items Needed to Maintain Equal Precision $L = 1000/I$
-0.6, 0.3	1	.50	100	10
0.4, 0.8	2	.33	90	11
0.9, 1.2	3	.25	75	13
1.3, 1.4	4	.20	65	15
1.5, 1.4	5	.17	55	18
1.7, 1.8	6	.14	50	20
1.9, 2.0	7	.12	45	22
2.1	8	.11	40	25
2.2	9	.10	36	28
2.3	10	.09	33	30
2.4	11	.08	31	32
2.5	12	.08	28	36
2.6	14	.07	25	40
2.7	15	.06	23	43
2.8	17	.06	21	48
2.9	18	.05	20	50
3.0	20	.05	18	55
3.1	22	.04	16	61
3.2	25	.04	15	66
3.3	27	.04	14	73
3.4	30	.03	12	83
3.5	33	.03	11	91
3.6	37	.03	10	100
3.7	41	.02	9	106
3.8	45	.02	9	117
3.9	50	.02	8	129
4.0	55	.02	7	142
4.1	60	.02	6	156
4.2	67	.02	6	172
4.3	74	.01	5	189
4.4	81	.01	5	209
4.5	90	.01	4	230
4.6	99	.01	4	254

Note that values of increasing unexpectedness $(z^2) = \exp(|B-D|)$ correspond to an increasing difference between person ability and item difficulty.

Column 3 is $p = 1/(1+z^2)$, the improbability of an observed response. This p provides a significance test for the null hypothesis of fit for any particular response.

When accumulated for the kind of data most often encountered, each z^2 is distributed approximately χ^2 with 1 degree of freedom each. When z^2s are accumulated over items for a person or over persons for an item, the resulting sums are approximately χ^2 with $(L-1)$ degrees of freedom for a person responding to L items and $(N - 1)$ degrees of freedom for an item responded to by N persons.

As p decreases, the difference between person ability and item difficulty increases. Examples of misfit analysis are given in Chapter 4 of *Best Test Design* (Wright & Stone, 1979).

Column 4 is an information index $I = 400p(1-p)$ which indicates the relative efficiency which an observation at any $|B-D|$ provides about that person and item interaction. The index is scaled by 400 to give the amount of information provided by the observation as a percentage of the maximum information that one observation at $|B-D| = 0$, i.e., right on target, would provide. This index can be used to judge the value of any particular item or items used in measuring a person.

It requires five 20% items at $|B-D| = 2.9$ to provide as much information about a person as would be provided by one 100% item at $|B-D| = 0$. When $|B-D|$ reaches 3.0 logits, it takes four to five times as many items to provide as much information as could be had from items that fall $|B-D| < 1$ region within one logit of the person. As $|B-D| > 2.8$, the probability of an unexpected response such as $X = 0$ when $(B\text{-}D) > 2.8$ or $X = 1$ when $(B\text{-}D) < -2.8$ drops to $P = .05$. This produces the possibility of a statistically significant "misfit," a probable invalidity in that response to that item.

The test length necessary to maintain a specified level of measurement precision is inversely proportional to the relative efficiency of the items used. Column 5 gives the number (L) of less efficient items necessary to match the precision of 10 "right-on-target" items. As items go increasingly "off target" the number of items required to maintain equal precision increases. The increase from on-target items at a minimal difference between person ability and item difficulty to off-target items at a two logit difference between person ability and item difficulty is two-fold. Twice the number of items are required. An increase to a three logit difference requires more than five times the number of on-target items and an increase to a four logit difference requires more than 14 times the number of on-target items! Off-target items are extremely inefficient, they require an inordinate number of additional items to maintain equal precision.

Figure 10.1 summarizes and facilitates the use of the data in Table 10.1.

The values in Table 10.1 can also be pictured as the logistic curves shown in Figure 10.2. The

Figure 10.1

Summary and interpretation of data provided in Table 10.1.

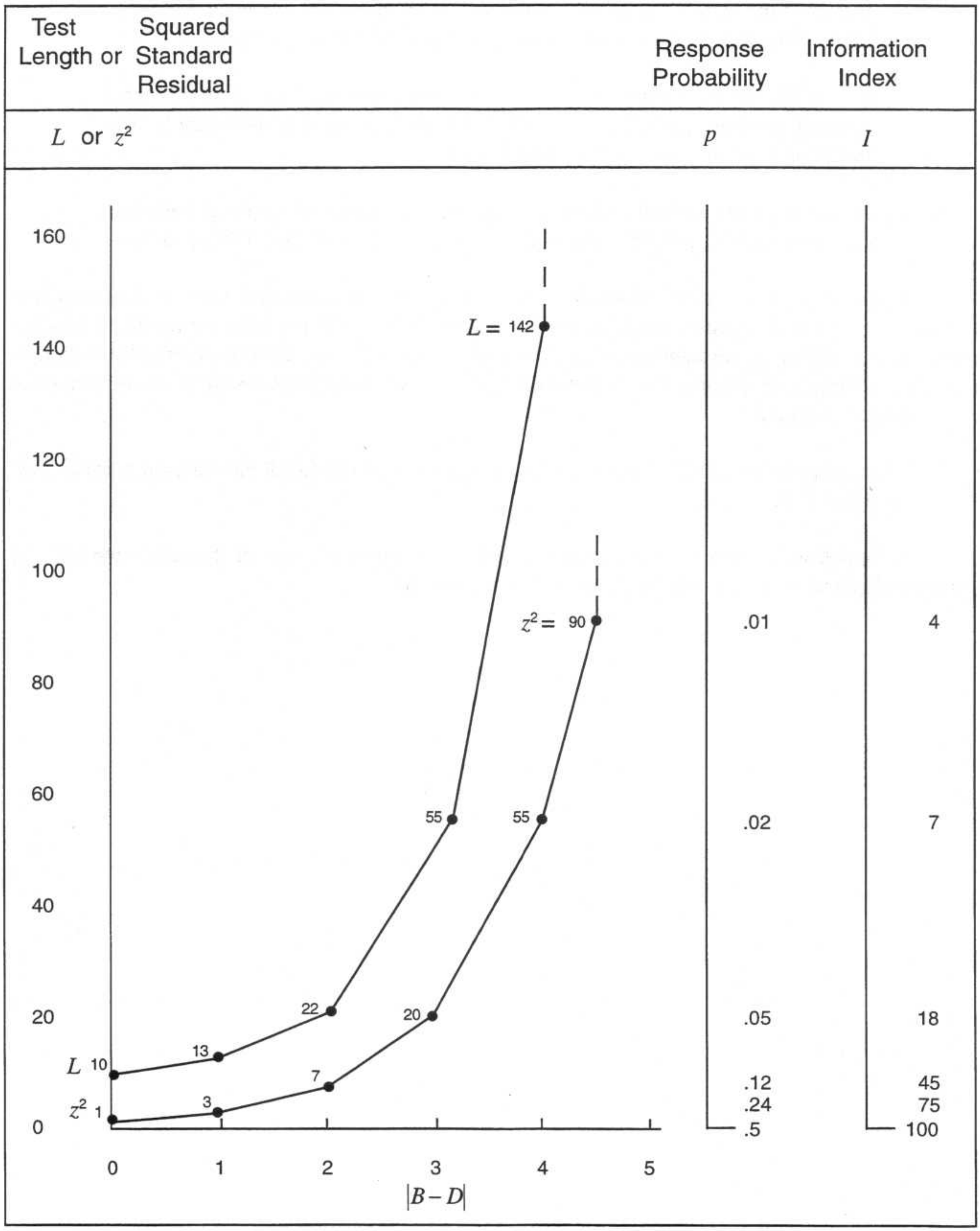
Relative Location of Item	(Ability-Difficulty) Difference	Item Efficiency and Sensitivity of Misfit Detection
Right on Target	$\lvert B-D\rvert<1$	- Excellent efficiency, 75% or better. No misfit analysis possible.
Close Enough	$1<\lvert B-D\rvert<2$	- Good efficiency, 45% or better. No misfit analysis possible.
Slightly Off	$2<\lvert B-D\rvert<3$	- Poor efficiency, less than 45%. Misfit detectable when unexpected responses accumulate.
Rather Off	$3<\lvert B-D\rvert<4$	- Very poor efficiency, less than 18%. - Even single unexpected responses can signal significant response irregularity.
Extremely Off	$4<\lvert B-D\rvert$	- Virtually no efficiency, less than 7%. - Unexpected responses always require diagnosis.

horizontal axis gives $\lvert B-D\rvert$. The vertical axis gives the values from Table 1 of z^2, L, p, and I. Because L is scaled to an intercept of +10, the L curve is located slightly to the left of the z^2 curve. Both curves show the same function. The slope, i.e., loss of information, for $\lvert B-D\rvert=0$ to 2 logits is modest. This slope increases greatly after 2 logits, and even more so after 3 logits. The progressive increase in slope after $\lvert B-D\rvert=2$ logits shows clearly how information changes as a function of $\lvert B-D\rvert$.

"Information," as a concept and statistic, was formulated by Sir Ronald Fisher, (1921, pp. 316-317) in conjunction with his formulations of efficiency and sufficiency.

> In the logical situation presented by problems of statistical estimation, I have shown that a mathematical quantity can be identified which measures the quantity of information provided by the observational data, relevant to the value of any

Figure 10.2
Functions of z^2, L, p, and I.



> particular unknown parameter. That it is appropriate to speak of this quantity as the quantity of information is shown by the three following properties:
>
> (i) The quantity of information in the aggregate of two independent sets of observations is the sum of the quantities of information in the two sets severally; each observation thus adds a certain amount to the total information accumulated.
>
> (ii) When, on increasing our observations, the sampling error of an efficient estimate tends to normality, the quantity of information is proportional to the precision constant of the limiting distribution.
>
> (iii) The quantity of information supplied by any statistic or group of statistics can never exceed the total contained in the original data (Fisher, 1934, p. 6-7).

Fisher's quest was for the best statistic. He reasoned that such a statistic must be *consistent*, tend to estimate its parameter more closely as sample size increases; *efficient*, have variance less than any other statistic estimating the same parameter and *sufficient*, incorporate all of the information available in the sample regarding its parameter. When a sufficient statistic exists, it can be obtained by the method of maximum likelihood.

> "A statistic which fulfills the criterion of sufficiency will also fulfill the criterion of efficiency" (Fisher, 1921, p. 317).

Information is simple in Rasch measurement. It is just a function of the difference between person ability and item difficulty (*B-D*) as shown in Table 10.1.

11. CONNECTING TESTS

In this chapter we describe the basic strategies for connecting tests intended to measure on the same variable so that the separate measures each test implies are expressed together on one single common scale. The process begins by understanding how to link two tests. Next we consider how to connect several tests and from there we proceed to plans for connecting all possible tests.

CONNECTING TESTS

The traditional method for connecting two tests is by equating the equal-percentile scores of a sample of persons who take both tests simultaneously. This process requires a large sample of persons with scores broadly enough distributed to assure an adequate representation of each score-to-percentile connection.

Rasch measurement enables a more economical and better controlled method for connecting tests and building item banks. Links of 10 to 20 common items are embedded in pairs of tests composed of otherwise different items. Each test is administered to its own sample of persons. No person need take more than one test. But all items in all tests can be subsequently connected through the network of common item links.

The traditional approach to equating two 60-item tests, say Test A and Test B, is to give both tests simultaneously to a sample of many, say 1200, persons as in Figure 11.1. The large sample is to assure the detailed representation of score percentiles necessary for successful percentile equating. Each person takes Test A and Test B, a total of 120 items.

In contrast, the Rasch approach can do the same job with each person taking only one test of 60 items. To accomplish this a third 60-item test, C, is made up of 30 items from each of the original tests A and B. Then each of these three tests is given to a sample of 400 persons as depicted in the lower half of Figure 11.1. Now each person takes only one test, but all 120 items are calibrated together through the two 30-item links connecting the three tests. The testing burden on each person is one half of that required by the equal percentile plan. But the equating of the tests is under far better control. In actual practice, the three samples can also be halved to 200 each without loss of control. This reduces the amount of data to one fourth of that required for the equal percentile equating.

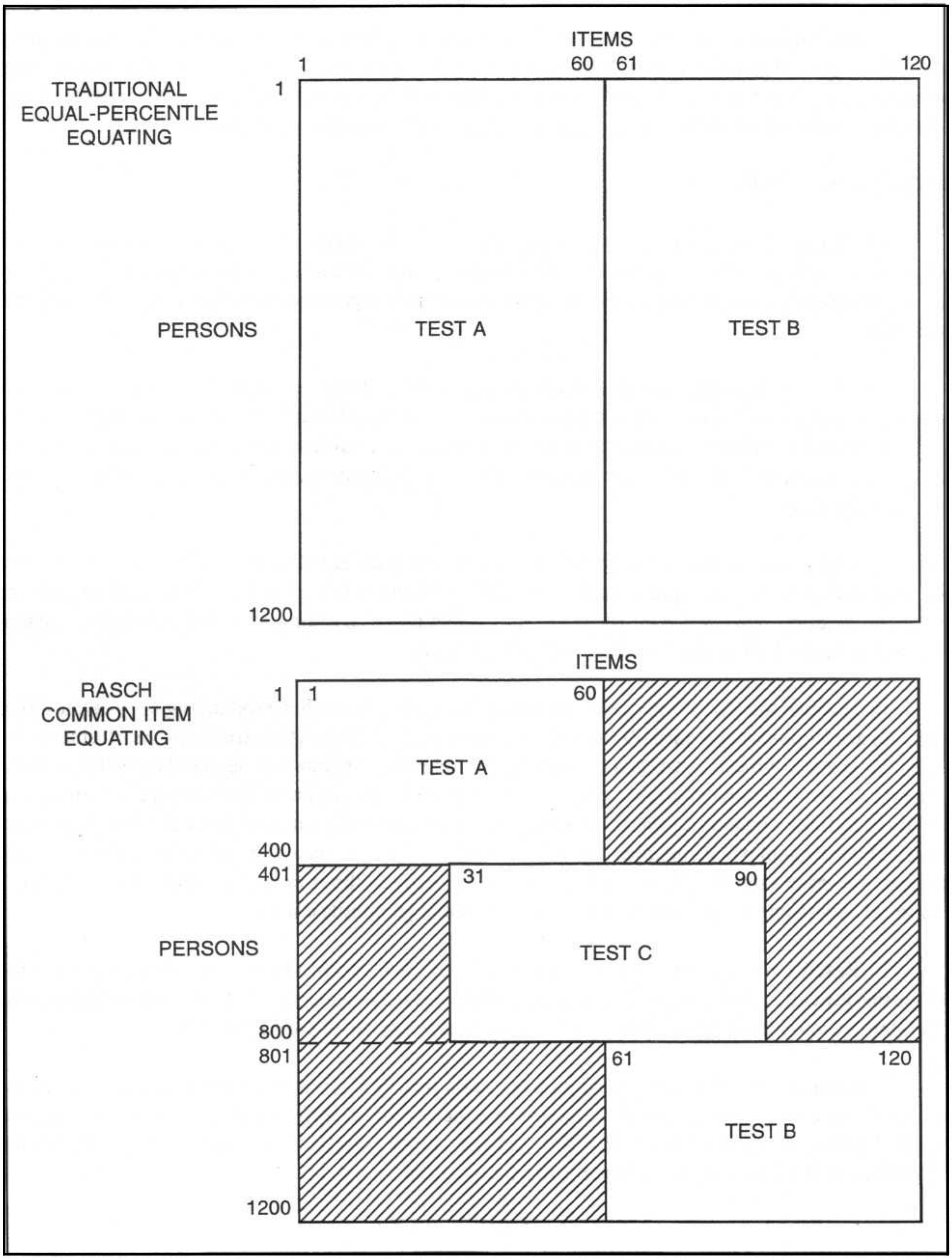
In Rasch equating, the separate calibrations of each test produce a pair of independent item difficulties for each linking item. The equating model asserts that each pair of estimates are statistically equivalent except for a single constant of translation common to all pairs in the link.

If two tests, A and B, are joined by a common link of K items and each test is given to its own sample of N persons, then d_{iA} and d_{iB} can represent the estimated difficulties of item i in each test with standard errors of approximately $2.5/N^{1/2}$ and the single constant necessary to translate all item difficulties in the calibration of Test B onto the scale of Test A is

$$G_{AB} = \sum_{i}^{K} \left(d_{iA} - d_{iB} \right) / K$$

Figure 11.1

Traditional and Rasch equating designs.



with standard error of approximately $3.5/(NK)^{1/2}$ logits.

In contrast to traditional equating, in which no quality control is available, the quality of this Rasch link can be evaluated by the fit statistic:

$$\sum_{i}^{K}(d_{iA}-d_{iB}-G_{AB})^2(N/12)[K/(K-1)] \sim X_K^2$$

which, when the two tests do fit together, will be distributed approximately chi-square with K degrees of freedom.

In addition, the individual fit of each item link can be evaluated by

$$(d_{iA}-d_{iB}-G_{AB})^2(N/12)[K/(K-1)] \sim X_1^2$$

which, when the performance of that item is consistent with the equating, will be approximately chi-square with one degree of freedom.

These simple fit statistics enable detailed, item by item control and remediation of test equations.

When using these chi-square statistics to judge link quality we keep in mind how they are affected by sample size. When N exceeds 500 these chi-squares can detect link flaws too small to make any noteworthy difference in G_{AB}, too small to matter. (When calibration samples are large, the root mean square misfit is more useful. This statistic can be used to estimate the logit increase in calibration error caused by link flaw.)

In deciding how to act on evaluations of link fit, we also keep in mind that random uncertainty in item difficulty of less than .3 logits has no discernible bearing on person measurement (Wright & Douglas, 1975, 35-39).

Because of the way sample size enters into the calculation of item difficultly and hence into the evaluation of link quality, we can deduce from these considerations that samples as small as 200 persons and links of as few as 10 good items will always be more than enough to supervise link validity at better than .3 logits. In practice we have found that we can construct useful and stable item banks with sample units as small as 50 persons.

THE COMMON LINK

The basic structure required to calibrate many items onto a single variable is the common item *link* in which one set of linking test items is shared by and so connects together two otherwise different tests. An easy and a hard test can be linked by a common set of intermediate items. These linking items are the "hard" items in the easy test but the "easy" items in the hard test (Figure 11.2).

With two or more test links we can build a *chain* of the kind shown in Figure 11.3.

The representation in Figure 11.3 can be conveyed equally well by the simpler scheme shown in Figure 11.4 which emphasizes the links and facilitates diagraming more complicated

Figure 11.2

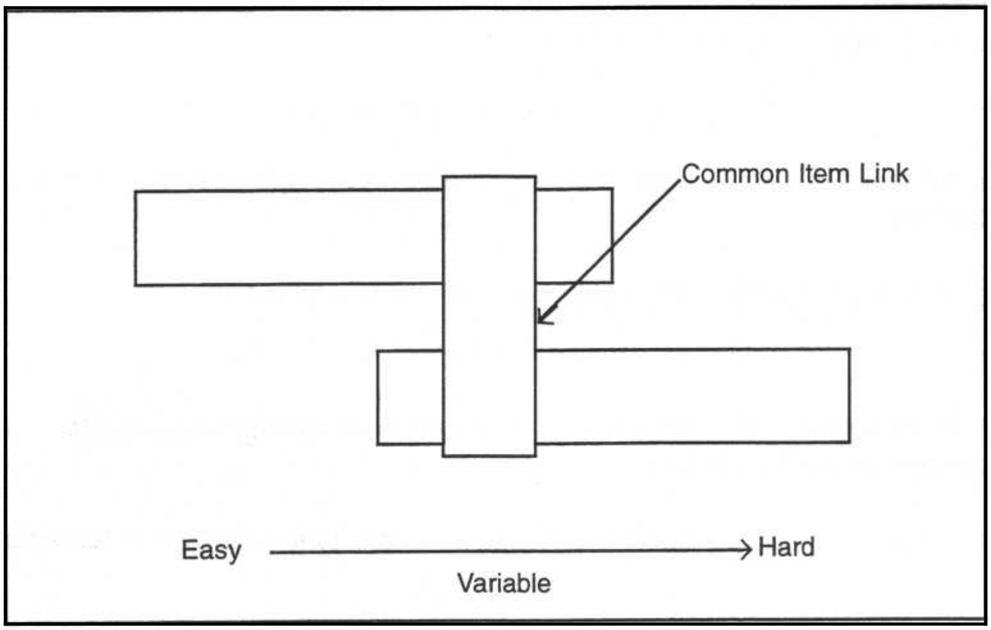
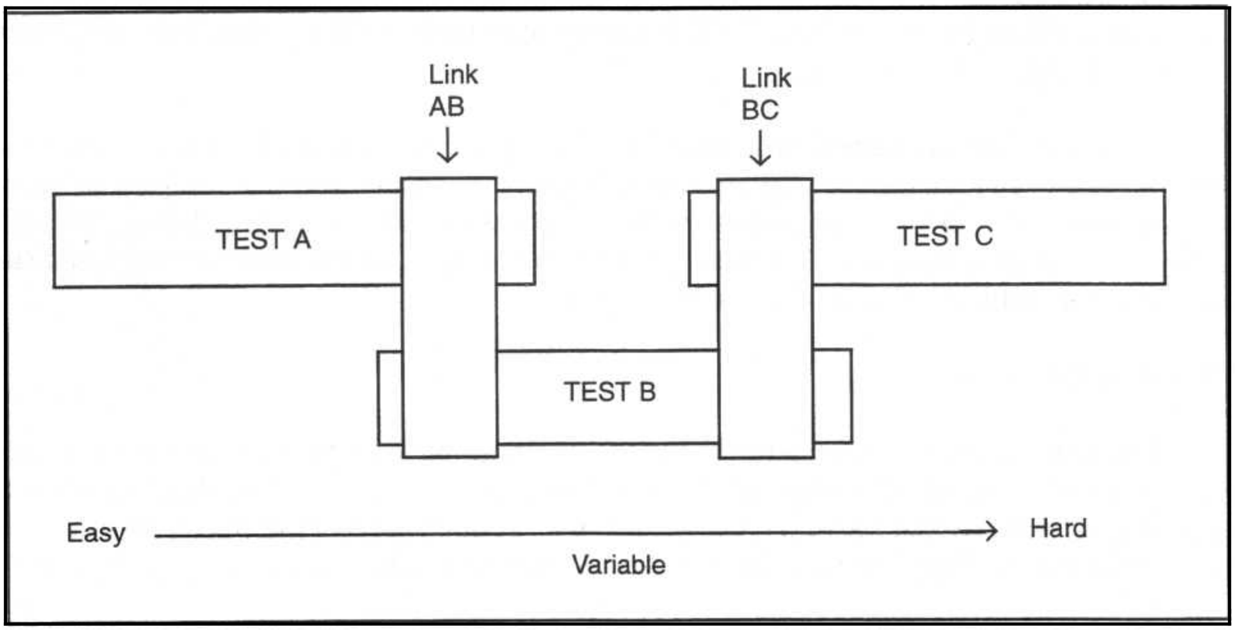


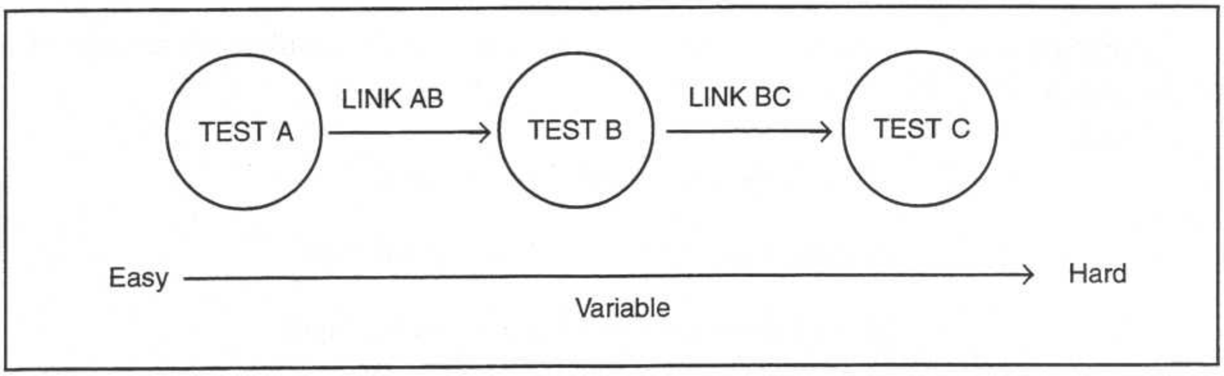
Figure 11.3



linking structures. Each circle indicates a test sufficiently narrow in range of item difficulties to be manageable by a suitably chosen sample of persons.

Figure 11.4

A chain with two links (simplified).

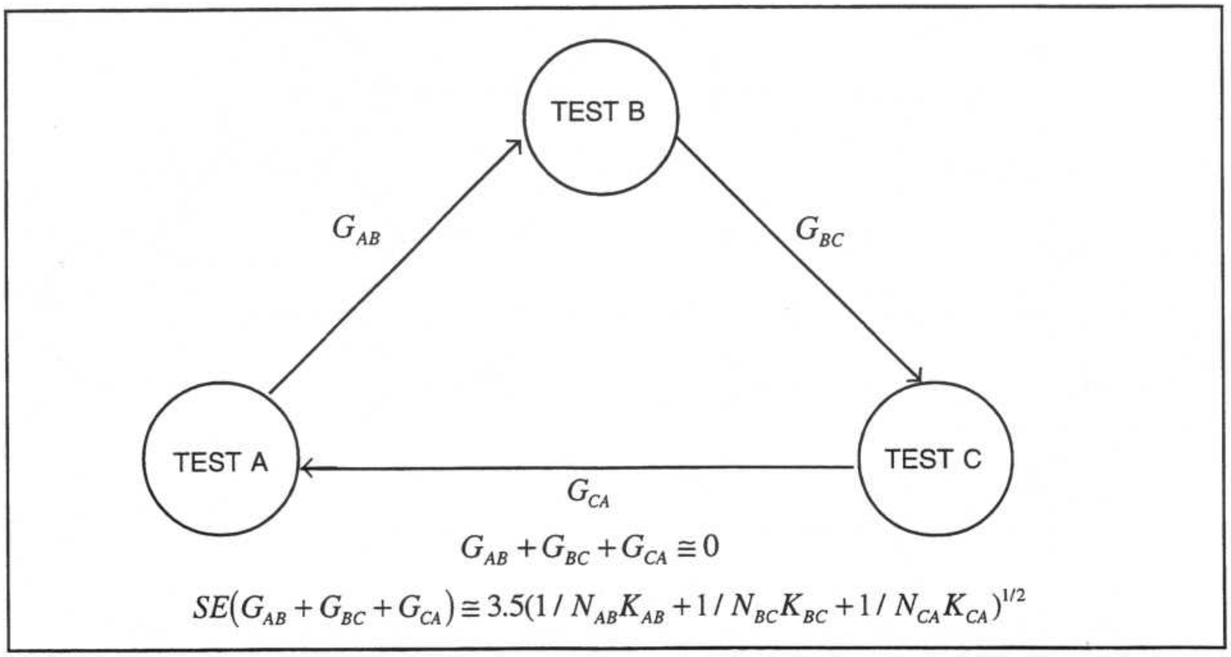


Each line connecting a circle represents a link of common items shared by the two tests it joins. Tests increase in difficulty horizontally along the variable and are comparable in difficulty vertically.

Links can be constructed to form a *loop* as shown in Figure 11.5.

Figure 11.5

A loop of three links.



$$G_{AB} + G_{BC} + G_{CA} \cong 0$$

$$SE(G_{AB} + G_{BC} + G_{CA}) \cong 3.5(1 / N_{AB}K_{AB} + 1 / N_{BC}K_{BC} + 1 / N_{CA}K_{CA})^{1/2}$$

The loop is an important linking structure because it yields an additional verification of link coherence. If the three links in a loop are consistent, then the sum of their three link translations should estimate zero.

$$(G_{AB} + G_{BC} + G_{CA}) \cong 0$$

where G_{AB} means the shift from Test A to Test B as we go around the loop so that G_{CA} means the shift from Test C back to Test A.

Estimating zero statistically means that the sum of these shifts should come to within a few standard errors of zero. The standard error of the sum $(G_{AB}+G_{BC}+G_{CA})$ is:

$$3.5(1/N_{AB}K_{AB} + 1/N_{BC}K_{BC} + 1/N_{CA}K_{CA})^{1/2}$$

in which

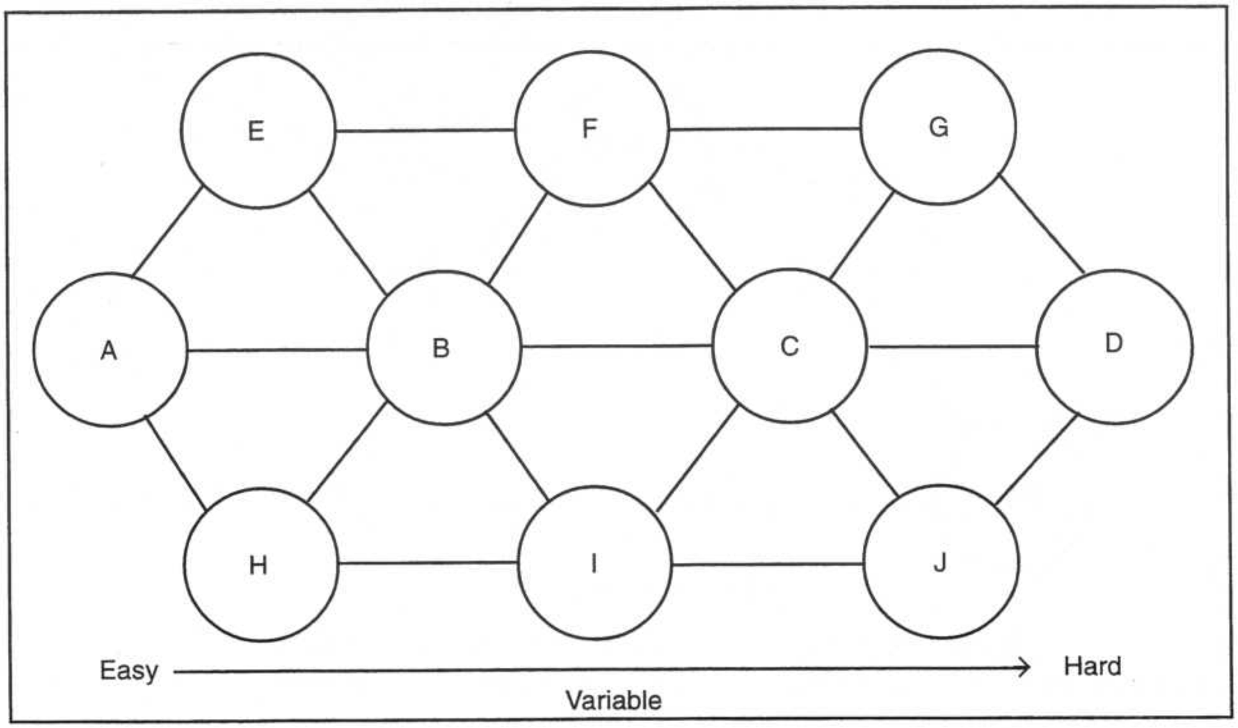
N = the various calibration sample sizes and

K = the various numbers of items in each link.

With four or more tests we can construct *networks* of loops. Figure 11.6 shows ten tests marking out several levels of difficulty from Tests A through D. This network could connect ten 60-item tests by means of nineteen 10-item links to construct a bank of 600-190=410 commonly calibrated items. If 100 persons took each test, then 410 items could be evaluated for possible calibration together from the responses of only 1,000 persons. Even persons at 50 per test would provide a substantial purchase on the possibilities for building an item bank out of the best of the 410 items.

Figure 11.6

A network connecting ten tests with nineteen links.



The building blocks of a test network are the loops of three tests each. When a loop fits the Rasch model, then its three translations will sum to within a few standard errors of zero. The success of the network at linking item calibrations can be evaluated from the magnitudes and directions of these loop sums. Shaky regions can be identified and steps taken to avoid or improve them.

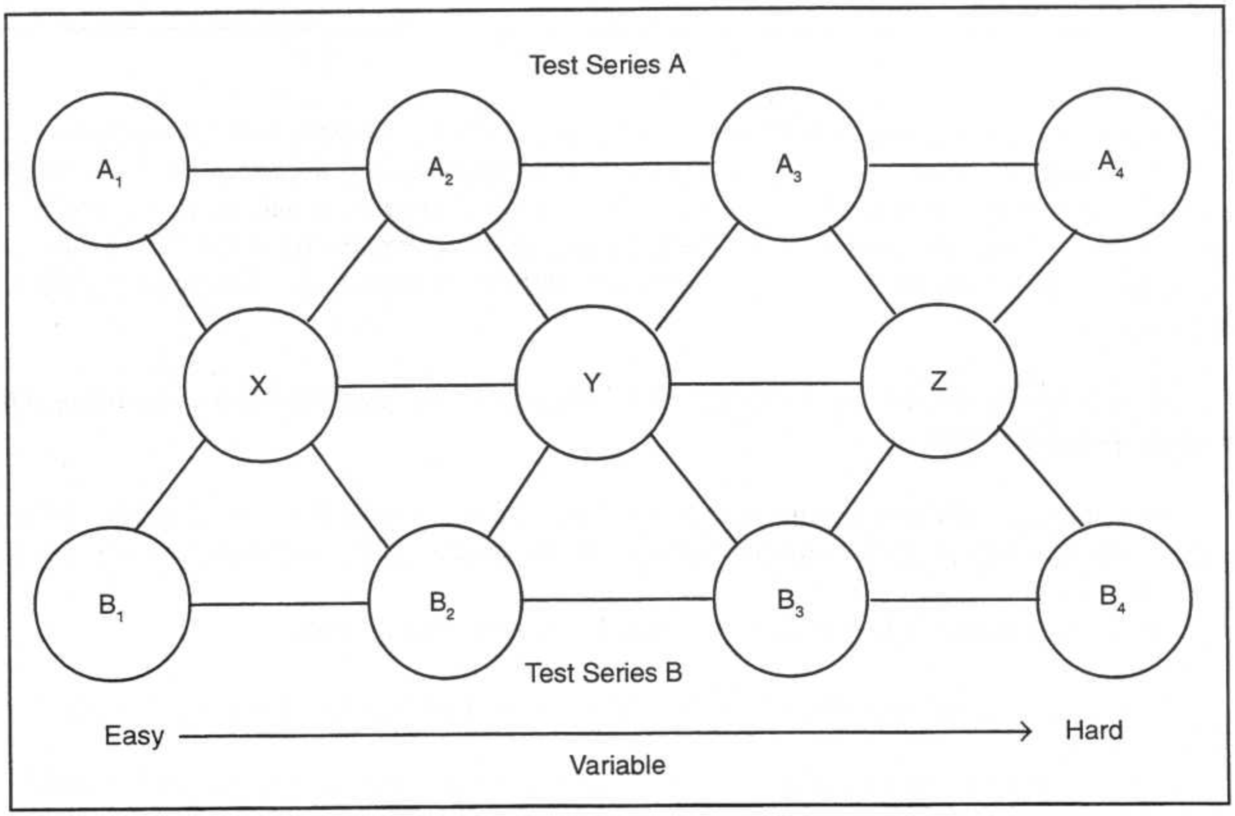
The implementation of test networks leads to banks of commonly calibrated items far larger in number and far more dispersed in difficulty than any single person could ever handle. The resulting item banks, because of the calibration of their items onto one common variable, provide the item resources for a prolific family of useful tests, long or short, easy or hard, widely spaced in item difficulty or narrowly focused, all equated in the measures they imply.

BANKING EXISTING TESTS AND ITEMS

These methods for building item banks can be applied to existing tests and items, if they have been carefully constructed. Suppose we have two non-overlapping, sequential series of tests A1, A2, A3, A4 and B1, B2, B3, B4 which we want to equate. All eight tests can be equated by connecting them with a new series of intermediate tests X, Y and Z made up entirely from items common to both series as shown in Figure 11.7.

Figure 11.7

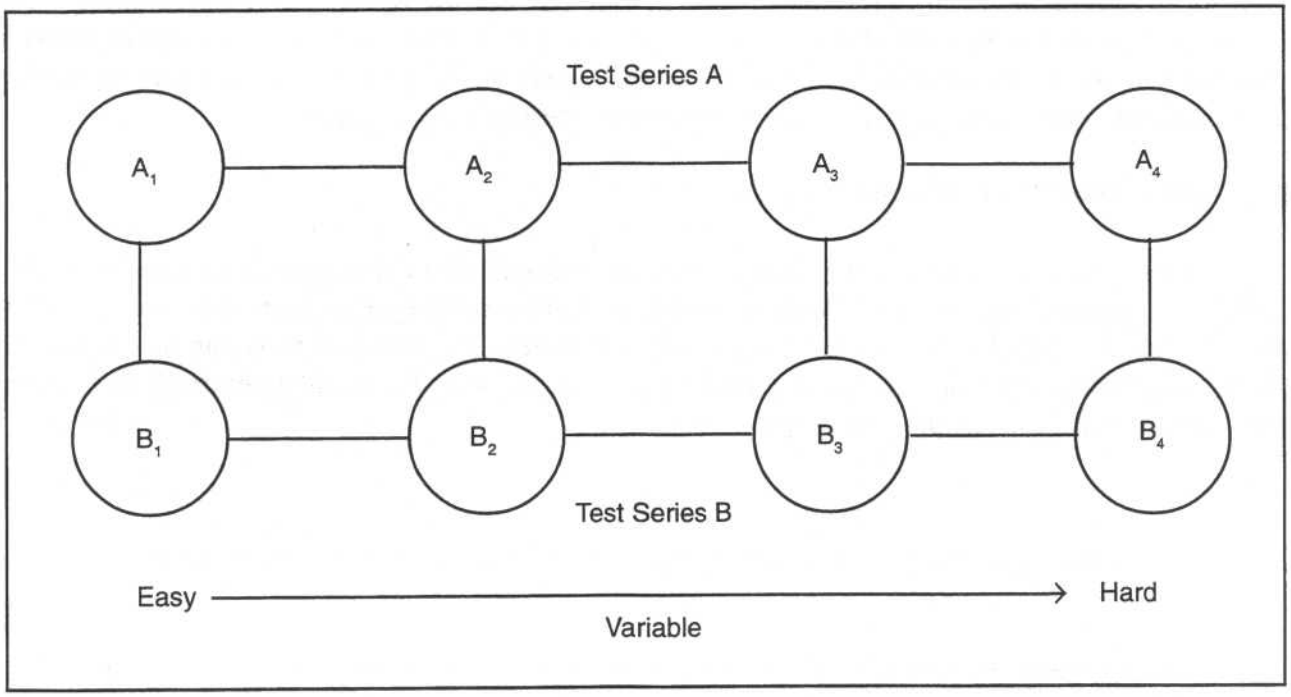
Connecting two non-overlapping test series by intermediate linking tests.



Were the A and B series of tests in Figure 11.7 still in the planning stage, they could also be linked directly by embedding common items in each test according to the pattern shown in Figure 11.8.

Figure 11.8

Connecting two test series by embedding common links.



Networks maximize the number of links among test forms because each form is linked to as many other forms as possible. To illustrate, take a small banking problem where we use 10 items per form in a *web* in which each one of these 10 items also appears in one of 10 other different forms. The complete set of 10+1=11 forms constitutes a web woven out of 11 x 10/2=55 individual linking items. Every one of the 11 forms is connected to every other form. The pattern is pictured in Figure 11.9.

The number entered in each cell is the identification of the item linking the two forms which define the position of that cell.

In this design, the web is complete because every form is connected to every other form. In the use of webs to build banks, however, there are three constraints which affect their construction:

1) the total number of items we want to calibrate into the bank,

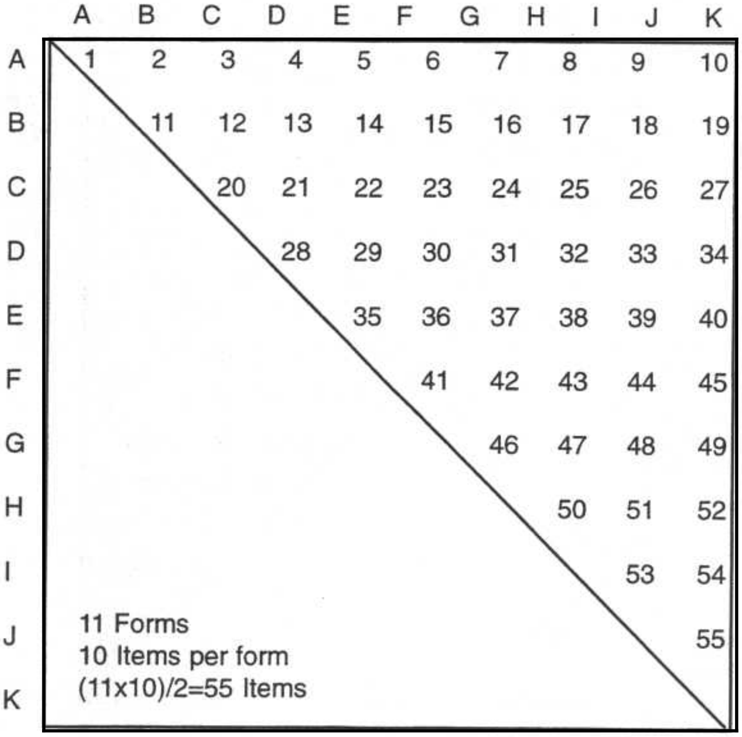
2) the maximum number of items which we can combine into a single form and

3) the extent to which the bank we have in mind reaches out in difficulty beyond the capacity of any one person.

The testing situation and the capacity of the persons taking the test forms limit the number of items we can put into a single form. Usually, however, we want to calibrate many more items that we can embed in a complete web like the one illustrated in Figure 11.9. There are two possibilities for including more items.

Figure 11.9

A complete web for parallel forms.



	A	B	C	D	E	F	G	H	I	J	K
A		1	2	3	4	5	6	7	8	9	10
B			11	12	13	14	15	16	17	18	19
C				20	21	22	23	24	25	26	27
D					28	29	30	31	32	33	34
E						35	36	37	38	39	40
F							41	42	43	44	45
G								46	47	48	49
H									50	51	52
I										53	54
J											55
K											

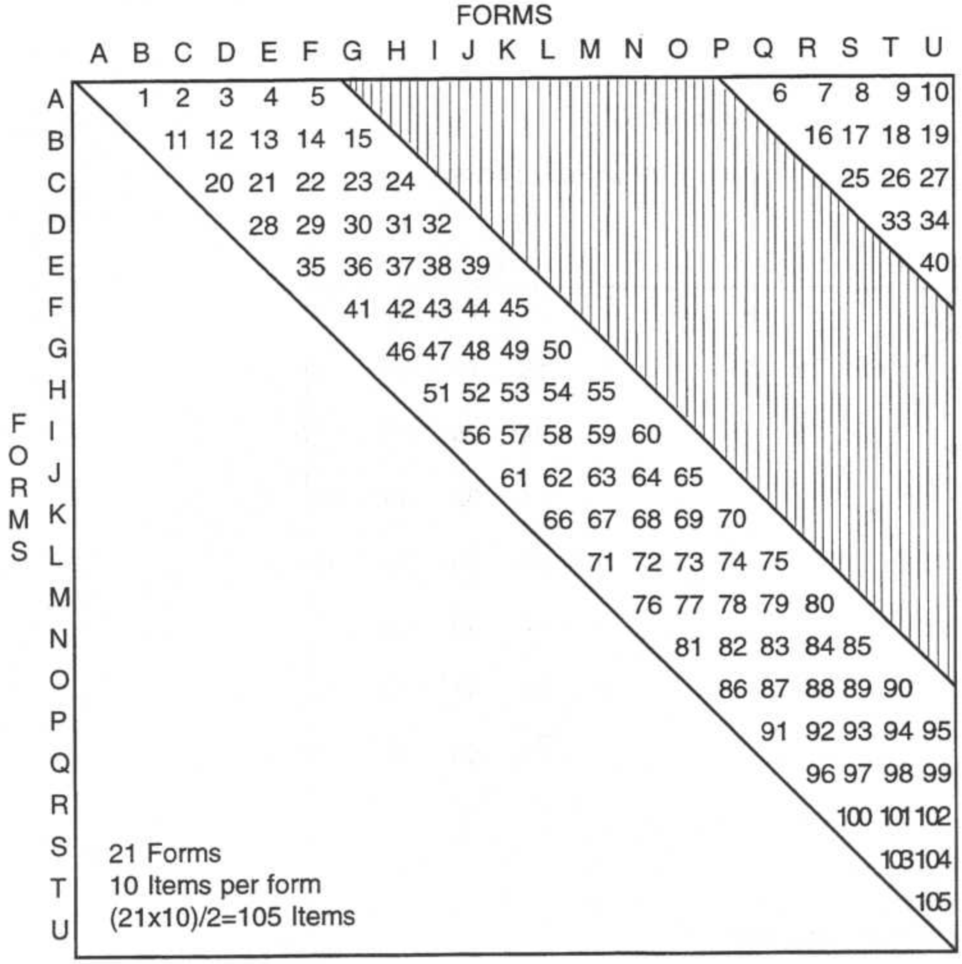
11 Forms
10 Items per form
(11x10)/2=55 Items

The simplest, but not the best, is to design a "nuclear" complete web which uses up some portion of the items we can include in a single form. Then we fill out the required form length with additional "tag" items. These tag items are calibrated into the bank by means of the link items in their form. Unlike the link items, however, which always appear in two forms, the tag items appear in only one form and so give no help with linking forms together into one commonly calibrated blank.

Another possibility, which is better statistically, is to increase the number of forms used while keeping the items per form fixed at the required limit. This makes the web incomplete but in a systematic way. The paired data on every item appearing twice can be used to evaluate the coherence of bank calibrations. Figure 11.10 shows an "incomplete" web for a 21 form design with 10 items per form, as in Figure 11.9, but connecting nearly twice as many items.

Figure 11.10

An incomplete web for parallel forms.



Formulation:
$N = ML/2$

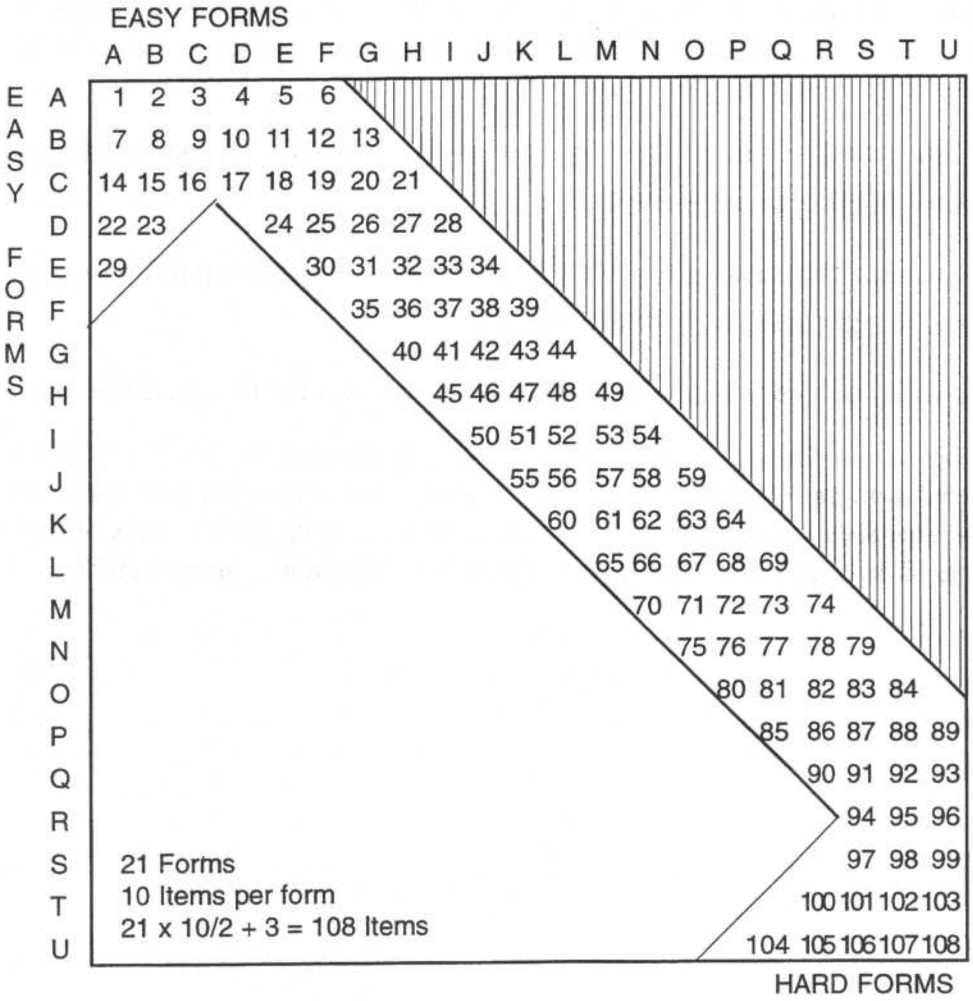
Where:
N = number of items
M = number of forms
i.e., $2\ N/L$
L = number of items (or links) per form must be even

The incomplete web in Figure 11.10 is suitable for linking a set of parallel test forms. When the reach of the bank goes beyond the capacity of any one person, however, neither of the webs in Figures 11.9 and 11.10 will suffice, because we will be unable to combine items from the easy and hard ends of the bank into the same forms. The triangle of linking items in the upper right corners of Figures 9 and 10 will not be functional and will have to be deleted. In order to maintain the balance of linking along the variable we will have to do something at each end of the web to fill out the easiest and hardest forms so that the extremes are as tightly linked as the center.

Figure 11.11 shows how this can be done systematically for a set of 21 sequential forms. We still have 10 items per form, but now only adjacent forms are linked together. There are no common items connecting the easiest forms directly with the hardest forms. But over the range of the variable the forms near to one another in difficulty level are woven together with the maximum number of links.

Figure 11.11

An incomplete web for sequential forms.



21 Forms
10 Items per form
21 x 10/2 + 3 = 108 Items

Formulation:
$N = ML/2 + K$

Where:
N = number of items (or links) in the bank
M = number of forms i.e., 2 $(N\text{-}K)/L$
L = number of items (or links) per form must be even
K = $L/4$, if $L/2$ is even
K = $(L + 2)/4$, if $L/2$ is odd

Each linking item in the webs shown in Figures 11.8, 11.9, 11.10, and 11.11 could in fact refer to a cluster of two or more items which appear together in each of the two forms they link. Sometimes the design or printing format of items forces them into clusters. This happens in reading comprehension tests where clusters of items are attached to reading passages. It also occurs on math and information retrieval tests where clusters of items refer to common exhibits. Clustering increases the item length of each form by a factor equal to the cluster size.

The statistical analysis of a bank-building web is simple if the web is complete as in Figure 11.9. The row means of the corresponding matrix of form links are least square estimates of the form difficulties. We need only be careful about signs. If the web cell entry $G_{jk.}$ estimates the difference in difficulty $(\delta_i - \delta_k)$ between forms j and k and the form difficulties are centered at zero so that $\delta\cdot = 0$, then

$$G_{j\cdot} = \sum_{K}^{M} G_{jk} / M \approx \delta j$$

the row means of the link matrix calibrate the forms onto their common variable. Once form difficulties are obtained, they need only be added to the item difficulties within forms to bring all items onto the common variable shared by the forms.

The incomplete webs in Figures 11.10 and 11.11 require us to estimate row means from a matrix with missing data. The skew symmetry of link matrices helps the solution to this problem which can be done satisfactorily by iteration or regression.

When cells of the link matrix of $G_{jk.}$ are missing, then initial values for $G_{j.}$ can be obtained from Equation [6] by using zero's for the missing $G_{jk.}$.

The next step is to replace the missing $G_{jk.}$ with estimates from the corresponding $G_{j.}$ and $G_{k.}$ using $G_{jk} = G_{j.} - G_{k.}$ and recalculating $G_{j.}$ by Equation [6].

Iterations of this process will converge to stable values for the test form difficulties $G_{j.}$.

An even simpler but less informative solution is to express the data for all forms in one large matrix in which every item has its own column, every person has their own row and every intersection, at which a person does not address an item, is recorded as blank. This matrix, with its missing data, can be analyzed directly in one step with BIGSTEPS (Wright & Linacre, 1996).

12. BUILDING SCHOLASTIC VARIABLES

Any professional in a position of responsibility in a school must have a way to keep track of what the school produces in student achievement. The only way to account for what is produced is to have a means for measuring scholastic growth in the areas for which the school is responsible such as arithmetic and reading. We will call these growth areas, variables and speak of the school's job as the increase of students' standings on scholastic variables.

Although there is a great deal of information about how children are supposed to develop, and what kinds of stimulation is supposed to encourage them, unless school effects can be made explicitly quantitative, it is impossible to evaluate school success. A school has to account for its educational efforts. To do this, the school has to construct scholastic variables on which the results of teaching can be measured and devise ways to measure these results.

THE MEASUREMENT PROBLEM

No school can neglect the measurement problem. Schools have to deal with it because it is the only way they can report to themselves or to the people to whom they are responsible the extent to which they are accomplishing their reason for existing. Schools must be able to measure their students' achievement.

How can school variables be defined and measured? We are deluged with tests from competing publishers who claim that their products relate the scores of increasingly difficult tests and provide indications of growth in particular areas. We believe that scholastic growth can be manifest by performance on test items. There seems no doubt that useful and relevant information can be obtained by giving students carefully selected questions to answer and then observing how they answer them. Thus we expect to use test performance to infer students' standing on the scholastic variable provoked by the test questions.

The trouble is that test publishers offer contradictory systems for quantifying test performance. The translation from one system to the other is neither definite nor agreed upon. Connecting test publishers' measures over the years of development is also difficult. Their equating systems are not convincing. Their reporting units of percentiles and grade equivalents are misleading.

Disagreement among test publishers is not the only problem. Difficulty in equating forms over the years of development is another. Local school dissatisfaction with national test items is yet another. The definition of an educational variable provided by a publisher, although marginally acceptable in New Jersey, may not be relevant to a school in Oregon. But no school dares to go off on its own without maintaining some connection to other schools. Neither does any school want to capitulate to a "national" standard imposed by some publisher. National tests offer a kind of comparability but lack relevance and flexibility. Local tests offer relevance and flexibility, but lack comparability. What is needed is a measurement system based on students' responses to test questions as the essential observation with tests made up of items focused on common scholastic variables of interest to the school but with results that can be compared from school to school. The ingredients of these tests must come from the school using them as well as from other reputable sources.

This flexibility, however, requires an objective method of constructing scholastic variables and checking consistencies that is accessible and workable for any school. We cannot know ahead of time whether there will be agreement among local definitions of the scholastic variables. Whether or not local schools are working with the same scholastic variables as state or national agencies is something that can only be established empirically. There also has to be a routine and objective way to find out how each test variable is working from moment-to-moment and place-to-place. Since the only way disagreements among differing agencies can be resolved is by an empirical check, the system of checking must be acceptable to all parties, even though they may disagree on the content of some items.

The way in which the relevance of items for a test is determined must be equally agreeable to national and local groups. It must have a methodological basis which transcends arguments about content. It must result in an objective measure which is immune to political manipulation.

Measurements can only be made through some kind of test situation. Tests can be valid if they are properly constructed. To be generally accepted, the test ingredients must represent both local and national wisdom and intention. The validity of items must be verifiable in some way equally satisfactory to all. Also, any measure, being an estimate rather than the thing itself, must be qualified by a standard error, the relevant index of its reliability as an agent of measurement.

To accomplish this it is necessary to develop banks of calibrated and validated items. These banks must consist of items which can be connected together in such a way that any selection from them can be used as a reasonable test for the common scholastic variable they, and all of the other items in the bank, define.

ITEM BANKS

This leads us to the concept of the item bank, with items contributed by local as well as national sources. National items would be items developed by expert teams. (See Choppin, 1968; Wright, 1977; Wright and Stone, 1979 and Wright and Bell, 1984 for an introduction to item banking.) Local items would be those items developed by school systems, by schools and even by an inspired teacher of the fourth grade who has insight into the scholastic development of the children in her class.

There must be room for all of these ingredients in the item bank. But having allowed this flexibility, there must be a method for checking whether each item is valid. It must also be possible for items that are valid to make up a test suitable to the occasion. Such a test must be equatable to any other test that might be constructed.

When a bank is well made and covers a wide range of the variable, then it is possible to have comparable measures available for individual children with whatever set of items they take and hence to follow student scholastic development longitudinally from the early grades. This requires an easy test that a second grader can take and another hard test measuring on the same scholastic variable but so much further along the variable that the same student can take it 10 years later and yet obtain a measure on the same scale and hence quantitatively comparable to the earlier measure. Items from these two tests could hardly be taken by both second and twelfth graders. Nevertheless, since we intend to compare the measures implied by each of these tests and to be able to say in an objective way how much a student has grown on the scholastic variable in those 10 years, we must find a way to connect these items so widely separated in difficulty to the scale of a single common scholastic variable.

A school system cannot escape the responsibility of measurement. But measurement needs to have certain characteristics in order to be useful to the school system. An item bank, solves a number of crucial problems. The developmental range problem and the equated forms problem is solved, and when the bank consists of local as well as national items, the relevance problem is also solved.

ITEM ANALYSIS

The occasion on which a student responds to an item, which we are relying on to show us where the student stands scholastically, is fraught with a variety of potential influences. But when we actually ask a student to answer a specific question, we would like to arrange things so that almost all that occurs at the moment is just an expression of that student's particular latent ability on the variable probed by that item. We are trying to provoke in the student's response a clear instance of this latent ability by means of the latent difficulty of the item that has been chosen. How well a student does on items of known difficulty can then be used to infer the student's measure on the latent variable.

However, when a student answers an item, there are the inevitable influences of motivation and distraction, as well as incidental elements in the item itself, which impede and facilitate the student's ability to solve it. Suppose it is a mathematics word problem. If the student is a good reader, it may be easier to do this item than if the student is a poor reader. It would be unfortunate if we failed to learn about a student's mathematical competence because reading difficulties on math items obscured the evidence the student would otherwise provide about math competency.

There are also administration and targeting difficulties which affect how students respond to items: guessing (on items too hard for them), sleeping (on items too easy for them), fumbling (on how the form is to be filled in), plodding (too slowly for the testing time and so not finishing) and bias (for and against success), all of which can interfere with measurement.

The system used for measurement must have a way to protect itself and its users against being mislead by unexpected disturbances in the observations from which the measure is estimated. The system must be able to detect spoiled measures. Once a test has been administered, we must be able to detect improbable divergence from expectation, to catch and correct for the influences of guessing, sleeping, fumbling, plodding and bias. We must be able to identify any secondary factors which interfere with performance on each item.

The measure estimated from a score on a test is an inexact estimate. We need to know not only the validity of the item responses on which the measure is based but the reliability, the error, of the measure.

A measurement project has two parts, item banking and person measuring. What is needed to manage these two parts is a common system which underlies both of them and so connects them together. The only hope we have of succeeding with a measurement project is to deduce a model for what we want to happen when a person encounters an item, a model formulated in the simplest practical terms, which also implements the basic requirements of measurement.

If we do not have a model, we cannot tell how to connect items together in the bank or how to free individual measurements from the particular items which happen to be used on a test. If we do not know what to expect, we have no way to tell whether a response is unexpected. We must be able to calculate from a model what we expect the answer to be so that we can observe whether a particular answer is surprising. The detection of irregularities requires a frame of reference by which a surprise can be defined.

This leads to the realization that, as far as measurement is concerned, it is not only sufficient but also necessary to pursue and enforce the fiction that each item can be characterized by a difficulty and nothing else and each person can be characterized by an ability and nothing else. We know that other factors always play a part, but with a simple model as our guide, we can always tell whether or not those other factors have spoiled the use of our simple model as a means for calibrating items and measuring persons.

When a simple model is put forward, that is not to say that what it is applied to is thought to be simple. Rather it is to assert that only through the construction of successful approximations to a simple model have we any chance of proceeding coherently and of making progress in managing a measurement project.

It is also not to say that when a student takes an item nothing is observed but the student's ability and the item's difficulty. Instead, it is our plan to make an effort to arrange and maintain things so that when a student takes an item most of what is observed is the expression of the student's ability against the difficulty of the item so that the observed response is dominated by student ability and item difficulty. Then, if something else happens, we can use the frame of reference of our simple model to identify the disturbance and to make correction for it.

THE MEASUREMENT MODEL

The traditional true score model specifies the observed score of a person taking a test as the sum of a true score and an error term:

$$x = t + e$$

where

x = OBSERVED SCORE
t = TRUE SCORE
e = ERROR

But we know that raw scores cannot be linear in what they represent and there is no useful theory for how big the true score error term should be. What we need, instead, is a different model which not only specifies that the person has an ability which is expressed in his behavior, but also that each item has a particular difficulty which is also expressed in any responses to that item, including the given response. Finally, we want a model which specifies how much deviation from expectation is reasonable and how much is excessive.

THE RASCH MODEL

The Rasch model (Rasch, 1960/1980) is a binomial probability model for a dichotomous right/wrong response. The Rasch model specifies that the probability of a right answer is defined by the difference between person ability and item difficulty. Then, when the probability of a right answer is calculated to be near zero, but a right answer is nevertheless observed, that right answer is obviously surprising. Being able to estimate the probability of a right answer enables us to be precise about the extent of our surprise.

The discrepancy between observation and expectation can be put into a standard form so that we can have a standard reference distribution for it. This quantifies the extent of our surprise. We will be surprised when a person of low ability achieves something that requires exceptional ability. When a person attempts an item many units harder than he is able and nevertheless gets it right, that right answer might have a probability of occurring less than five in 100 times. In that case, we might take the position that our surprise has become too large for comfort. Thus we have a means for being explicit about the extent of our surprise and, if we can agree among ourselves as to what level of improbability is unacceptable, then we have an explicit and public rule which we can apply to validate any observed response.

This enables us to take an objective stand with respect to what to do about correct answers to items too many units above a person's ability. Using the natural log odds units (logits) of Rasch measurement, a difference of three logits would produce an improbability of .05. In particular, we may decide to use such improbable answers only for diagnostic purposes and to exclude them from our measure of the person.

When an unexpected response occurs, we do not ignore it, what we do is to decide what to do with it. We might decide to use it in the score, or to delete it. We might decide to use it to diagnose the person or to diagnose the item. Both can be useful. When only one person uses one item unexpectedly, that, in itself, will not tell us whether the person or the item produced the unexpected condition. If we suspect it was the item, we will look at the responses of other people to see whether that item continues to behave poorly, e.g. for many boys, or for many fourth grade boys, or for whatever condition we suspect might make the item irregular. If, on the other hand, we are making an individual study of a child and are concerned about brain damage, emotional disturbance, a fixation, or an inhibition, then we could become especially interested in the diagnostic potential of unexpected responses, and might even seek to provoke such responses for diagnostic reasons.

A careful study of items is beneficial to any school system. It can produce uniform content-free public decision rules that can be applied fairly and without prejudice.

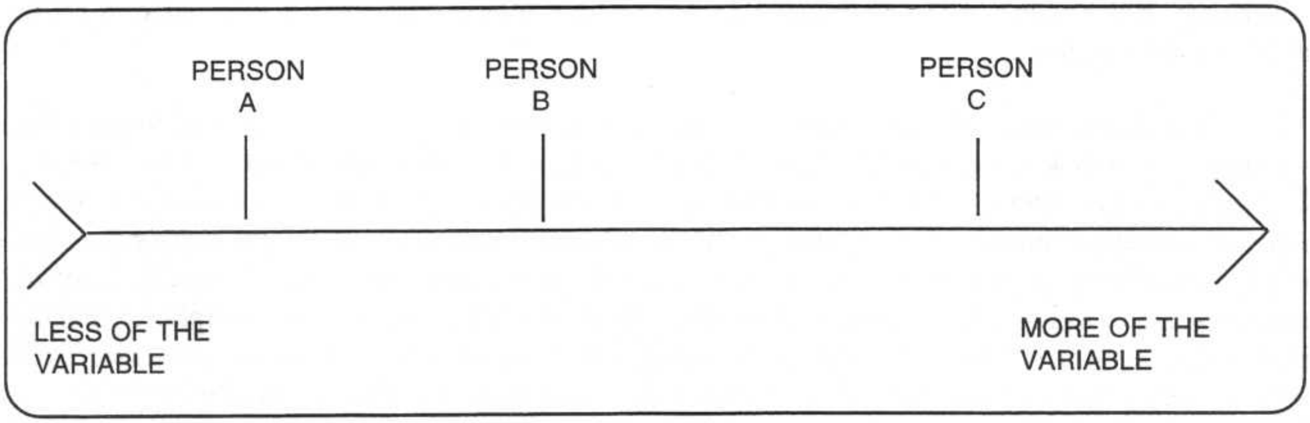
MEASUREMENT CRITERIA

LINEARITY

When we think about a variable, we have in mind the straight line so well represented by the familiar yardstick. One direction of this line represents more of the variable; the other, less. Person measures are locations along the interval scale of this line. This simple idea is illustrated in Figure 1.

That we employ the idea of a straight line when we think about variables like height and weight is obvious. But the relevance of this idea may not be as obvious when we speak of constructs such as intelligence or attitude. Nevertheless, we betray our reliance on this simple and useful idea whenever

Figure 12.1
Positions of persons A, B, C on the line of a variable.



we say that one person has a more positive attitude than another, or whenever we report an intelligence score for a child.

Our inevitable reliance upon this simple idea was noted long ago by L. L. Thurstone:

> 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 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 and less" judgment is conceptual, that it does not necessarily have the physical existence of a yardstick (Thurstone, 1928a, p. 532).

INVARIANCE OR OBJECTIVITY

When we measure a variable such as verbal ability, the measures we obtain must not depend upon the particulars of the items administered. Our ability measures must be freed of the particulars of the items taken in the same way that measures of height have a meaning which is independent of the particular yardstick used to obtain them.

Thurstone saw the necessity of this in 1926, and described the following requirements of a satisfactory measuring method:

> It should be possible to omit several test questions at different levels of the scale without affecting the individual score. It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point, being selected by the examiner, should not directly affect the individual score (Thurstone, 1926, p. 446).

Thurstone also pointed out the accompanying necessity of being able to obtain difficulty estimates for items which are freed from the particulars of the calibrating sample:

> One of the first requirements of a solution is that the scale values of the statements of opinion must be as free as possible, and preferably entirely free, from the actual opinions of individuals or groups. If the scale value of one of the statements should be affected by the opinion of any individual person or group, then it would be impossible to compare the opinion distributions of two groups on the same base (Thurstone, 1928b, p. 416).

And in the same year:

> The scale must transcend the group measured. One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid.
>
> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement (Thurstone, 1928b, p. 547).

The criteria for "measurement" are: logical ordering, linear scales, and objective comparisons. A model is needed which enables observations to be transformed into measures which meet these requirements.

In the early 1950's Georg Rasch (1960/1980) undertook to obtain measures of reading ability which were independent of the difficulty of the test taken:

> In a concrete formulation of this problem I imagined - in good statistical tradition - the possibility that the reading ability of a student at each stage, and in each of the two above-mentioned dimensions, could be characterized in a quantitative scale, but by a positive real number defined as regularly as the measurement of a length (Rasch, 1977, p. 59).

Rasch coined the term "specific objectivity" to describe comparisons among persons which are independent of the item parameters, and comparisons among items which are independent of the person parameters.

THE ITEM BANKING MODEL

Item banking can be accomplished with Rasch's psychometric methods. His measurement model describes the probable outcome of any encounter between a person and an item as entirely determined by two parameters - the "ability" of the person, represented by b; and the difficulty of the item, represented by d. If we use the numeric labels $x = 1$ to represent a correct answer and $x = 0$ to represent an incorrect answer, then Rasch's model for the probability of response x is:

$$P\{x=0,1|b,d\}=\exp[x(b-d)]/[1+\exp(b-d)]$$

$$\text{or} \qquad \log\left[\frac{P_{x=1}}{P_{x=0}}\right]=b-d$$

Rasch specifies that the log odds (logits) that a person with ability *b* answers correctly an item with difficulty *d* correctly be dominated by the difference (*b-d*) between person ability *b* and item difficulty *d*. This positions persons by their ability and items by their difficulty on the interval scale of a single variable which they share. The result is probabilities of potential interactions between persons and items which are positioned along one common line and specifications of expectations for all possible responses.

Because the parameters *b* and *d* in Rasch's model appear as separate terms in a linear function, they can be separated in the application of the model. The difficulty calibrations of the items can be estimated in a way which frees them from the ability distribution of the persons used and the ability measures of the persons can be estimated in a way which frees them from the difficulty distribution of the items they happen to take. This produces the "sample-free" item calibration and "test-free" person measurement (Wright, 1968) which Thurstone demanded.

The sufficient statistics for these results are the test score for each person and the number of persons who respond correctly to each item, the sample score for each item. But these scores are not yet calibrations or measures because they are nonlinear on the variable they are intended to measure and also sample and test dependent. The Rasch measurement procedure, however, can use these familiar raw scores to construct sample-free item calibrations and test-free person measures on a common linear scale.

Each item's raw score is specific to the ability distribution of the sample used on that item, but the linear Rasch item calibrations are adjusted so that the effects of this ability distribution are removed. The resulting sample-free item difficulties can be used to define a general variable of meaning which can reach beyond the particular occasion of calibration.

Each person's raw score is specific to the pattern of item difficulties in the particular test he or she takes, but the linear Rasch ability measures are adjusted so that the effects of this item difficulty distribution are removed and the person's ability is generalized onto the variable defined by the whole set of calibrated items.

Whether any particular set of calibrations and measures are in fact test-free and sample-free can be verified at each step by simple methods (Wright and Stone, 1979). Verification of fit to the Rasch measurement model provides an explicit quantitative definition of item function validity and person performance validity and enables continuous quality control over item calibration and person measurement.

With a workable calibration procedure and a method for the evaluation of fit, it becomes practical to turn our attention to a critical examination of the calibrated items to see what it is that they imply about the possibility of a variable of some useful generality. We can find out whether our calibrated items spread out in a way that shows coherent and meaningful direction. We can examine the hierarchy of item content and evaluate the extent to which this order indicates a line of increasing competence of recognizable meaning.

DEFINING A VARIABLE

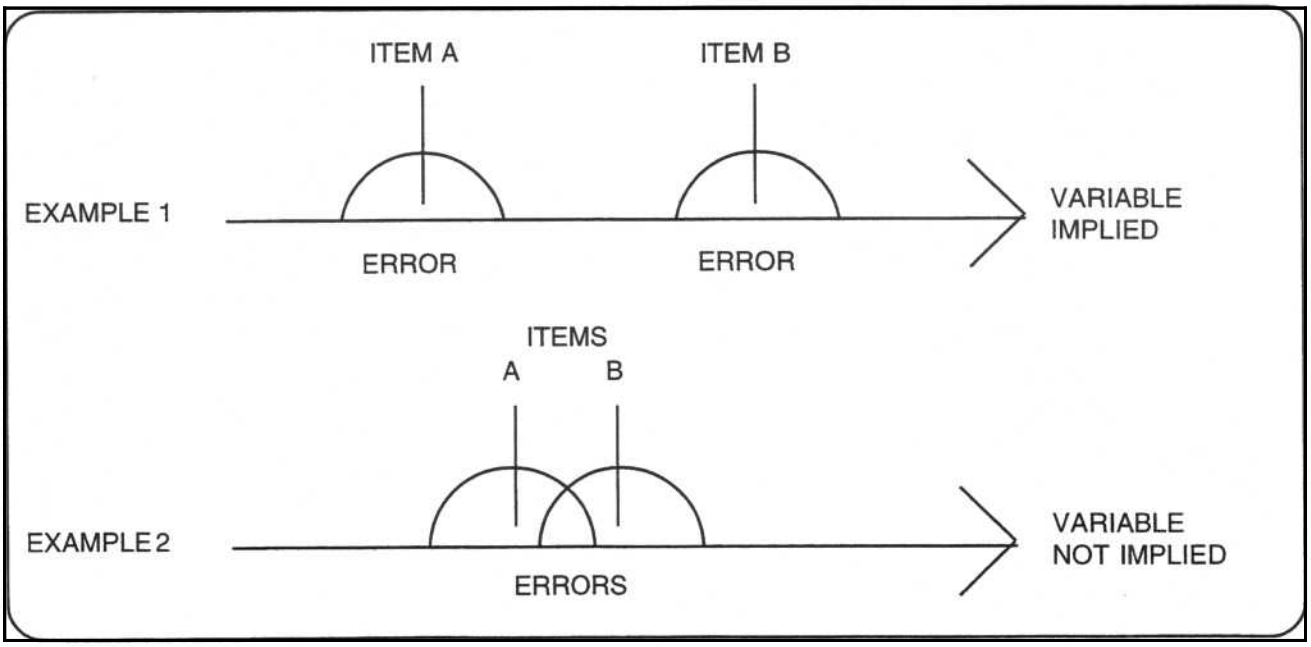
Our intention now is to show how calibrated items can be used to define a variable and how to find out whether the resulting operational definition of the variable makes content and construct sense. We begin by examining the degree to which the spread of item difficulties exceeds the standard error of their estimates, that is, the degree to which the data have given a direction to the variable. Consider, for example, the estimates of two item difficulties with their respective standard errors of estimation. In order for these two items to define a line between them, the difference between their estimates must be substantially greater than the standard error of this difference! Only when the two estimates are well separated by several calibration standard errors can we begin to see a line between the two items suggesting a direction for the variable defined by their content and order.

If, however, when we compare two item difficulty estimates, each bracketed by a standard error or two, they overlap substantially, then we cannot assume that the two values differ in difficulty, and as a result, cannot see a direction for the variable. Instead, the items define a point without direction. If the items do not spread out, then what have we defined? Only a point, perhaps on some variable, perhaps not. But the extent and hence the meaning of the variable is still missing.

Figure 12.2 illustrates this idea. In the first example we have items A and B separated from each other by several standard errors. Even with two items we see a direction to the variable as pointed out by these two items. In the second example, however, we find the two items so close to each other that, considering their standard errors, they are not separable. We have a point. But no direction has been established and so no quantifiable concept of the variable has as yet been implied. Only when items can be separated along the line representing the variable of interest have we begun to realize a construct.

Figure 12.2

Defining a variable.



In this discussion we have introduced a method by which objective scholastic variables can be constructed. Developing banks of Rasch calibrated items is the method. Item analysis is the tool by which these banks are built. The measurement model of George Rasch provides the means by which we construct these measurements. It provides a workable calibration procedure and a method for the evaluation of fit. Successful item bank construction can meet the criteria that Thurstone stated in defining the requirements of measurement - valid ability scales which transcend their particular items.

In the accompanying chapters we explore each of the above areas in detail.

13. ITEM BANKING

This chapter discusses the curricular implications of item banking and its usefulness to all who depend on tests to evaluate educational achievement. We review the psychometric basis of item banking and give equations for building a bank. We conclude by showing how item quality control can be maintained over a bank of items.

THE IDEA OF ITEM BANKING

A mere collection of items is not an item bank. An item bank is a set of carefully composed and jointly calibrated items that develop, define and quantify a single common theme and hence provide an operational definition of one variable.

The first step in building an item bank is to develop its specifications. If we are building a scholastic variable it will be necessary to define the curriculum area and then to determine which items explicate it. To do so requires the expertise of professionals familiar with that curriculum area: teachers and curriculum experts.

We need a plan for the scholastic variable which is sufficiently detailed to specify how the items are expected to be ordered by difficulty along one main line of scholastic growth. This is important because it is in this beginning step that we demonstrate our understanding of the line of inquiry that is intended to define the scholastic variable under construction. If we discover that we do not have a clear enough understanding of the items to arrange them by difficulty order, then we have discovered that we do not know enough about what we are trying to do to succeed.

To accomplish item development:

1. Choose or write an item that you consider clearly on the line of the scholastic variable to be constructed.

   Intended Difficulty: ----1----> Harder

2. Add a second item written to be <u>easier</u> than the first item.

   Intended Difficulty: ----2----1----> Harder

3. Add a third item written to be <u>harder</u> than the first item.

   Intended Difficulty: ----2----1----3----> Harder

4. Next, add a fourth item positioned between items one and two and a fifth positioned between items one and three.

   Intended Difficulty: ----2----4----1----5----3---> Harder

5. Continue this stepwise process by positioning successively easier and harder items which extend the line of existing items and by filling in the spaces between these items with additional items positioned in difficulty between pairs of existing items.

This process of constructing the variable with items can be refined by re-positioning items upon further consideration and by review by other experts. The final line of items should show an ordering of items positioned by their intended difficulty from the easiest to the hardest. Successful construction of such a line of ordered items is an indication that the essence of the variable is understood by the item writers, and that the growth line implied by the scholastic variable and the items which define it belong together and lead somewhere. When we are not able successfully to position items along a line of growth by their difficulty, that is a sign that we do not understand our idea of the variable or the items required to describe it well enough to proceed.

Each item must represent an element in the strand of the scholastic variable we are building and each item must test some knowledge, skill or behavior at a specified position along the increase of that variable. When the items are empirically calibrated, these "conceptual" positions can be verified and improved. When, finally, the items are well-located along the line of a scholastic variable, then the scholastic variable has acquired a meaningful and useful operational definition.

Items with low calibration values entail easy tasks that define the low end of the variable. Items with high calibration values entail difficult tasks that define the high end of the variable. The arrangement of items by their order of calibrations from easy to hard describes the path of learning that most students follow as they progress along the line of the scholastic variable. The empirical item calibrations can be obtained by applying the Rasch model for what ought to happen when a student attempts an item (Rasch, 1960/1980; Wright & Stone, 1979). This probability model imposes an orderly response process on the data. The probabilities obtained specify what is expected to occur, with some give and take, because no student will follow the expected line exactly.

The process of item planning, writing and positioning, along with the confrontations and revisions provoked by subsequent item calibrations, is an integrated and constructive dialogue between the item construction phase of bank development and the item calibration phase - between theory and practice. This dialogue will progress in successive stages as better and better confirmation of item positions is achieved and the operational definition of the scholastic variable evolves. Continual monitoring of the bank building process is both required and beneficial.

When a scholastic variable is well understood, the task of constructing its item bank is straightforward. But when the variable is newly conceived or not clearly understood, the interactive process between item positioning and item calibration may require many stages before useful agreement between intention and realization, between idea and experience is achieved.

It is important to recognize that the agreement to be achieved between theory and practice is the method for control over item development quality. Creative item writing is required to capture and implement the essence of a scholastic variable. The empirical calibration of these items gives the item writers feedback on the utility of their creative efforts.

Reviewing the evolving line of items from easy to hard along the intended variable promotes communication between the specialists of curriculum and teaching and those of test construction. The resulting marriage of these two specialty areas can produce valid scholastic variables defined by operationally efficient items.

THE USEFULNESS OF ITEM BANKS

A well constructed and organized item bank enables a wide variety of tests. Each test can be tailored to the objectives of its use and yet be quantitatively connected to the common core of bank items. Additional items can be added whenever their calibrations are found to fit the growing common core of calibrated items.

A well constructed item bank provides the elements necessary for designing the best possible test for any assessment purpose. It is not necessary for every student to take the same test in order to be able to compare results. Students can take only those items closest to their level of development as in computer assisted instruction. The number of items, level, range of difficulty and content can be selected individually from the bank. Each individualized test maintains quantitative comparability because any test formed from calibrated bank items, on which a valid pattern of performance is obtained, can be automatically equated through the calibration of the test items to all of the items in the bank and so to all of the measures produced by every other test that has ever been or might sometime be formed from this bank.

A very wide-range test for general screening can be formed as well as narrow tests for specific purposes. The two procedures of wide-range screening and narrow-range measuring can be combined to implement adaptive testing. The wide-range test locates the student's general area on the line of the scholastic variable and the narrow-range test pinpoints the location for the most efficient measurement of that student.

BUILDING AN ITEM BANK

TO CONSTRUCT AN ITEM BANK:

1. Begin with a pool of items dominated in their content by a common curriculum line. These items are best when constructed and arranged according to a clear hierarchy of increasing conceptual difficulty.

2. Apportion these items among test forms so that there is a web of common items which forms a network of connections among all test forms. This web can reduce the test size of each form to manageable length and yet distribute all items over the many forms connected by the web of shared items.

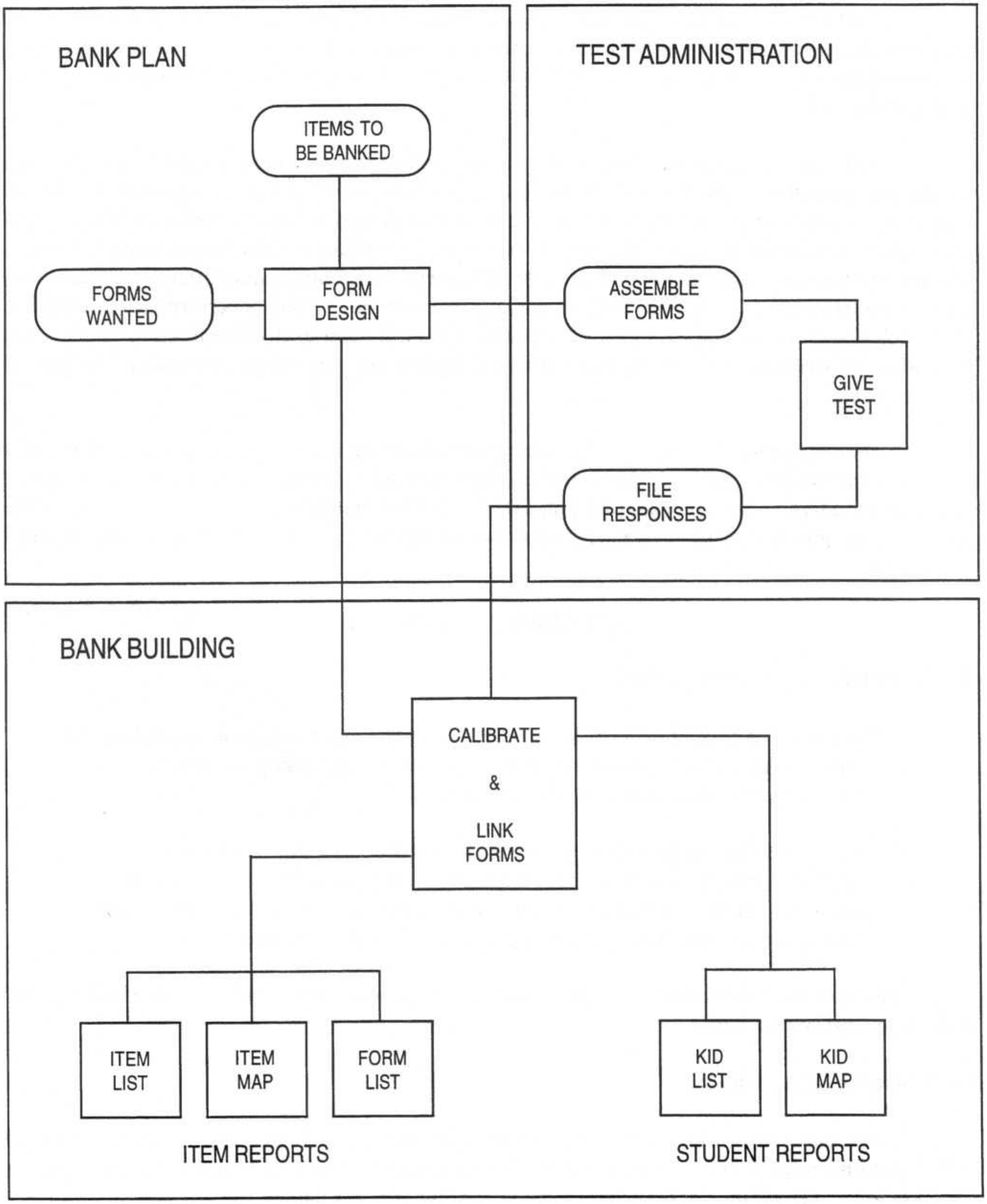
The flow chart in Figure 13.1 outlines the basic steps necessary to build a pool of coordinated items into a calibrated bank.

DESIGNING TEST FORMS

Items must be distributed among test forms so that there is a web of common item connections which maximizes the statistical strength of the linking structure, while meeting the practical requirements of the test situation (for details see Wright & Stone, 1979, Chapter 5).

Figure 13.1

Flowchart for building an item bank.



Design input includes the number of items to be calibrated, the number of items desired per form, the number of items desired per link, the expected difficulty of each item and whether the pattern of form difficulties is to be horizontal or vertical. The design determines the number of links per form, total number of links and total number of forms necessary for an optimal web.

The design process constructs a file of item specifications from which the banking system works. This list includes item identification number, name, link number, expected difficulty, correct responses, and associated forms so that item test form placements can be checked and listed item-by-form and also form-by-item in their within-form position in order to facilitate the verification of content coherence and form assembly.

CALIBRATING TEST FORMS

When forms are designed, assembled and administered, student responses are collected, recorded and filed in an individual record for each student that includes student identification, form taken, and the student's item response string. This student file is the form calibration input. The item file prepared during form design and the student file obtained from testing, are used to calibrate items within each form in order to analyze within-form item and student fit and then to calibrate all items and measure all students simultaneously on one common linear variable. (A useful computer program for this is *BIGSTEPS*, Wright & Linacre, 1997.)

The form equating, accomplished by the single simultaneous analysis of all forms, can be evaluated in detail by explicitly linking the separate analyses of each form in which item difficulties are still relative to the local origin defined by each form. Connections among forms can be made explicit by a link analysis of the connections of all forms to the single common scale.

ANALYSIS OF FIT

Analysis of fit evaluates the degree of consistency between observation and expectation and the extent to which any subdivisions of observed data (by group, grade level, sex, etc.) produce statistically equivalent item and form calibrations. There is a hierarchy of fit statistics available to implement fit analysis.

ITEM WITHIN-FORM FIT

A routine check on whether item difficulties are sample-free is done during form calibration. If item estimates are invariant with respect to student abilities, student sample subdivisions will give statistically equivalent item difficulties. One way to evaluate sample-freeness is to divide the sample into raw score subgroups and then to compare the observed successes on each item *i* in each raw score subgroup *g* with the number of successes predicted for that subgroup. If the general parameter estimates are adequate for describing score group *g*, then the observed number correct in group *g* will be near the estimated model expectation

$$R_{gi} = \sum_{r \in g} N_r p_{ri} \qquad 13.1$$

with model variance

$$s_{gi}^2 = \sum_{rEg} N_r p_{ri}[1-p_{ri}]$$

where 13.2

$$p_{ri} = \frac{\exp(b_r - d_i)}{1+\exp(b_r - d_i)}.$$

N_r is the number of students with raw score *r* and p_{ri} is the estimated probability of success for a student with score *r* on item *i*, given the general ability and difficulty estimates b_r for score *r* and d_i for test item *i*.

If observed and expected numbers correct are statistically equivalent, given the model variance of the observed, then there is no evidence against the conclusion that the subgroup concurs on the estimated difficulty of item *i*. The statistical precision (reliability) of this estimate can be specified with its modeled standard error. Similar analyses can be done for student subgroups defined in other ways.

Another way to check within-form item fit is to evaluate the agreement between the variable manifested by item *i* and the variable defined by the other items. A useful statistic for this is an "infit" mean square in which the standard squared residual of observation *x* from its expectation *p*, $z_{ni}^2 = (x_{ni} - p_{ni})^2 / [p_{ni}(1-p_{ni})$, for each student *n*'s response to item *i*, is weighted by the information in the observation, $q_{ni} = p_{ni}(1-p_{ni})$, and summed over the *N* students.

$$V_i = \frac{\sum_n^N [z_{ni}^2 q_{ni}]}{\sum_n^N q_{ni}} [N/(N-1)] \quad 13.3$$

This "infit" statistic is useful because it is robust with respect to idiosyncratic outliers. The alternative "outfit" statistic that detects outliers is the unweighted mean square,

$$U_i = \sum_n^N z_{ni}^2 / (N-1) \ . \quad 13.4$$

When data fit the model, these statistics estimate one with variance of order $[2/(N-1)]$.

For more exact estimates of these variances see Rasch, 1980, pp. 193-194 or Wright & Masters, 1982, p. 100.

CALCULATING TEST FORM LINKS

When the items in each form have been calibrated separately within each form, there are as many difficulty estimates for each item as there are forms in which it appears. The items that appear in more than one form provide the linking data. The differences observed between within-form item calibrations and the model requirement that each item be characterized by a single difficulty, regardless of form or sample, estimate the relative difficulty of each form. This form difficulty is then added to the within-form item calibrations to place every administration of every item onto one common bank scale.

CALIBRATING FORMS ON THE BANK

To estimate the shift in difficulty between two forms, *k* and *j*, a weighted average of difficulty differences is calculated for the items linking them

$$t_{kj} = \frac{\sum_{i}^{n} [d_{ij} - d_{ik}] w_{ikj}}{\sum_{i}^{n} w_{ikj}} \qquad 13.5$$

where d_{ik} and d_{ij} are the estimated difficulties of linking item *i* in forms *k* and *j*, *n* is the number of items in this link, and $w_{ikj} = 1/(se_{ik}^2 + se_{ij}^2)$ is an information weight based on the item calibration standard errors, se_{ik}, and se_{ij}. The standard error of the difficulty shift t_{kj} is

$$se_{kj} = 1 / \left[\sum_{i}^{n} w_{ikj} \right]^{1/2} . \qquad 13.6$$

The shift t_{kj} estimates the difference in origins of forms *k* and *j*. A shift is calculated for every pair of forms linked by common items. When every possible pair of forms is linked, then the difficulty T_k of form *k* is the average shift for form *k* over all forms.

$$T_k = \frac{\sum_{j}^{M} t_{kj}}{M} \qquad 13.7$$

where *M* is the number of forms and $t_{kk} = 0$. The standard error of form difficulty T_k is

$$se_k = \left(\sum_{j}^{M} se_{kj}{}^2 \right)^{1/2} / M \qquad 13.8$$

Equations 13.5 through *13.8* assume every form is linked to every other form. When links are missing between some forms, as is usually the case, an iterative procedure can be used to bridge the empty cells. Empty cells can be started at

$$t_{kj} = t_{jk} = 0 \qquad 13.9$$

and the form calibrations T_k improved step-by-step by calculating temporary form difficulties with *Equation 13.7*, adjusting empty cells to

$$t_{kj} = T_k - T_j \text{ and } t_{jk} = T_j - T_k \qquad 13.10$$

and then reapplying *Equation 13.7* iteratively until the successive values of T_k stabilize. This process works as long as every form can be reached from every other form by some chain of links.

This procedure sets the bank origin at the center of all forms so that form difficulty T_k is the difference between the center of form *k* and the center of the bank.

ITEM WITHIN-LINK FIT ANALYSIS

To verify the extent to which the linking items perform adequately within their forms combine the item-within-form fit statistics of *Equation 13.3* into a within-form fit statistic for the link.

$$\text{Within form link fit } = \frac{\sum_{i}^{n}(V_{ik} + V_{ji})}{2n} \qquad 13.11$$

where V_{ik} is the fit of item *i* in form *k*

V_{ij} is the fit of item *i* in form *j*, and

n is the number of items in the link.

This statistic estimates one with variance of order $[1/n(N-1)]$ when the link items fit within forms.

ITEM BETWEEN-LINK FIT ANALYSIS

To check the extent to which link items agree on the relative difficulties of their two forms, calculate the ratio of observed to model variance.

$$\text{Between form link fit} = \frac{\sum_{i}^{n}(d'_{ik} - d'_{ij})^2}{\sum_{i}^{n} w_{ikj}} \qquad 13.12$$

where now $w_{ikj} = [se_{ik}^2 + se_{ij}^2]$ and the within form item difficulties, d_{ik} have been translated to their bank values d'_{ik} by

$$d'_{ik} = d_{ik} + T_k \qquad 13.13$$

Values substantially greater than one, given expected variance $[2/(n-1)]$, signify that some items operate differently in the two forms. A plot of d_{ik} versus d_{ij} over i facilitates the evaluation of link status and the identification of aberrant items (see Wright & Stone, 1979, pp. 92-95; Wright & Masters, 1982, pp. 114-117).

LINK WITHIN-BANK FIT ANALYSIS

To check the extent of agreement among links with respect to form difficulties review the extent to which each entry in the matrix of observed shifts between forms is close to the difference in bank difficulties of the forms. To evaluate whether a link fits the bank, calculate the link residual

$$y_{kj} = t_{kj} - [T_k - T_j] \qquad 13.14$$

where t_{kj} is the observed shift between forms *k* and *j*, and T_k and T_j are their bank difficulties.

These link residuals can be standardized to mean zero and variance one by dividing them by the standard errors, se_{kj} of their t_{kj} of *Equation 13.5* and multiplying by $[M/(M-2)]^{1/2}$ where *M* is the number of forms in the linking network.

FORM WITHIN-BANK FIT ANALYSIS

To check the fit of each form to the bank as a whole calculate

$$V_k = \frac{\sum_{i}^{L} [y_{kj} / se_{kj}]^2 [M/(M-2)]}{L-1} \qquad 13.15$$

where *L* is the number of t_{kj} observed for form *k*.

The criterion value of V_k is also one, this time with variance of order [*2 / (L* - 1 *)*].

The fit of a link or a form into the bank is related to how well linking items fit within their own forms. When the number of students taking a form is large, the item fit statistic variances can become unrealistically small and must be taken with a grain of salt. Careful investigation of doubtful items is always instructive and invariably leads to insight into the nature of the variable. The misfit of links within the bank is usually associated with particular forms. This can occur when a form is inadvertently administered to a sample of students for whom it is inappropriate. The best items for estimating form difficulties are those that satisfy the various fit analyses.

REVIEWING THE RESULTING BANK

At this point an ITEM LIST (Figure 13.1) which gives each item in the bank by sequence number, legitimate alternatives, correct responses, item name, bank difficulty, between difficulty root mean square, and within form fit mean square is useful.

Bank difficulty is the average of the item's difficulties in the forms in which it was calibrated, adjusted for these forms' local difficulties. A between difficulty root mean square, the square root of the average squared difference between an item's bank equated difficulties in each form and its bank difficulty is useful to tag potentially errant items. Items showing between difficulty root mean squares greater than 0.5 logits are frequently found to have been miskeyed or misprinted in one of the forms in which they appear.

The within-form item fit mean square of *Equation 13.3* can be standardized to mean zero and variance one so that the average square of these standardized within-form fits can summarize item performances within forms. Its sign is taken from the sign of the standardized fit with the largest absolute value to distinguish between misfit caused by unexpected disorder, indicated by large positive standardized fits, and misfit caused by unexpected within-form inter-item dependence, indicated by large negative standardized fits. It is useful to tag items producing values greater than 2 or less than -2 for further examination.

An ITEM MAP (Figure 13.1) which displays the variable graphically by plotting the items according to their bank difficulties along the line of the variable which they define, will enable teachers to examine the relationship between the content of the items and their bank difficulties in order to review the extent to which the empirical item order defines a curriculum strand that agrees with their curriculum expectations and so has construct validity for them. The item map provides a framework for writing new items to fill gaps that appear in the definition of the curriculum strand and for choosing items for new tests.

A FORM LIST (Figure 13.1) which gives each form by form number, name, number of items and bank difficulty is useful. Each item is listed by form position, item name, key, within form difficulty and standard error, total within form standardized fit, and bank difficulty. This facilitates the review of each form as a whole and the identification of form specific anomalies.

A KID LIST (Figure 13.1) which gives each student by identification, ability measure, error and fit statistic indicates which students misfit by displaying their response string and its residuals from expectation, so that teacher and student see the specific item sources of misfit.

A KID MAP (Figure 13.1) produces a graphical representation of each individual student's performance. The map for each student shows where that student and the items they took stand on the curriculum strand, which items were answered correctly, the probability of each response, and the student's percent mastery at each item. This provides teacher, student and parent with a picture of the student's performance which combines in one easy to read picture specification of criteria mastery with the identification of unexpected strengths and weaknesses.

ITEM QUALITY CONTROL

Once items have been banked, the identification and study of misfitting items follows. The irregularities most often identified are mechanical and clerical such as miskeying, misprinting, misscoring, more than one right answer and no right answer. Sometimes, however, item misfit brings out anomalies in student performance which leads to new and unexpected understanding of how the subject matter contained in the item is learned and used.

The item infit and outfit mean square fit statistics of *Equations 13.3* and *13.4* indicate the degree to which an item functions as intended. Mean square statistics greater than 1.4 imply noise in item use, outbreaks of guessing or carelessness, or the presence of secondary variables correlated negatively with the intended variable. Mean square statistics less than 0.6 imply inter-item dependencies or the presence of secondary variables correlated positively with the intended variable.

MISFIT PATTERNS

Miskeying and scanner errors usually cause an item to appear more difficult than anticipated, making item fit too large.

Misfit caused by student behavior, such as guessing and carelessness, is not diagnosed well by item fit statistics because item statistics lump together students behaving differently. Disturbances that are the consequences of individual student behavior are best detected and best dealt with through the fit analysis of individual students (Wright & Stone, 1979, Chapters 4 & 7). But item statistics can call attention to items that tend to provoke irregular behavior in many students.

Guessing is a problem only when students inclined to guess are also provoked to guess on items that are too difficult for them and then only when those particular students happen to guess correctly. This is more probable for low ability students but may occur for others depending upon the value given to an outcome or success on the test and the time allowed. Problems of guessing are best addressed by targeting test administration so that it does not provoke guessing by allowing enough test time so that students are not rushed and by reviewing each student's response pattern for the presence of improbable right answers which might have been achieved by lucky guessing.

Carelessness occurs when a high ability student fails an easy item. The pattern in item statistics is low difficulty and high fit. This, too, is most usefully and accurately detected through the identification of improbable wrong answers in individualized person fit analyses.

OTHER SOURCES

When the disturbance in a misfitting item is not mechanical or clerical, the cause is usually special knowledge. Interactions with curriculum specifics affect the shape of the response curve. Dependence on a skill that only high-ability students are taught can make an item unfairly easier for these high ability students. This will cause the item to have a fit statistic that is improbably low and an unusually high point biserial. On the other hand, dependence on a skill that is negatively related to instruction, so that low-ability students possess more of it, can make an item unfairly easier for low-ability students and, hence, give it a fit statistic that is improbably high. Either way, the interaction disqualifies the item for use with students who are unequal in their exposure to the special skill. When fit is too high, the item is unfair to more able students. When fit is too low, the item is unfair to less able students.

One-step implementation of an item bank can be done using a computer program like *BIGSTEPS* (Wright & Linacre, 1997). But the data layout must be organized so that the separate forms flow into one standard frame of reference. An integrated item banking system like *SAMS* (Wright, Linacre & Schultz, 1991) can be used for general school applications.

14. VARIABLE MAPPING

This primer shows how variable mapping is fundamental to measurement. Mapping is where test development begins (as an idea) and mapping represents its realization (through empirical validation) and its actionable interpretation.

The map of the variable begins as the blueprint for the design of a test. When the map is well conceived, its test design implementation will be a straightforward representation in ordered items. Later, as the map is empirically verified by candidates' responses, it results in the successful implementation and realization of an idea. The map of the variable pictures the idea and its realization.

Alfred Binet is better known for the American versions of his tests than for his ideas on test construction and measurement. But Binet's work in test development represents an excellent example of what is implied in mapping. The first (1905) edition of his test consisted of thirty items arranged in difficulty order. This item arrangement enabled measuring ability by locating a child "along" the ordered item scale.

Gould (1981) faults Binet for a "hodgepodge of diverse activities" in item selection. But, however diverse, Binet's items were not haphazard.

> First of all, it will be noticed that our tests are well arranged in a *real order of increasing difficulty.* It is *as the result of many trials,* that we have established this order; we have by no means imagined that which we present. If we had left the field clear to our conjectures, we should certainly not have admitted that it required the space of time comprised between four and seven years, for a child to learn to repeat 5 figures in place of 3. Likewise we should never have believed that it is only at ten years that the majority of children are able to repeat the names of the months in correct order without forgetting any; or that it is only at ten years that a child recognized all the pieces of our money (Binet, 1905, p. 185).

Binet relied upon "numerous" replications of his ordered items to give him the measurement accuracy he desired.

> "One might almost say, "It matters very little what the tests [items] are, so long as they are numerous'" (1911, p. 329).

Furthermore, Binet writes that his scale:

> "properly speaking, does not permit the measure of intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured" (1905, p. 40).

In recognizing this deficiency in his test, Binet shows that he knew that linearity was necessary for measurement.

Binet's essential ingredients for the construction of measurement were:

1. Item arrangement by difficulty order
2. Numerous items to insure precision
3. Need for, but recognized difficulty in producing, linear measures.

How else could one build a test? There is no other way except to begin as Binet did: with an idea for a variable illustrated by items arranged by intended difficulty, and measures of persons according to their locations among the items along the variable. Every attempt at test construction is made along such lines whether successful or not.

The hallmark of Binet's efforts is his attempt to benchmark items and persons. The idea of a benchmarked line of increasing amounts is fundamental in constructing a variable, and the map of intentions is the blueprint for item selection and/or item construction. Its realization gives us a picture of the variable and a means for seeing the locations of items and persons along the variable.

A contemporary example is WRAT3 (Wilkinson, 1993). This test implements the achievement measures: (1) word naming, (2) arithmetic computation and (3) spelling from dictation. The item arrangement is developmental and indicates the sequence of instruction and learning. The items for each scale proceed from items at the most elementary levels to those of increasing difficulty at higher levels.

Arrangement of items in difficulty order is exactly what we want in any measuring tool. The locations of items along the variable are determined by teacher judgment, curriculum, and learning experts. Validation is rendered by subjecting the initial item arrangement to empirical testing. When the arrangement of items is sensible it will be supported by data gathered from students' response to these items.

The arrangement of items should also correspond to the arrangement of persons. Less able persons should be located below more able persons. A hierarchical correspondence between items and persons will show easy items in company with less able persons, more difficult items associated with more able persons.

Determining the calibrations of items and the measures for persons will either substantiate the original item placement or suggest revisions. This leads to a continuous dialogue between the idea for the variable and its data. A good initial plan generally results in fewer cycles between idea and data before an acceptable definition of the intended variable is achieved.

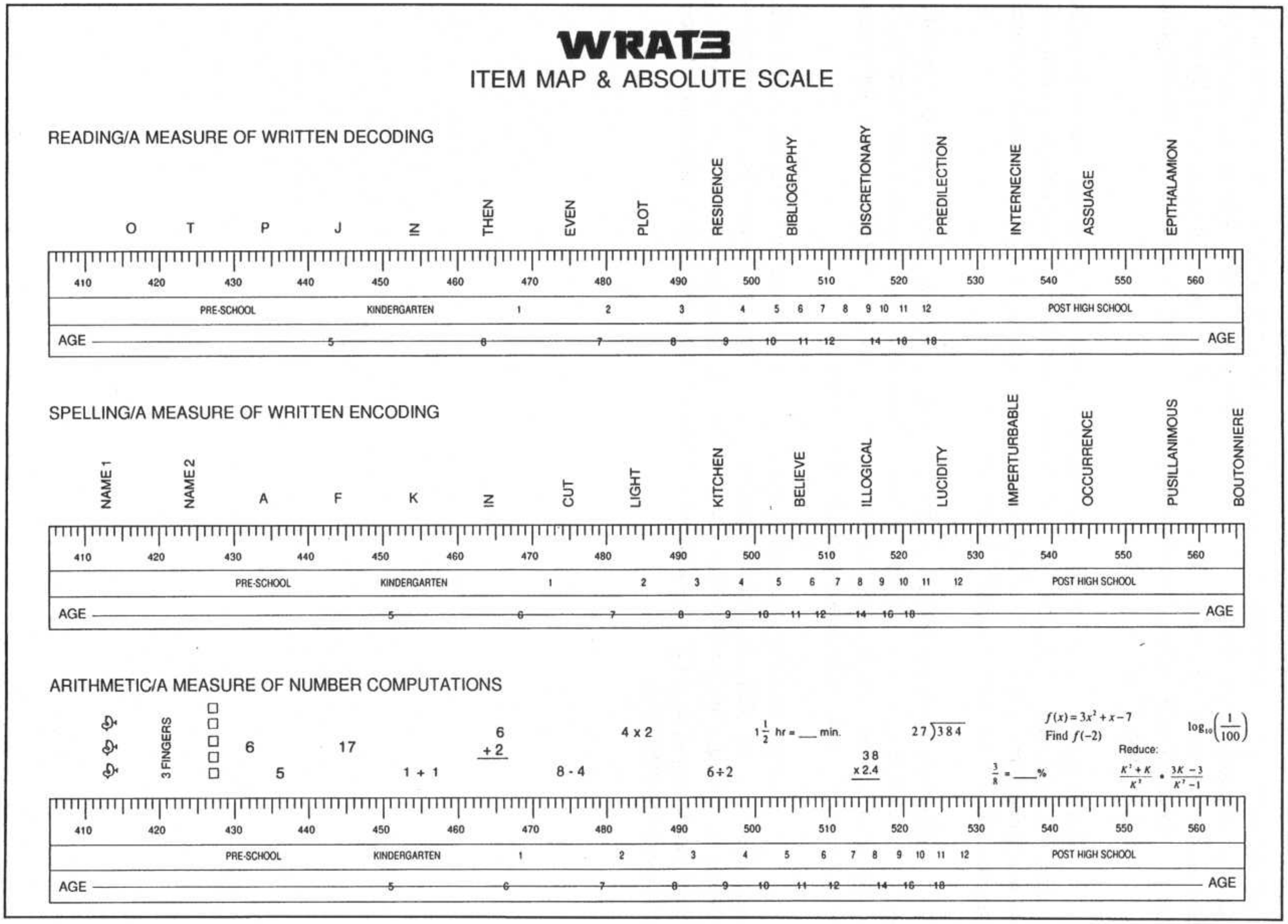
Successful item calibration and person measurement leads to a map of the variable. The resulting map is no less a "ruler" than one constructed for measuring length. It can be applied in a similar way to produce measures as useful as those of any yardstick.

Figure 14.1 is a map of the WRAT3 variables for word reading, arithmetic computation and spelling from dictation. It proceeds from left to right in a progressive order of difficulty for items and ability for persons.

The map of each variable gives sample items showing their progressive difficulty. Below the items is an absolute equal interval scale providing measures. The location of average grade and age are also given.

Figure 14.1

Variable map.



WRAT3
ITEM MAP & ABSOLUTE SCALE
READING/A MEASURE OF WRITTEN DECODING
O T P J IN THEN EVEN PLOT RESIDENCE BIBLIOGRAPHY DISCRETIONARY PREDILECTION INTERNECINE ASSUAGE EPITHALAMION
410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560
PRE-SCHOOL KINDERGARTEN 1 2 3 4 5 6 7 8 9 10 11 12 POST HIGH SCHOOL
AGE 5 6 7 8 9 10 11 12 14 16 18 AGE
SPELLING/A MEASURE OF WRITTEN ENCODING
NAME 1 NAME 2 A F K IN CUT LIGHT KITCHEN BELIEVE ILLOGICAL LUCIDITY IMPERTURBABLE OCCURRENCE PUSILLANIMOUS BOUTONNIERE
410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560
PRE-SCHOOL KINDERGARTEN 1 2 3 4 5 6 7 8 9 10 11 12 POST HIGH SCHOOL
AGE 5 6 7 8 9 10 11 12 14 16 18 AGE
ARITHMETIC/A MEASURE OF NUMBER COMPUTATIONS
3 FINGERS
□ □ □ □ □ □
6 5 17 1 + 1
6 +2
8 - 4
4 x 2
6÷2
$1\frac{1}{2}$ hr = ___ min.
38 x 2.4
27)384
$\frac{3}{8}$ = ___%
$f(x) = 3x^2 + x - 7$ Find $f(-2)$
Reduce: $\frac{K^2+K}{K^2} \cdot \frac{3K-3}{K^2-1}$
$\log_{10}\left(\frac{1}{100}\right)$
410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560
PRE-SCHOOL KINDERGARTEN 1 2 3 4 5 6 7 8 9 10 11 12 POST HIGH SCHOOL
AGE 5 6 7 8 9 10 11 12 14 16 18 AGE

The map has immediate appeal and application. Like the marks of increasing height of a child on the door jamb, the map can show student progress on these three scales. Especially helpful is the overall view that the map provides, giving a sense of order and coverage to the entire variable of interest. The map shows the order implied in the variable and it can be used to show the location and subsequent progress of pupils along the ruler. The absolute scale gives values useful in data analysis. The grade and age norms show the progress we expect to see at increasing grade and age levels. The wider spacing observed on the left of the ruler compared to the right indicates the accelerated growth occurring among younger children.

Gathering data on all three scales across persons allows comparisons to be made between scales. We notice that successfully reading the word "residence" corresponds in general to spelling the word "kitchen" from dictation and computing the problem "6-2= ." These comparisons enable further diagnosis and make the map a useful diagnostic tool.

Although each ruler appears orderly, that order is in need of continual reappraisal and revalidation. Variable definitions are never finished. While there is consistency and order to the WRAT3 scales, a fact demonstrated by five successive editions, it remains necessary to monitor continuously the variable in order to keep the map coherent and up-to-date.

Continuous monitoring is required for any variable. Concerns for reliability and validity do not rest in historical coefficients, but in continuing successful demonstrations that can be referenced by test consumers in order to determine the extent to which the test is relevant to their intended application. Such indications of applicability must be continuously provided in order to maintain the variable map and assure its relevancy.

15. ESTIMATING ITEM CALIBRATIONS AND PERSON MEASURES

INTRODUCTION

In this chapter we will work through a mathematical approach to the estimation of Rasch model item and person parameters (Rasch, 1960). This approach is especially suited to computer implementation and most of the computer programs in use employ versions of the algorithms to be described. The procedure is called UCON, for unconditional maximum likelihood estimation (MLE) (Wright & Panchapakasan, 1969; Wright & Douglas, 1977a; Wright & Stone, 1979; Wright, 1980). The term "unconditional" is used because there is another fully conditional maximum likelihood estimation (FCON) which uses conditional probabilities to estimate item difficulties directly without involving any simultaneous estimation of person abilities (Wright, 1968, 1980; Wright & Douglas, 1977b). FCON has desirable theoretical properties, but it is difficult to implement when there are more than a few items. UCON, on the other hand, approximates the results of FCON closely—and UCON seldom has any trouble giving useful results.

Although calibration of item difficulties is the first stage in the implementation of the model, and, in principle, precedes the measurement of persons, it is convenient to estimate item difficulties and person abilities simultaneously. The analysis of fit is expedited by the computation of expected responses of persons to items so that these expected responses can be compared with the observed responses. These expected responses can be determined most easily when we have simultaneous estimates of item difficulties and person abilities.

The estimation of statistical model parameters is the fundamental step of applied statistics. When we view calibration as a problem in statistical estimation, the question arises as to which estimation procedure to use. There are many estimation procedures: least squares, mean value, minimum chi-square, maximum likelihood. The last procedure, MLE, developed by Ronald Fisher in the 1920's, has a number of useful properties. The Rasch model lends itself to MLE and the useful properties of MLE translate into substantive fundamentals of measurement.

RASCH MLE PROCEDURES

Once a statistical model is specified, an equation for the probability of occurrence of any observation follows. From this equation, the joint probability of any data set may also be specified and this equation used to answer the question: What is the probability that this particular set of data occurred when this set of items was given to this group of persons? This joint probability is known as the likelihood of the data. It is a function of the observed data and also of the initially unknown but soon to be estimated parameters of the model (the item difficulties and person abilities). The MLE principle is to select for the estimates of the parameters that particular set of values which makes the likelihood of the data in hand as large as possible - a maximum.

The likelihood of the data is viewed as a function of known data and unknown parameters. The parameters become the variables. Calculus is employed to find the particular values of these unknown parameters that make the likelihood of these data a maximum. This is done by taking the derivative of the likelihood with respect to each unknown variable and setting this derivative equal to zero. This produces equations which may be solved for the unknown values, which, when obtained, make the likelihood of these data as large as it can get.

To review, we:

1) derive an expression for the likelihood of the data,

2) differentiate this expression with respect to each of the unknown parameters,

3) set each result equal to zero and

4) solve the resulting set of equations for the ML item difficulty and person ability estimates.

When the Rasch model is applied to test data there are a large number of unknown parameters to be estimated, many more than the one or two involved in the usual maximization problem. Nevertheless, the principles are the same and when the procedure is applied step-by-step to one item and then one person at a time no complications arise.

Usually when we solve equations for an unknown value in algebra, arithmetic operations like addition and division are sufficient to obtain an explicit solution. The equation $5X+6=20$, for example, requires one subtraction and one division to reach the exact solution of $X = 2.8$. Since this kind of equation can be solved by a finite number of simple arithmetical steps, it is called explicit.

In contrast, an equation like $X + 2 * \sin X = .73$ does not lend itself to simple arithmetic. To solve this "implicit" equation we must resort to another method. A good way to solve this kind of implicit equation was invented by Isaac Newton in the 1680's.

Newton's method:

1) a reasonable guess is provided for the unknown value of X,

2) the "closeness" of this guess to the best solution is determined by noting how much remains when this value for X is substituted in the equation,

3) the difference between the initial value for X and the remainder is then used to determine a next "better" value for X,

4) this process for improving the estimate of X continues until the remainder gets small. How small is left to the discretion of the person solving the equation.

Each step in this process is called an iteration. The iterative process will converge to a solution for a large class of implicit equations, among which are equations incorporating the exponential function $\exp(X)$. All that are needed to implement Newton's method are the derivatives of the equations to be solved and good initial guesses. For the Rasch model equations, there are very sensible initial guesses for the unknown item difficulties and person abilities.

If we let $f(x) = 0$ be the equation to be solved for the unknown X, and, if $f'(X)$ is its derivative with respect to X and, if X_0 is the initial guess for the value of X, then Newton's method specifies the next better value for X as

$$X_1 = X_0 - \frac{f(X_0)}{f'(X_0)} \tag{15.1}$$

where $f(x_o)$ and $f'(x_o)$ are values of these functions when we substitute the initial value x_o for X and x_t is the new, improved value for X at the end of the first iteration.

We can write a general expression for this relation which shows the value of x_t at the end of t iterations in terms of what it was on the previous iteration:

$$X_t = X_{t-1} - \frac{f(X_{t-1})}{f'(X_{t-1})} \tag{15.2}$$

Since we may continue iterating until our result is as accurate as we wish, when should we stop when estimating parameters for a Rasch model? Experience has shown that when reporting values for item difficulties and person abilities we never need accuracy greater than two decimal places. Enough accuracy is obtained when we settle for an x_t which makes the absolute difference between that x_t and its previous value x_{t-1} in the vicinity of 0.005, that is, "correct" to the second decimal place.

MAXIMUMLIKELIHOODESTIMATION

The Rasch probability of any observation x_{ni} for person n on item i is

$$P(X_{ni}|B_n,\ D_i) = P_{ni} = [\exp\ X_{ni}(B_n - D_i)] / [1 + \exp(B_n - D_i) \tag{15.3}$$

where x_{ni} is the observed data, and may be either 0 or 1,

B_n is the unknown person ability measure and

D_i is the unknown item difficulty calibration.

For a test of L items given to N persons for whom it is reasonable to think of the persons and items as functioning independently i.e. as specified by *Equation 15.3*, the joint probability (the likelihood) of all the data is found by multiplying together all N by L probabilities of the type in *Equation 15.3*.

The expression $(A**m)*(A**n)*(A**q)$ may be written with the single base A and an exponent which is the sum of the three exponents, $A**(m+n+q)$. When this notation is applied to the N x L exponents of the likelihood function, we have

$$\wedge = \prod_{i}^{L}\prod_{n}^{N} P_{ni}$$

$$= \prod_{i=1}^{L}\prod_{n=1}^{N}\left[\frac{\exp\left[X_{ni}(B_n - D_i)\right]}{1+\exp(B_n - D_i)}\right]$$

$$\left[\frac{\exp\left[\sum_{i}^{L}\sum_{n}^{N}\right](X_{ni}B_n - X_{ni}D_i)}{\prod_{i}^{L}\prod_{n}^{N}\left[1+\exp(B_n - D_i)\right]}\right] \qquad 15.4$$

where is the likelihood of the data, $\prod_{i}^{L}\prod_{n}^{N}$ is the continued product over n and i of all $N * L$ probabilities P_{ni} and $\sum_{i}^{L}\sum_{n}^{N}$ is the continued sum over n and i of all $N * L$ exponents $(X_{ni}B_n - X_{ni}D_i)$.

The double summation in the numerator can be distributed over the two terms with the result

$$\wedge = \left[\frac{\exp\left[\sum_{n}^{N} B_n \sum_{i}^{L}\right]X_{ni} - \sum_{i}^{L} D_i \sum_{n}^{N} X_{ni}}{\prod_{i}^{L}\prod_{n}^{N}\left[1+\exp(B_n - D_i)\right]}\right]$$

$$= \frac{\exp\left[\sum_{n}^{N} B_n R_n - \sum_{i}^{L} D_i S_i\right]}{\prod_{i}^{L}\prod_{n}^{N}\left[1+\exp(B_n - D_i)\right]} \qquad 15.5$$

where $\sum_{i}^{L} X_{ni} = R_n$ is the right answer count or the raw test score for person n,

and $\sum_{n}^{N} X_{ni} = S_i$ is the right answer count or raw sample score for item i.

With the likelihood in this form we see that the statistics required are not the separate person-to-item responses but only their accumulations into the person scores R_n and the item scores S_i. Further, the R_n's and S_i's are separated from each other. Each set multiplies its own parameters B_n's and D_i's in turn. This separation is the defining characteristic of a Fisher "sufficient" statistic (Fisher, 1958) and also the algebraic requirement for Rasch objectivity.

Although R_n and S_i are sufficient to estimate B_n and D_i these scores themselves are not satisfactory as measures. Person score is not free from the particular item difficulties encountered in the test. Nor is item score S_i free from the ability distribution of the persons who happen to be taking the item. Independence from these local factors requires adjusting the observed R_n and S_i for the item difficulty and person ability distributions they depend on. This adjustment is necessary to produce the test-free person measures and sample-free item calibrations we desire.

In order to obtain the maximum of this likelihood with respect to possible values of the unknown parameters, the likelihood needs to be differentiated with respect to the *B*'s and *D*'s in turn. This task is easier when we take the logarithm of the likelihood. We can do that because the values which make the logarithm of a function a maximum also make that function a maximum.

Since $\log(\exp X) = X$, the numerator of the log likelihood becomes simple. The denominator turns into a subtraction and the double product becomes a double sum of $\log\left[1+\exp(B_n - D_i)\right]$.

Thus $K = \log \wedge = \sum_n^N B_n R_n - \sum_i^L D_i S_i - \sum_i^L \sum_n^N \log\left[1+\exp(B_n - D_i)\right]$ is the log-likelihood. *15.6*

Since the derivative of the exponential function, exp *X*, reproduces itself and the derivative of the logarithmic function, log *Y*, is 1/*Y*, the differentials required to produce solutions for $\partial K / \partial B$ and $\partial K / \partial D$ are

$$\partial\left(D_i S_i\right) / \partial D_i = S$$

$$\partial\left(B_n R_n\right) / \partial B_n = R_{ni}$$

$$\frac{\partial \log\left[1+\exp(B_n - D_i)\right]}{\partial B_n} = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} = P_{ni}$$

$$\frac{\partial \log\left[1+\exp(B_n - D_i)\right]}{\partial D_i} = -\frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} = -P_{ni} \qquad 15.7$$

By differentiating the log likelihood *K* with respect to each D_i and then, separately, with respect to each B_n and equating each of these derivatives to zero to locate maxima, we obtain the two sets of equations.

$$\frac{\partial K}{\partial D_i} = -S_i + \sum_n^N \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)}$$

$$= -S_i + \sum_n^N P_{ni} = 0 \text{ for each } i = 1, L$$

$$\frac{\partial K}{\partial B_n} = +R_n - \sum_i^L \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)}$$

$$= +R_n - \sum_i^L P_{ni} = 0 \text{ for each } n = 1, N \qquad 15.8$$

Each of the first L equations contains N unknown B_n's and one unknown D_i. Each of the second N equations contains L unknown D_i's and one unknown B_n .

Newton's method uses the derivative of the equation to be solved, therefore we need to take the derivatives of the above implicit equations with respect to D_i and B_n once again in order to solve them by Newton's method. These derivatives are the second derivatives of the likelihood.

Since $$P_{ni} = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} = \frac{1}{1+\exp(D_i - B_n)} \qquad 15.9$$

the differentials needed to find the second derivatives of K with respect to B_n and D_i are

$$\frac{\partial K}{\partial D_i} = -S_i + \sum_n^N P_{ni}$$

$$\frac{\partial^2 K}{\partial D_i^2} = -\sum_n^N P_{ni}(1 - P_{ni}) = -\sum_n^N Q_{ni} \text{ for } i = 1, L$$

$$\frac{\partial K}{\partial B_n} = R_n - \sum_i^L P_{ni}$$

$$\frac{\partial^2 K}{\partial B_n^2} = -\sum_i^L P_{ni}(1 - P_{ni}) = -\sum_i^L Q_{ni}$$

$$\text{for } n = 1, N \text{ where } Q_{ni} = P_{ni}(1 - P_{ni}) \qquad 15.10$$

These second derivatives are the product of P_{ni} and its complement (1-P_{ni}) combined in $Q_{ni} = P_{ni}(1 - P_{ni})$ where P_{ni} is the probability that person *n* gets item *i* correct.

Since $Q_{ni} \geq 0$, these second derivatives are always negative. This tells us that the solutions to *Equation 15.8* must be maxima.

Before we apply Newton's method to solve these equations, three uncertainties need to be resolved.

1. What shall we use for initial values of the estimates? Although Newton's method is usually robust with respect to the choice of an initial estimate (meaning we will get to the same final estimate no matter where we start), we will get convergence most rapidly if we use initial estimates which are not far from the final estimates.

   We can do this for items, by approximating the abilities of all persons at zero.

Then the MLE's of *Equation 15.8* have the explicit solution:

$$\sum_{n}^{N} P_{ni} = \sum_{n}^{N} \frac{\exp(-D_i)}{1+\exp(-D_i)} = \frac{N\exp(-D_i)}{1+\exp(-D_i)}$$

so $$-S_i + \sum_{n}^{N} P_{ni} = -S_i + \frac{N\exp(-D_i)}{1+\exp(-D_i)} = 0$$

$$\exp(-D_i) = S_i / (N - S_i)$$

and $$D_i = -\log\left[\frac{S_i}{N-S_i}\right] = +\log\left[\frac{N-S_i}{S_i}\right] \text{ for } i = 1, L \qquad 15.11$$

This initial estimate is a simple logarithmic transformation (the logit) of the item scores.

By approximating the difficulties of all items at zero in the equations for the B's in *Equation 15.8*, we find a similar explicit solution for B_n as a simple logarithmic transformation of the raw scores for person n

$$B_n = \log\left[\frac{R_n}{L-R_n}\right] \text{ for } n = 1, N \qquad 15.12$$

2. Although it may appear that the equations in 15.8 have N unknowns $B_1, B_2, \ldots, B_N$ only the statistics $R_1, R_2, \ldots, R_{L-1}$ are available to estimate them. When data is complete the values of B_n which can be estimated from a test of L items may therefore be indexed by R rather than by n. Indexing persons by their raw scores highlights the fact that a raw score for a person is the sufficient statistic for estimating that person's ability.

In general, there will be more than one person with a given raw score. Since as far as ability estimation is concerned, we are unable to distinguish among persons who took the same items and earned the same raw score. We may group persons who took the same items according to their raw score. If we let N_r be the number of persons who scored R on the test, we may rewrite *Equations 15.8* and *15.10* as

$$-S_i + \sum_{R=1}^{L-1} N_R P_{Ri} = 0 \qquad 15.13$$

and $$\frac{\partial}{\partial D_i}\left(\frac{\partial K}{\partial D_i}\right) = -\sum_{R}^{L-1} N_R P_{Ri}(1 - P_{Ri})$$

$$= -\sum_{R}^{L-1} N_R Q_{Ri} \qquad 15.14$$

and *Equation 15.12* as
$$B_R = \log\left[\frac{R}{L-R}\right] \qquad 15.14$$

$$R=1,\ L\text{-}1$$

where
$$P_{Ri} = \exp(B_R - D_i)/\left[1+\exp(B_R - D_i)\right]$$

and
$$Q_{Ri} = P_{Ri}(1-P_{Ri})$$

3. Were we to apply Newton's method to these equations as they stand, we would find that the iteration process would not converge. This is because our set of equations contains one too many unknowns to be uniquely estimated.

The Rasch model specifies the probability of a response by a person to an item as a function of the difference between their locations on a variable.

The probability that a person with ability B_n gets an item with difficulty D_i correct, is exactly the same as the probability for a person with ability $(B_n + 3)$, say, responding to an item with difficulty $(D_i + 3)$, because $(B_n + 3) - (D_i + 3) = B_n - D_i$. Since our choice of 3 was arbitrary, we see that an infinite set of B's and D's will satisfy our equations providing only that they maintain their differences $(B_n - D_i)$.

This problem of too many unknowns can be overcome by placing one restriction on the set of B_n's and D_i's. The particulars of this restriction are not important algebraically. We could set any person, say B_1, equal to a constant or any item, say D_3, equal to some other constant. Any constant will do. We have found it convenient for calibration to use the restriction that the sum of our set of estimated item difficulties $\sum_i^L D_i \equiv 0$ be zero. This centering on the test has the effect of reducing our unknowns from $(L-1)+L = 2L-1$ to $(L-1)+(L-1) = 2L-2$.

In order to maintain the possibility of convergence, we must implement this restriction each time we derive an improved set of (D_i) values. Centering is accomplished by finding the mean of the current estimates of the D_i's and subtracting this mean from each D_i. This is done at each iteration.

Thus the initial centered set of D_i's are

$$D_i = \log\left[\frac{N-S_i}{S_i}\right] - \sum_i^L \left\{\log\left[\frac{N-S_i}{S_i}\right]\right\} / L \qquad 15.16$$

SOLVING THE MAXIMUM LIKELIHOOD EQUATIONS

Here is a systematic procedure for solving these equations and hence obtaining estimates of item difficulty and person ability (Once all perfect and zero scores have been removed from the data matrix).

1. Determine the initial item estimates from *Equation 15.16.* Items are centered.

2. Determine the initial person estimates from *Equation 15.15*. Persons do not need to be centered. In fact, they must not be.

3. Using all person estimates and the current estimate for each item i, apply Newton's method to *Equation 15.13* until differences between successive estimates of each D_i that is, $(D_i' - D_i)$ are less than, say, .005 logits. The process is,

$$D_i' = D_i - \frac{S_i - \sum_R^{L-1} N_R P_{Ri}}{\sum_R^{L-1} N_R Q_{Ri}} \qquad 15.17$$

   in which D_i is the current estimate and D_i is the next improved estimate. The successive differences are $(D_i' - D_i)$.

4. Repeat step 3 for all items, $i = 1, L$. When we have finished, we have a new and better set of D_i estimates.

5. Center these new D estimates.

$$D. = \sum_i^L D_i / L \qquad\qquad D_i' = D_i - D. \qquad 15.18$$

6. Using these new centered D_i estimates and the person estimate for a score of $r = 1$ (that is, B_1), apply Newton's method to *Equation 15.10* expressed in terms of r instead of n.

$$r - \sum_i^L P_{ri} = 0 \qquad 15.19$$

   until differences between successive estimates of B_r that is $(B_r' - B_r)$ are less than, say, .005 logits.

   The process is,

$$B_r' = B_r + \frac{r - \sum_i^L P_{ri}}{\sum_i^L Q_{ri}} \qquad 15.20$$

in which B_r is the current estimate and B_r' is the improved estimate. The successive differences are $(B_r' - B_r)$.

7. Repeat step 6 for the second, then third, etc., raw score. When we have reached $r = L - 1$ we have a new and better set of B_r estimates. Do *not* center these B_r.

8. At this stage we have reached the end of the first major "loop". This loop comprised L minor loops on the items and L-1 minor loops on the person scores.

At the end of each major loop we determine whether the likelihood has been sufficiently maximized by reviewing our convergence criterion for all 2L-1 estimates. Since it is unlikely that satisfactory convergence will have been achieved in one major loop, we proceed to additional major loops.

9. Using the latest person estimates and the current value for D_1, apply Newton's method as in Step 3 until convergence. Repeat for all items.

10. Center the latest set of D_i's.

11. Using these latest centered D_i and the current value for B_1, apply Newton's method, Step 6, until convergence. Repeat for all raw scores from 2 to L-1.

12. Determine whether a satisfactory overall convergence has been obtained at the end of this second major loop and so on.

This estimation procedure usually converges to a criterion of .005 logits in 5 or 6 major loops. There are rare circumstances in which an MLE cannot be obtained. When there are one or two items or persons separated from the nucleus of the data by many logits, then round-off problems can occur with the procedure outlined above. When this procedure fails it is almost always due to inaccurate editing of the original data or to failure to center items each time a new set of estimates is produced.

Because of the way estimates are calculated in UCON there is a slight bias. This bias can be corrected by shrinking all values of B and D by the factor (L-1)/L (Wright, 1988).

STANDARD ERRORS OF ESTIMATES

A key benefit of a good estimation procedure is the simultaneous estimation of standard errors for its estimates. These standard errors specify the modeled degree of precision (reliability) with which the estimates can be obtained.

A familiar example of this is estimating a mean from a random sample of N observations. The sample mean is, in many ways, a "best" estimate of the location parameter of the distribution from which the random sample was drawn. The standard error of a mean is given by $S / N^{1/2}$, where S is an estimate of the dispersion of the distribution calculated from the standard deviation of the N observations. Notice that this standard error, or precision of estimation, is dominated by the size of the sample N; the larger the sample size, the smaller the standard error and so the greater the precision.

With respect to the MLE procedure just described for the Rasch model, Ronald Fisher proved that as long as sample size is reasonably large, the standard error of a ML estimate is well estimated by the inverse negative square root of the second derivative of the likelihood function.

Fisher also proved that replicates of a MLE will, with sufficiently large sample size, have a normal distribution with expected value equal to the parameter itself and with a standard deviation equal to this standard error. We will use this result to set confidence limits on our MLE's.

The second derivative of the likelihood, which served as a scaling factor for Newton's method, now plays an important statistical role. It gives us the standard errors for our estimates.

Thus from *Equations 15.10* and *15.14*

$$SE(D_i) = \left[\frac{-\partial}{\partial D_i}\left(\frac{\partial K}{\partial D_i}\right)\right]^{-1/2} = \left(\sum_r^{L-1} N_r Q_{ri}\right)^{-1/2} \sim 2.5 / N^{1/2} \qquad \textit{15.21}$$

$$SE(B_r) = \left[\frac{-\partial}{\partial B_r}\left(\frac{\partial K}{\partial B_r}\right)\right]^{-1/2} = \left(\sum_i^{L} Q_{ri}\right)^{-1/2} \sim 2.5 / L^{1/2} \qquad \textit{15.22}$$

These standard errors are determined by substituting into $P_{ri} = \exp(B_r - D_i) / \left[1 + \exp(B_r - D_i)\right]$ the converged values of the B_r and D_i estimates and finding $Q_{ri} = P_{ri}(1 - P_{ri})$.

Item calibration and person measurement is now complete. Here is a summary of the results of this MLE.

A set of L items estimates D_i is obtained whose sum has been set to zero so that the measuring system under construction has been centered on the calibrations of these test items. Some values will be negative, indicating relatively easy items and some will be positive, indicating relatively hard items.

Associated with each of the D_i estimates is its estimated standard error. We will see that D_i's with values far from the sample of persons have relatively large standard errors and that standard errors get smaller (and hence items more precisely estimated) as we get closer to items with D_i values near the B_r values of the majority of the persons. This is a consequence of the formula for the standard error in *Equation 15.21*. When D_i is equal to B_r, the value of P_{ri} is 0.5 and so Q_{ri} is 0.25, its maximum possible value. This is where the standard error is nearest to its theoretical minimum of $2 / N^{1/2}$.

Equation 15.22 shows the way the standard error of B_r is a function of the number of items near that B_r. The standard errors of the person abilities depend on how many of the item difficulties are near the location of that person. The more items near the person, the smaller the standard error of the measure. In a test which is well centered on its target group we would expect the standard errors of person abilities to be symmetric around a central B near zero, corresponding to about half the items correct, and that these standard errors would be large at both ends of our variable line and get smaller towards the center.

Since MLE produces values for all scores r, for $r = 1$ to $r = L - 1$, we will have ability estimates and standard errors for all possible scores, even when, in the calibrating sample, there were no persons who actually obtained a particular score.

16. PARAMETER ESTIMATION

Parameter estimation for Rasch measurement is usually done by a computer program like BICAL (Wright & Mead, 1976), BIGSCALE (Wright, Linacre & Schulz, 1990) or BIGSTEPS (Wright & Linacre, 1997).

The PROX procedure, however, is a method of estimation so easy to apply that it is completely manageable by hand. The simplicity of PROX is useful because it details exactly how the Rasch model works in practice. PROX accomplishes the primary aims of Rasch item analysis:

1) linearization of item raw scores (P-values) onto an interval scale with relevant errors of calibrations and

2) adjustment for the sampling effects of person ability.

In so doing, PROX almost always approximates the results obtained by more elaborate procedures extremely well.

The simplification which enables PROX is to approximate the effects on item calibration of sample ability with a sample mean and standard deviation and the effects on person measurement of test item difficulty with a test mean and standard deviation. This simplification makes PROX easy to apply by hand. Nothing more than the observed distributions of item and persons scores, a hand calculator and pencil and paper are needed.

PROX is as applicable to large assessment problems like national item banking as it is to evaluating a small classroom of examinees. The PROX algorithm is the working basis of most successful computer assisted testing (CAT) programs.

In this chapter we

1) outline the PROX equations, and

2) explain how these equations implement Rasch measurement.

The chapter can be used to guide the user in calibrating items and measuring persons from their own data. An example worked out in numerical detail can be found in the second chapter of *Best Test Design* (Wright & Stone, 1979, pp. 28-45).

THE PROX ESTIMATION EQUATIONS

PROX simplifies the representation of person abilities b_n to a normal distribution with sample mean ability M and sample ability standard deviation σ and the representation of item difficulties d_i to a normal distribution with test item mean difficulty H and test item difficulty standard deviation w.

When that is done, then the measure b_n for person n with person score r_n on a test of L items becomes

$$b_n = H + X\log\left[r_n / (L - r_n)\right] \tag{16.1}$$

where

H = mean difficulty of the L items taken,
X = the scaling necessary to adjust for the difficulty standard deviation w of these L items, r = the raw test score of person n ,

and the calibration d_i for item i with item score s_i from a sample of N persons is

$$d_i = M + Y\log\left[(N - s_i) / s_i\right] \tag{16.2}$$

where $d_i = M + Y\log\left[(N - s_i) / s_i\right]$
M = mean ability of the N persons taking the test,
Y = the scaling necessary to adjust for the ability standard deviation σ of these N persons,
s_i = the raw sample score of item i.

and

$$X = \left[1 + (w^2 / 2.89)\right]^{1/2} \cong \left[1 + (w^2 / 5.8)\right] \tag{16.3}$$

$$Y = \left[1 + (\sigma^2 / 2.89\right]^{1/2} \cong \left[1 + (\sigma^2 / 5.8)\right]. \tag{16.4}$$

The divisor $2.89 = 1.7^2$ comes from the scaling factor 1.7 which, because the logistic ogive for values of $1.7z$, is never more than one percent different from the normal ogive for values of z, brings the cumulative logistic distribution into approximate coincidence with the cumulative normal distribution. PROX uses this coincidence to obtain its simplification.

The estimates b_n and d_i have standard errors

$$SE(b_n) = X[L / r_n(L - r_n)]^{1/2} \cong 2.5 / L^{1/2} \tag{16.5}$$

$$SE(d_i) = Y[N / s_i(N - s_i)]^{1/2} \cong 2.5 / N^{1/2}\ . \tag{16.6}$$

APPLYING THE PROX ESTIMATION EQUATIONS

This estimation method can be applied to observed item scores s_i by calculating the sample score logit of item i as

$$x_i = \log[(N - s_i) / s_i] \tag{16.7}$$

and to the observed person scores r_n by calculating the test score logit of person n as

$$y_n = \log[r_n / (L - r_n)] \tag{16.8}$$

The scaling coefficients X and Y can be estimated from

$$X = \{[1 + (U / 2.89)] / [1 - (UV / 8.35)]\} \tag{16.9}$$

for the person logit scaling coefficient and

$$Y = \{[1 + (V / 2.89)] / [1 - (UV / 8.35)]\} \tag{16.10}$$

for the item logit scaling coefficient.

Where $U = \left(\sum_i^L x_i^2 - Lx_.^2 \right) / (L-1)$ and $8.35 = 2.89^2 = 1.7^4$,

$$U = \left(\sum_i^L x_i^2 - Lx_.^2 \right) / (L - 1) \tag{16.11}$$

is the item logit variance,

$$V = \left(\sum_n^N y_n^2 - Ny_.^2 \right) / (N - 1) \tag{16.12}$$

is the person logit variance, $x_\bullet = \sum_i^L x_i / L$ is the item logit mean, and $y_\bullet = \sum_n^N y_n / N$ is the person logit mean.

To complete the estimation, we anchor the scale "zero" at the center of the test by defining $H \equiv 0$ so that

$$d_i = N| + Yx_i = Y(x_i - x_\bullet) \text{ because } M = -Yx_\bullet \tag{16.13}$$

for each item difficulty and

$$b_n = H + Xy_n = Xy_n \text{ because } H \equiv 0 \tag{16.14}$$

for each person ability.

STANDARD ERRORS

The standard errors of these person and item estimates are

$$SE(b_n) = X[L / r_n(L - r_n)]^{1/2} \cong 2.5 / L^{1/2} \tag{16.15}$$

and

$$SE(d_i) = Y[N / s_i(N - s_i)]^{1/2} \cong 2.5 / N^{1/2} \, . \tag{16.16}$$

SAMPLE STATISTICS

The estimates of the person sample mean M and standard deviation σ are

$$M \approx -Yx_{\bullet} \qquad 16.17$$

$$\sigma \approx 1.7(Y^2 - 1)^{1/2} \qquad 16.18$$

ANALYSIS OF RESIDUALS

When we have estimated b_n and w we can use them to obtain the difference between the model's prediction and the data observed. Residuals from the model are calculated by estimating from b_n and d_i the model expectation at each response χ_{ni} and the subtracting this expectation from the $\chi_{ni} = 0$ or 1 which was actually observed.

The model expectation for χ_{ni} is Π_{ni} where the Rasch model for Π_{ni} is $\Pi_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$ and B_n and D_i are the parameters which b_n and d_i estimate.

The standardized residual from expectation is

$$z_{ni} = (\chi_{ni} - \Pi_{ni}) / [\Pi_{ni}(1 - \Pi_{ni})]^{1/2} \text{ .} \qquad 16.19$$

When the data fit the model this standardized residual is distributed with mean zero and variance one. Although the expected sampling distribution of D_i is not normal, we have found that values below -2 and above +2 are useful as indicators of noteworthy misfit.

We estimate Π_{ni} from

$$P_{ni} = \exp(b_n - d_i) / [1 + \exp(b_n - d_i)] \qquad 16.20$$

where b_n and d_i are the estimates of B_n and D_i, and then use the sampling distributions for $z_{ni} \approx (\chi_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})]^{1/2}$ of $z_{ni} \sim N(0,1)$ and $z_{ni}^2 \sim X_1^2$ as guidelines for evaluating the extent to which any particular set of data is sufficiently coherent to construct useful measurement.

PERSON FIT

To measure the validity of a person's performance, calculate the sum of squared residuals z_{ni}^2 for that person. When the person's behavior is useful for measurement because their response pattern fits the measurement model, then their sum of square standard residuals will approximate a chi-square statistic

$$\sum_{i}^{L} z_{ni}^2 = C_n^2 \sim X_{fn}^2 \qquad 16.21$$

with degrees of freedom

$$f_n = (L-1)(N-1)/N \qquad 16.22$$

and a mean square statistic

$$V_n = C_n^2 / f_n \sim F_{fn,\infty} \qquad 16.23$$

with degrees of freedom f_n and ∞ .

ITEM FIT

To measure the validity of an item's usage calculate the sum of squared residuals z_{ni}^2 for that item. When the item is useful for measurement because the pattern of its responses fits the measurement model, then its sum of squared standard residuals will approximate a chi-square statistic

$$\sum_{n}^{N} z_{ni}^2 = C_i^2 \sim X_{fi}^2 \qquad 16.24$$

with degrees of freedom

$$f_i = (N-1)(L-1)/L \qquad 16.25$$

and a mean square statistic

$$U_i = C_i^2 / f_i \sim F_{fi,\infty} \qquad 16.26$$

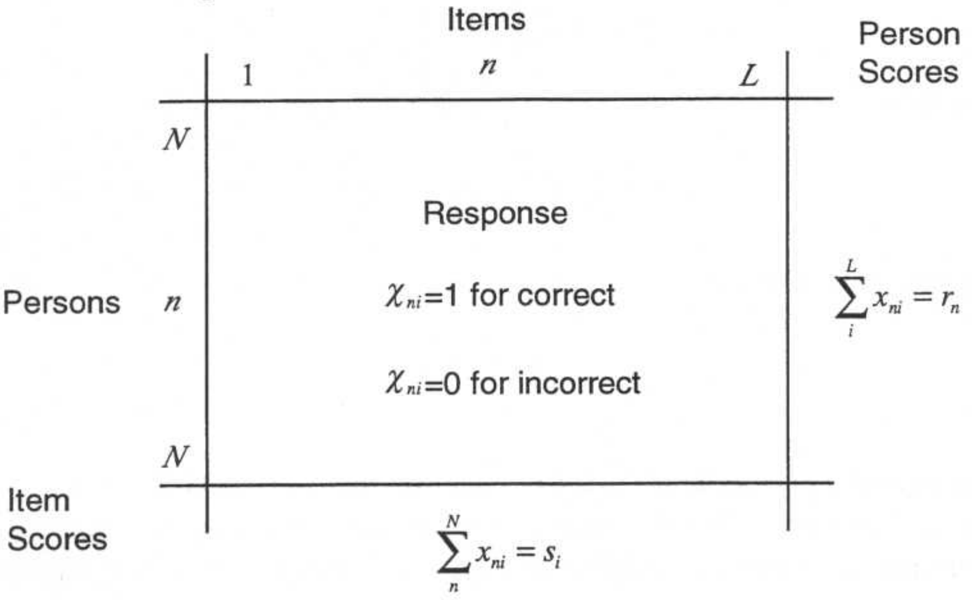
with degrees of freedom f_i and ∞.

UNDERSTANDING HOW THE RASCH MODEL WORKS

Now we will reexamine the details of PROX to discover what is accomplished in its application. The PROX formula enables a simple and intuitive approach to understanding how Rasch item calibration and person measurement work.

THE RESPONSE MATRIX

Consider the response matrix:


Items
Person Scores
1
n
L
N
Response
Persons
n
χ_{ni}=1 for correct
χ_{ni}=0 for incorrect
$\sum_{i}^{L} x_{ni} = r_n$
N
Item Scores
$\sum_{n}^{N} x_{ni} = s_i$

When there is no missing data* so that this response matrix is complete, then every row total gives a person score for the same set of items and every column total gives an item score for the same set of persons. These item scores are reported as "P-values" in traditional item analysis. "P-values" are calculated by dividing the item score s_i by the number of persons N so that the P-value for item i becomes $P_i = S_i / N$, the proportion of correct responses to item i.

If the person raw scores r_n and item P-values $P_i = S_i / N$ were linear objective quantifiers of person ability and item difficulty, our work would seem done at this point. Indeed, person raw scores and item P-values are as far as traditional person measurement and item analysis go. As a consequence, most of the research results in educational measurement have been limited to what little can be done when raw counts of right answers are mistaken for measures.

As measures, however, raw counts have two serious drawbacks.

* In the application of PROX (or any other Rasch model procedure), the data need not be complete. Missing data can be accommodated without trouble. This happens because the measurement structure specified by the model needs only enough data to identify a finite estimate for each person and each item. As a result none of the estimation procedures need complete data to obtain good estimates.

RAW SCORES ARE NON-LINEAR

Raw counts are bounded in range between none right and all right. Because of this they cannot represent abilities or difficulties on a linear (interval) scale. But, if they are not linear, then the results of the arithmetic used in statistical analysis become misleading. In order to enable the substantial benefits of statistical analysis, we must transform the non-linear scores into linear measures.

RAW SCORES ARE TEST AND SAMPLE DEPENDENT

Each observed response is the result of a person of some ability attempting an item of some difficulty. Because of this, the magnitudes of item scores s_i and P-values P_i, which are summed over persons, depend on the particular abilities of this particular sample of persons and so are sample dependent.

The magnitudes of person scores r_n, which are summed over items, depend, in turn, on the particular difficulties of this particular set of items and so are test dependent.

In order to make general use of the information about person measures and item difficulties which is contained in the test item response data, we must liberate the numerical representations of person measures and item difficulties from the local effects they have on one another. We must construct objective person measures which are test-free and objective item calibrations which are sample-free.

CONVERTING NON-LINEAR, LOCALLY DEPENDENT ITEM AND PERSON SCORES TO LINEAR, INDEPENDENT ITEM CALIBRATIONS AND PERSON MEASURES

The way the Rasch model linearizes raw scores and frees them of sample and test dependency can be seen in the following two PROX formulae:

For Items:

$$\underset{\substack{\text{[log odds}\\ \text{linearize item}\\ \text{P-values]}}}{\log\left[\frac{(1-P_i)}{P_i}\right]} * \underset{\substack{\text{[scales out}\\ \text{the sample}\\ \text{variance } \sigma^2\text{]}}}{\left[1+\frac{\sigma^2}{2.7}\right]^{1/2}} \underset{\substack{\text{[adjusts out}\\ \text{the sample } \Rightarrow\\ \text{mean } H\text{]}}}{+\; M} \;\Rightarrow\; \underset{\substack{\text{[Test-freed}\\ \text{item}\\ \text{calibration]}}}{d}$$

For Persons:

$$\underset{\substack{\text{[log odds}\\ \text{linearize person}\\ \text{raw scores]}}}{\log\left[\frac{r_n}{(1-r_n)}\right]} * \underset{\substack{\text{[scales out}\\ \text{the test item}\\ \text{variance } w^2\text{]}}}{\left[1+\frac{w^2}{2.7}\right]} \underset{\substack{\text{[adjusts out}\\ \text{the test item } \Rightarrow\\ \text{mean } H\text{]}}}{+\; H} \;\Rightarrow\; \underset{\substack{\text{[Test-freed}\\ \text{person}\\ \text{measurement]}}}{b_n}$$

For more detail see the example of PROX item calibration and person measure in Chapter 2 of Best Test Design (Wright & Stone, 1979, pp. 28-45).

17. INFORMATION AND MISFIT ANALYSIS

THE STANDARD ERROR OF A MEASURE

The most immediate and also most convenient quantification of the precision of an empirical measure is the standard error (SE) of the measure's estimate. The magnitude of the SE is either determined from the factual structure of the measuring instrument (as in "to the nearest sixteenth of an inch" on a ruler) or calculated from the measurement model used to calibrate the instrument. It is usually estimated from the same data used to estimate the measure. This SE estimates the standard deviation of innumerable independent replications of the data collecting process, when the only disturbances imagined are those anticipated by the measurement model.

The convenience of the SE quantification of precision is that it is in the units of the measure and so can be used directly to specify:

1) an "identification of misfit", as in outside three standard errors (SE);

2) a "region of confidence", as in within two standard errors (SE);

3) an "allowance for error", as in one standard error (SE).

The inconvenience of the SE is that when several samples of independent data bearing on a common quantity to be estimated are combined to form a "better" estimate or when it is useful to keep track of the sequential improvement of "precision" during a stepwise process of data collecting, the corresponding SE's are not additive.

INFORMATION

Ronald Fisher devised a cure for this inconvenience in the 1920's (e.g. 1935, p. 182 ff.). While the SE's of a series of independently obtained, but commonly bearing, commensurable measures are not additive, their inverse squares are. When applied to Rasch measurement "Fisher information" can be defined as

$$I = C / SE^2$$

where C = a constant chosen to specify convenient "information" units.

For dichotomous data, as in test item responses scored 0 or 1, the inverse square of each measure's SE is proportional to a count of how many "standard" items inform that measure.

In particular, when $C \equiv 4$, then

$$I = 4 / SE^2$$

becomes the minimum number of perfectly targeted (i.e. maximally informative) items it would take to produce this SE. I is the "information" in the estimate. We will call the units of $I = 4 / SE^2$ "EQUITS," for EQUivalent on-target ITemS.

The additivity of $I = 4 / SE^2$ can be seen in the algebraic definition of SE for Rasch modeled dichotomous data.

Then

$$SE^2 = 1 / \left[\sum_i^L P_i (1 - P_i) \right]$$

where

$$P_i = \exp(B - D_i) / \left[1 + \exp(B - D_i) \right]$$

is the probability of a right answer given person measure B and item difficulty D_i, and $\sum_i^L$ signifies summation over the L items taken. Thus

$$I = \left(4 / SE^2 \right) = 4 * \sum_i^L \left[P_i (1 - P_i) \right]$$

is an expression which adds $\left[P_i (1 - P_1) \right]$ over items.

When every item is perfectly targeted, then

$$P_i = 1/2,$$

$$P_i (1 - P_i) = 1/4$$

and so

$$I = 4 * \left[\sum_i^L P_i (1 - P_i) \right] = L$$

the number of responses to perfectly targeted items necessary in order to obtain this particular

$$SE = \left\{ 1 / \left[\sum_i^L P_i (1 - P_i) \right] \right\}^{1/2}$$

COMPARING INFORMATION

When we wish to compare the information value of a pair of measures, we can use their corresponding SE's and this definition of information, $I = (4 / SE^2)$, to find out which measure contains more information and by how many "equivalent on-target items" or "equits".

Thus for measures B_1 and B_2 we have

$$I_1 = 4 / SE_1^2 \text{ equits}$$

$$I_2 = 4 / SE_2^2 \text{ equits}$$

for which the advantage in equits of measure B_2 over measure B_1 is

$$I_2 - I_1 = 4\left(1 / SE_2^2 - 1 / SE_1^2\right)$$

$$= 4(SE_1^2 - SE_2^2) / SE_1^2 * SE_2^2$$

Comparison of the information values of a pair of measures can also be calculated from the ratio of their error variances,

$$RE_{21} = I_2 / I_1 = SE_1^2 / SE_2^2$$

This ratio, RE_{21}, gives the "information" provided by the second measure B_2 in units of the "information" provided by the first B_1, i.e., it is the "Relative Efficiency" of the second measure with respect to the first.

MISFIT ANALYSIS

Maximum information is obtained when $P = .5$ so that $P(1 - P) = .25$. While this would appear ideal, there is a catch. Fit analysis requires the possibility of improbable, and hence unexpected responses - responses for which

$$P \rightarrow 1 \text{ but } X = 0 \text{ or } P \rightarrow 0 \text{ but } X = 1.$$

Then, when the highly probable response is *not* observed, misfit and hence invalidity is implied. Were all items targeted successfully near $P = .5$, this kind of fit analysis for verification of response validity would not be possible. Since $X = 0$ or 1 would be equally likely, no improbable condition with which to detect misfit could be observed.

THE EFFICIENCY - FIT PARADOX

1. Responses to items which provide maximum information because $P \rightarrow .5$ allow minimum misfit detection.
2. Responses to items which allow maximum misfit detection because $P \rightarrow 0$ or 1 provide minimum information.

Best test design requires a compromise between these extremes. The simultaneous avoidance of both extremes benefits greatly from prior knowledge concerning the relative locations of items and persons.

BEST TEST DESIGN

We want the items to elicit maximum information from the person. But we must balance the amount of information (reliability, precision) gained against the concomitant loss of opportunity to detect misfit and, hence, to verify validity. Where we have no knowledge of a person's ability, then items must be of a difficulty range sufficiently wide to cover the reasonable possibilities. This means that while some items will identify the location of the person between items passed and items failed, other items will inevitably turn out to be too far from the person's discovered ability to contribute much information about that ability. The off-target items, however, will be useful for identifying misfit and thus verifying validity.

When we have a useful expectation about where a person is on the variable to be measured, then item selection can be accomplished with maximum utility and efficiency by focusing most of the items on the interval in which we expect the person to be located, but including some additional intentionally off-target items to verify the validity of this location.

We use enough targeted items to "fix" the person's location with sufficient precision ($SE^2 \approx 4 / L$), where this L is the number of on-target items) for our testing purpose. Then we add enough additional off-target items (2 logits above and below where we expect the person to be located) to verify the validity of our measure.

The efficiency of this design depends on the extent of our knowledge of the person prior to the test. Without some prior focusing knowledge, we must use a wide range of items. This will guarantee enough off-target items to validate the measure, but will cost more items than a narrow on-target test to reach equivalent precision.

Targeting an educational test to a particular student requires both the art of knowing the student and the science of measurement. Teaching intuition can guide expectations in the absence of quantitative knowledge. When previous measurements are also available, they too can be utilized.

INFORMATION, EFFICIENCY AND PRECISION

The way information and efficiency enter into judging the value of an observation is through their bearing on the precision of measurement. Measurement precision depends on the number of items in the performance and on the difference in logits $|B - D|$ between each item difficulty and the person's ability. We can simplify the evaluation of each item's contribution to our knowledge of the person by calculating what percent of a best possible item the item in question contributes. These are the values of INF entered in Column 2 of Table 17.1.

We call this information index $INF = 400[P(1 - P)]$

the "relative efficiency" of the observation.

The relative efficiency (INF) is the I defined in *Equation 17.8* but scaled by the factor 100 so

Table 17.1

Information and Misfit Statistics

1 LOGIT DISTANCE BETWEEN PERSON AND ITEM	2 % EFFICIENCY OF AN OBSERVATION AT $\lvert B-D \rvert$	3 NUMBER OF ITEMS L NEEDED TO MAINTAIN EQUAL PRECISION	4 IMPROBABILITY OF AN UNEXPECTED ANSWER AT $\lvert B-D \rvert$
$\lvert B-D \rvert$	$INF = 400P(1-P)$	$L = 1000 / INF$	$P = 1/\left[1+\exp(\lvert B-D \rvert)\right]$
0.0, 0.3	100	10	.50
0.4, 0.8	90	11	.33
0.9, 1.2	75	13	.25
1.3, 1.4	65	15	.20
1.4, 1.5	55	18	.17
1.7, 1.8	50	20	.14
1.9, 2.0	45	22	.12
2.1	40	25	.11
2.2	36	28	.10
2.3	33	30	.09
2.4	31	32	.08
2.5	28	36	.08
2.6	25	40	.07
2.7	23	43	.06
2.8	21	48	.06
2.9	20	50	.05
3.0	18	55	.05
3.1	16	61	.04
3.2	15	66	.04
3.3	14	73	.04
3.4	12	83	.03
3.5	11	91	.03
3.6	10	100	.03
3.7	9	106	.02
3.8	9	117	.02
3.9	8	129	.02
4.0	7	142	.02
4.1	6	156	.02
4.2	6	172	.02
4.3	5	189	.01
4.4	5	209	.01
4.5	4	230	.01
4.6	4	254	.01

Wright & Stone, 1979. *Best Test Design.* Chicago: MESA Press. Pages 73 and 216.

that it will give the amount of information provided by the observation at $|B-D|$ as a percentage of the maximum information that one observation "exactly on target" at $|B-D|=0$ would provide.

The relative efficiency (INF) of an observation can be used to estimate the potential value of any particular item for measuring a particular person. This can be done by considering how much information would be lost by removing that item from the test. Thus, $\text{INF}=23\%$ for $|B-D|=2.7$ indicates how much of a perfectly targeted item we gain by including that item in the measurement of the person and conversely how much we lose by omitting that item. The "how much" is 23% of the most we could get from one item exactly on target at $|B-D|=0$.

When an item and person are close to one another $|B \quad D| \quad 0$; i.e., on target, then the item contributes more to the measure of the person than when the item and person are far apart $|B \quad D|$. The greater the difference between item and person, the greater the number of items needed to obtain a measure of comparable precision and as a result, the less efficient each item.

Once we have estimates of person ability B to combine with our knowledge of item difficulty D, we can determine the relative efficiency of any item. Column 2 of Table 17.1 gives the percent relative efficiency (INF) by which any observation at the absolute difference $|B-D|$ given in Column 1, provides information about that person-item interaction.

It requires five $\text{INF}=20\%$ items at $|B-D| \to 2.9$ to provide as much information about a person as could be provided by one $\text{INF}=100\%$ item at $|B-D| \to 0$.

When $|B-D|$ is three, it takes four times as many items to equal the information to be had from items in the $|B-D|<1$ region, within one logit of the person.

The test length necessary to maintain a specific level of measurement precision is inversely proportional to the relative efficiency of the items used. The number L of less efficient items necessary to match the precision of 10 exactly-on-target, $|B-D|=0$, items is given in Column 3 of Table 17.1.

Column 3 shows $L=1000/INF$ the number of items needed to maintain equal precision over the range of possible values of $|B-D|$.

There is also, however, the verification or validation of test performance validity to keep in mind. When we are off-target because $|B-D|>2$ or 3, then we can use the possibility of unexpected (improbable) responses to evaluate response validity. Column 4 in Table 17.2 gives the probability of an unexpected response (i.e. the improbability of the observed response) for each value of $|B-D|$.

Note that as $|B-D|>2.8$, the probability of an unexpected response such as

$$X=0 \text{ when } (B-D)>2.8 \text{ or } X=1 \text{ when } (B-D)<-2.8$$

drops to P = .05. This produces the possibility of a statistically significant "misfit" and hence of a probable invalidity in that response to that item.

Detailed examples of misfit analysis are given in Chapter 4 of Best Test Design (Wright and Stone, 1979).

To standardize our use of Table 17.1, we use this guide:

Location of Item	(Ability-Difficulty) Difference	Item Efficiency and Misfit Detection
Right on Target	$\|B-D\| < 1$	- excellent efficiency, 75% or better - no misfit analysis possible
Close Enough	$1 < \|B-D\| < 2$	- good efficiency, 45% or better - no misfit analysis possible
Slightly Off	$2 < \|B-D\| < 3$	- poor efficiency, less than 45% - misfit detectable when unexpected responses accumulate
Rather Off	$3 < \|B-D\| < 4$	- very poor efficiency, less than 18% - even single unexpected responses indicate irregularity
Extremely Off	$4 < \|B-D\|$ -	- virtually no efficiency, less than 7% unexpected responses always require diagnosis

18. SEPARATION STATISTICS

The person and item separation statistics used in Rasch measurement are still unfamiliar to some practitioners (Wright, Mead and Bell, 1979; Wright and Stone, 1979; and Wright and Masters, 1982). This primer describes the separation statistics and illustrates some of the key concepts and procedures of separation by working through an example.

A variable can be thought of as a straight line. To measure successfully we must be able to locate both items and persons along this line. Items are located by the number of persons getting a specific item correct. Persons are located by how many items they were able to answer correctly. Items to the left on the line are easier than those to the right while persons to the left have less ability than others to the right.

It is necessary to locate persons and items along the variable line with sufficient precision to "see" between them. Items and persons must be separated along this line for useful measurement to be possible. But, separation that is too wide usually signifies gaps among item difficulties and person abilities. This leads to imprecise measurement. Separation that is too narrow, however, signifies redundancy for test items and not enough differentiation among person abilities to distinguish between them.

Items must be sufficiently well separated in difficulty to identify the direction and meaning of the variable. To be useful, a selection of items, a test, must separate relevant persons by their performance. The item locations are the operational definition of the variable of interest while the person locations are the application of the variable to measurement.

The item and person separation statistics in Rasch measurement provide an analytical tool by which to evaluate the successful development of a variable and with which to monitor its continuing utility. Person separation indicates how efficiently a set of items is able to separate those persons measured. Item separation indicates how well a sample of people is able to separate those items used in the test. Where these statistics are expressed as reliabilities, they range from 0.0 to 1.0. The higher the value the better the separation that exists and the more precise the measurement.

PERSON SEPARATION RELIABILITY

The data in Table 18.1 are from the calibration output of the BICAL program developed by Wright and Mead in 1976. This example is based upon the calibration of 14 Knox Cube tapping items taken by a sample of 34 persons. These data are also discussed in Wright, Mead and Bell, 1976 and Wright and Stone, 1979. At the bottom of the table the Person Separation Reliability (PSR) is given as 0.68. This indicates that the 14 items used in this version of the Knox Cube Test were able to separate the 34 people tested to a moderate degree.

The PSR reported by most Rasch computer programs is calculated by subtracting the ratio of the sample mean square person measure error (MSEp) to the sample person measure variance (SD_p^2) from one. The formula (Wright and Masters, 1982, p.106) is:

Figure 18.1

Original responses of 35 persons to 18 items on the Knox Cube Test.

	Item Name																		
PERSON NAME	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	PERSON SCORE
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	7
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
3	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	10
4	1	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	6
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
6	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
7	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	14
8	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
9	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
10	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
11	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	8
12	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	8
13	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	10
14	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
15	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	13
16	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	10
17	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	9
18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	11
19	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	9
20	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	11
21	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	12
22	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	12
23	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	12
24	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	14
25	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	5
26	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
27	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	7
28	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	10
29	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0	10
30	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	9
31	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	10
32	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	11
33	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	6
34	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	12
35	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ITEM SCORE	35	35	35	32	31	30	31	27	30	24	12	6	7	3	1	1	1	0	

Wright & Stone, Best Test Design (Chicago: MESA Press, 1979), p.31.

$$PSR = 1 - \left[\frac{MSEp}{SDp^2}\right] \qquad 18.1$$

Table 18.1 shows the data necessary to calculate the PSR for this example. Using these data the equation becomes:

$$PSR = 1 - \frac{32.99}{104.93} = 0.68 \qquad 18.2$$

This Person Separation Reliability is comparable to the KR20 measure of internal consistency. The PSR can be corrected for degrees of freedom [*L*/(*L*-1)] and yield a result which is very similar to that of the KR20. With this correction the PSR becomes:

$$PSR = \left[1 - \frac{32.99}{104.93}\right] \times \left[\frac{13}{12}\right] \qquad 18.3$$

$$PSR = 0.74$$

This "corrected" Person Separation Reliability is close in both formulation and value to the KR20 reliability coefficient. In order to compare the PSR to the KR20, we need to return to the original item-by-score matrix analyzed by the Rasch computer program, BICAL. This matrix is given in Figure 18.1 and comes from Wright and Stone, 1979, p.31. Calculation of the KR20 on this 14 x 34 matrix produces a KR20 of 0.72. The difference between this KR20 of 0.72 and the "corrected" PSR of 0.74 is due to the curvilinear relation between the nonlinear raw scores on which the KR20 is based and the linear logit measures on which the PSR is based.

These calculations of the PSR of Rasch psychometrics and the KR20 on the same data matrix illustrate their equivalence and elucidate the calculation of person separation reliability.

The similarity of these results can be seen in the similarity of the formulae:

$$KR20 = \left[\frac{L}{L-1}\right] \times \left[1 - \frac{\sum pq}{\sigma_y{}^2}\right] \qquad 18.4$$

and

$$\text{Rasch } PSR = \left[\frac{L}{L-1}\right] \times \left[1 - \frac{MSEp}{SDp^2}\right] \qquad 18.5$$

In these formulae *pq* is the error variance of the response score of a "person" for whom the sample item p-values apply, while MSE*p* is the sample average measure error variance in logits and $\sigma_y{}^2$ is the sample variance of the nonlinear raw scores, while $SD_p{}^2$ is the sample variance of the linear logit measures.

Table 18.1

Knox Cube Test Output from Bical

Raw Score	Count	Logit Ability	Standard Error
11	2	3.31	0.92
10	1	2.53	0.93
9	4	1.71	0.96
8	5	0.81	1.03
7	12	-0.22	1.07
6	3	-1.19	0.97
5	2	-1.96	0.86
4	2	-2.61	0.81
3	2	-3.21	0.81
2	1	-3.86	0.88

Person Separation Reliability = 0.68 (without D. F. Correction)

(Best Test Design, Page 57, Table 3.4.2)

ITEM SEPARATION RELIABILITY

KR20 is commonly calculated for items, but almost never for persons. When Hoyt (1941) published his paper on test reliability by ANOVA, he recognized both approaches saying "extended examination of the 'among' items variance would make it possible to decide on the heterogeneity of the respective difficulties of the items while a more extended examination of the 'among students' variance would make it possible to answer certain pertinent questions regarding the individual differences among students" (page 156). This good advice, however has never been followed in practice.

In Rasch measurement, however, the Item Separation Reliability (ISR) is routine (Wright and Masters, 1982, p.92). This ISR gives the test user an indication of how well items are separated by the persons taking the test. The formula for this index is:

$$ISR = 1 - \left[\frac{MSE_I}{SD^2{}_I}\right] \qquad 18.6$$

This is calculated in a fashion similar to the Person Separation Reliability. The higher the ISR, the better those particular items are separated by the persons taking the test.

It is not the algebraic and statistical similarity of the KR20 and Rasch PSR, however, that is of major importance now that it has been demonstrated. Instead, it is the decomposition of these single indices into their constituent parts that leads to a more detailed and more useful management of the test characteristic traditionally referred to as "reliability."

With Rasch calibration we are able to obtain the standard error of calibration for each individual item as well as the standard error of measurement for each person ability. With traditional methods, a standard error of measurement is provided only for measures at the group mean of person ability.

The standard error specific to each item (or person) statistic is far more useful than any single sample (or test) "average". The location of each item and person on a line representing the variable together with their standard errors shows us the definition and utility of the variable. The definition of the variable is specified by the location of the items. The utility of the variable for measuring persons is quantified by the standard error which accompanies each person measure.

19. RELIABILITY AND SEPARATION

Validity and reliability have been key concepts in measurement for eighty years. These two topics command Chapters 1 and 2 of the Standards for Educational and Psychological Testing (1985). The Standards define reliability as "the degree to which test scores are free from errors," (1985, p.19). The "errors" referred to are measurement errors. The magnitude of these errors and the specification of their source are necessary in order to determine the efficacy of a measuring instrument. The reliability coefficient is the traditional statistic intended to quantify reliability. Coefficients are commonly reported for test-retest, multiple form and split-half replications. The purpose of this primer is to discuss how these topics are dealt with in Rasch measurement and how this improves and, hence, supersedes traditional methods.

TRADITIONAL RELIABILITY

The KR20 for dichotomous responses (or its generalization, coefficient alpha) are estimates based upon a single administration of a test assumed to have homogeneous items. These coefficients are intended to be an estimate of the test's reliability with respect to a single attribute postulated to underlie all the test items. However, what any particular reliability actually refers to can only be whatever attribute the test items actually define. Sufficient time to answer the items is assumed (timed tests produce spuriously high coefficients). The KR20 and its variants (coefficient alpha and KR21) are calculated by comparing a numerator based on sampled item p-values with a denominator based on the sampled persons' raw scores, computed from the same response matrix of persons and items.

The statistics outlined in Figure 19.1 bring together the two contrasting elements which make up the KR20. One element summarizes the test items in terms of *pq* in which *p* comes from the sampled item p-values (where p = proportion correct) and $q = 1 - p$. Each item *pq* is the variance of a response to that item for a "person" for whom that p-value is their probability of succeeding on that item. Since the *p*-value for an item is the sample mean of the dichotomous person responses to that item, this *p*-value is what we expect of an "average person" from that sample on that item.

The *p*-value for an item describes a "sample average person's" probability of success on that item and can be used to estimate an "average" sample response variance for that item. When these variances are summed over the items they yield a score variance for a "person" who has exactly those *p*-values. This "average" test score variance is the numerator in KR20.

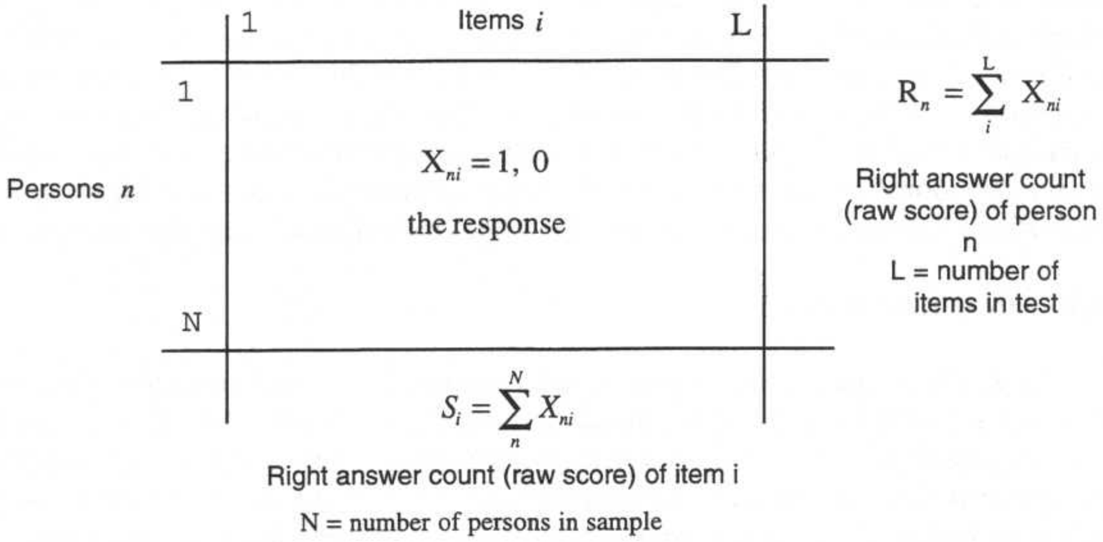
The KR20 denominator is the observed sample variance of person scores. Thus the KR20 combines a "test" characteristic for a "typical" person sampled, based on item p-values, with a "sample" characteristic from the observed sample variance of person raw scores.

CHARACTERISTICS OF THE KR20 STATISTIC

1. The item response variance used is that of an "average" person sampled. This is not the same as an average of the persons' test score error variances. If the sample score distribution is not symmetric, then the error variance of an "average" person must be different from the average of individual persons' error variances.

Figure 19.1

Traditional analysis of a response matrix.


1
Items i
L
1
Persons n
N
$X_{ni} = 1, 0$
the response
$R_n = \sum_i^L X_{ni}$
Right answer count (raw score) of person n
L = number of items in test
$S_i = \sum_n^N X_{ni}$
Right answer count (raw score) of item i
N = number of persons in sample

Items: **P-value of item *i:***

$$p_i = S_i / N \qquad q_i = (1 - p_i)$$

variance for item *i*: $p_i q_i = [S_i / N][(N - S_i) / N]$

Person raw score variance for a "person" for whom the sample item p-values are their probabilities of success:

$$VR = \sum_i^L p_i q_i$$

Persons: Sample raw score mean: $R = \sum_n^N R_n / N$

Sample raw score variance: $VS = (\sum_n^N R_n^2 / N) - R^2$

KR20 reliability: $KR20 = [L / (L - 1)][1 - (VR / VS)]$

2. While pq provides a test score error variance for an "average" person, we know that the sampled people vary, i.e., the variance of their raw scores is greater than zero. Persons with high or low scores have less score error variance than those with scores near fifty percent correct where the score error variance is maximum. Since the "average" person variance used in the KR20 formula is always larger than the lower score error variance of persons with extreme scores, it must always overestimate their score error variances.

3. If we want to anticipate reliability for a proposed application, a previously reported KR20 cannot be used as is, unless we know that the proposed sample will have the same score distribution as the sample used for the reported KR20. This is quite unlikely.

4. The use of raw scores as the data for calculating the sample variance is misleading to the extent that raw scores are not linear representations of the variable they are intended to indicate. Proof that raw scores cannot be linear representations can be seen by plotting the raw scores from a hard test against the raw scores from an easy test measuring the same attribute.

Figure 19.2 shows that the relationship between this pair of raw scores must be curvilinear. As a result, neither set of raw scores can be linear indicators of what they purport to represent. But the calculation of means and variances necessary to estimate reliabilities assumes linearity in the numbers used. Therefore, the calculation of these statistics from raw scores is always incorrect to some unknown degree.

If we expressed person measures in a linear, rather than curvilinear, form, then the sample variance estimates would be improved.

If person error variances were averaged instead of using the error variance of an "average" person, the information about sample test error conveyed by the reliability coefficient would also be improved.

RASCH RELIABILITY

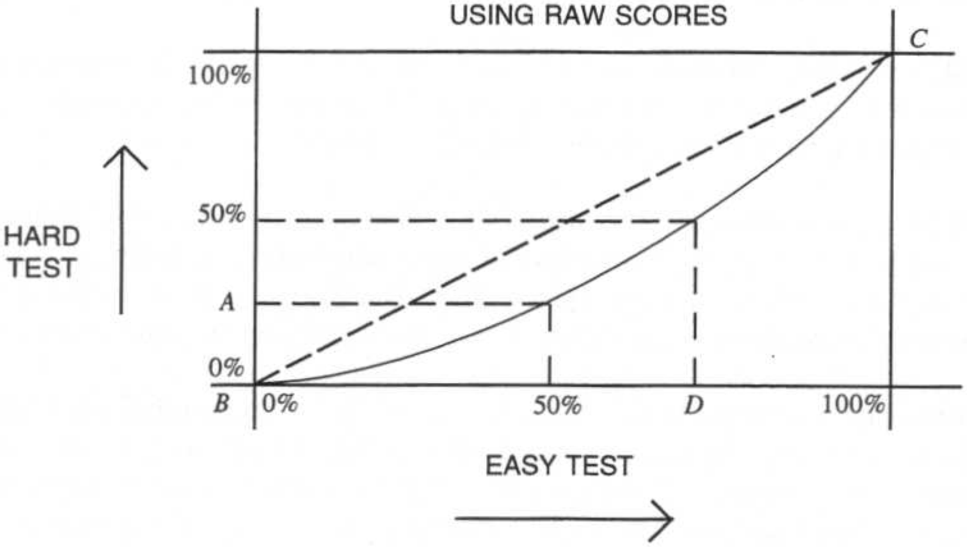
These shortcomings in KR20, or any other reliability coefficients based on raw scores, are remedied when a Rasch measurement analysis is made of the same data and reliability calculated from Rasch results. Rasch measurement produces a measure of each person's ability on a linear scale calculated from a logistic transformation of their raw score. The result is a linear comparison of the Hard and Easy tests as shown in Figure 19.3. These linear ability measures are numerically suitable for calculating sample variances.

We also have, for each person measured, an accompanying standard error of measurement. These individual errors can be squared and summed to produce a correct average error variance for the sample. When these results are substituted for those in the traditional KR20 formula, the result is a new formula which, while equivalent in interpretation, gives a better estimate of reliability than KR20, coefficient alpha, or any other reliability based on nonlinear raw scores.

When terms are replaced in this way, a better reliability coefficient results because (1) the numerical arguments are now linear rather than curvilinear, and (2) the actual average error variance of the sample is used instead of the error variance of an "average" person (see Figure 19.5).

Figure 19.2

Comparing scores from easy and hard tests.



Legend:

Pt. A = less than 50% on Hard Test.

Line BC = presumed linear relation between Easy and Hard test scores.

Arc BC = actual non-linear relation between Easy and Hard test scores.

Pt. D = more than 50% on Easy Test.

Explanation:

0% on the Easy Test implies 0% on the Hard Test (Pt. B).

100% on the Hard Test implies 100% on the Easy Test (Pt. C).

But 50% on the Easy Test implies *less* than 50% on the Hard Test (Pt. A)

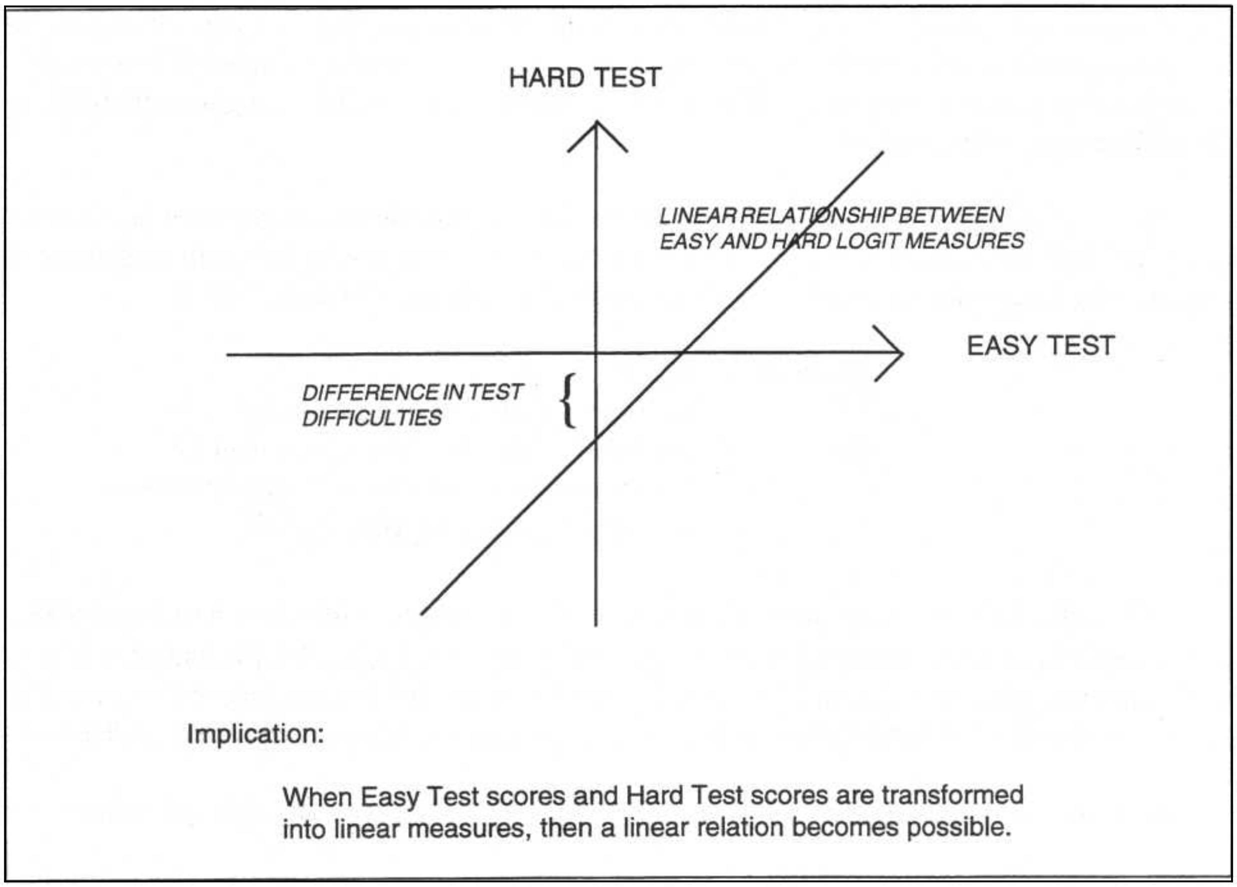
and 50% on the Hard Test implies *more* than 50% on the Easy Test (Pt. D).

Implication:

The score relationship between any pair of tests which differ in difficulty *cannot be linear.*

Figure 19.3

Comparing measures from easy and hard tests using logit measures.


HARD TEST
LINEAR RELATIONSHIP BETWEEN EASY AND HARD LOGIT MEASURES
EASY TEST
DIFFERENCE IN TEST DIFFICULTIES

Implication:

When Easy Test scores and Hard Test scores are transformed into linear measures, then a linear relation becomes possible.

PREDICTING RELIABILITY

In the application of a test, it is the characteristics of the new sample to which we intend to apply the test, rather than a description of some previous sample, that is our real concern. We want to know how the test will work with the new people who are about to take it. We want a relevant reliability coefficient which applies to the people we intend to test, rather than an obsolete one describing people who were previously tested. But few practitioners know how to use an old KR20 to estimate a new KR20 for a new sample.

In fact, it is easy to predict the reliability for a forthcoming sample, if we are willing to postulate an expected mean and variance for this sample. From these statistics and the Rasch targeting formula we can calculate the reliability of the test for the new application without reference to any previous sample (Wright and Stone, 1979, 129-140).

ADVANTAGES OF A SEPARATION INDEX

Correlation-based reliability coefficients, however, are also nonlinear in implication. For example, improvement of KR20 from .6 to .7 is not twice the improvement from .9 to .95. Although the difference in amount of reliability between .9 and .95 is half as much as the difference between .6 and .7. This half-as-much signifies twice the improvement in measurement precision. We can escape this shortcoming of KR20 by replacing the traditional reliability coefficient with a Separation Index (G) (See Figure 19.5).

The Separation Index (G) is the ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample. It is on a ratio scale in the metric of the root mean square measurement error of the test for the sample postulated. It quantifies "reliability" in a simple and direct way and has a clear interpretation. This expedites comprehension of what changes in reliability mean in terms of measurement precision.

The estimation of separations for new samples is easy. No reference to any previous samples is required. We need only estimate the expected standard deviation of our new target sample and then divide this estimate by the average standard error of the intended test for such a sample. As in:

Separation: $G = SDT/SET$
SDT: the expected *SD* of the target sample
SET: the test standard error of measurement for such a sample, a value which is almost always well approximated by $SET = 2.5 / \sqrt{L}$

SET can be estimated more precisely as $\mathrm{SET} = \sqrt{C / L}$ where *L* is the number of items in the test and *C* is a targeting coefficient (explained in Wright and Stone, 1979, pages 135-136 and tabled for most test and target relationships on pages 214-215). *C* varies between 4 and 9 depending on the range of item difficulties in the intended test and the target sample's expected average percent correct on that test.

Here are some values of *C* for typical item difficulty ranges and typical target sample mean percents correct:

Values of the Targeting Coefficient C

Test Item Difficulty Range in Logits

Expected Percent Correct of Target Sample	1	2	3	4	5	6
50	4.0	4.4	4.8	5.3	5.8	6.8
60	4.4	4.4	4.8	5.3	6.2	6.8
70	4.8	5.3	5.3	5.8	6.8	7.3
80	6.2	6.8	6.8	7.3	7.8	8.4

$\mathrm{SET} = \sqrt{C / L}$

L = Number of Items in Test

(See Wright and Stone, 1979, p. 214)

Thus, SET is easy to approximate well enough for the calculation of an expected target sample. Separation *G:* $G = SDT/SET$.

If an expected reliability is also desired, it can be obtained from: $R = G^2 / \left(1 + G^2\right)$.

Rasch Separation Indexes $G = \sqrt{[R/(1-R)]}$	*Corresponding Reliability Coefficients* $R = G^2 / \left(1 + G^2\right)$
1	0.50
2	0.80
3	0.90
4	0.94
5	0.96

We use the Rasch Model in our example. But this Separation Index is applicable to any latent trait model. With it, one can predict the reliability of a test with any sample to be used in a study, if one can specify an expected sample mean and variance. No information about any previous samples is necessary.

The Standards (1985, page 22) recommend that, "Standard errors of measurement be reported at critical score levels." Rasch measurement analysis routinely provides standard errors for every possible test measure along the variable as shown in Figure 19.4. Thus, the Rasch approach meets this recommendation completely. If reliability, as defined by the Standards, is the degree to which test scores are free from errors of measurement, then it follows that every ability measure should be accompanied by a standard error as an index of the degree to which this criterion is met for that measure.

The Rasch measurement errors satisfy this goal by providing individual errors of measurement for every observable measure. If a collective index of reliability is desired, the Rasch Separation Index is more useful in basis and numerical form than the traditional indices of reliability.

Figure 19.5 summarizes the calculation of the Separation Index.

Figure 19.5

Rasch person separation index.

$G = STB / RMSEB$ where

$STB^2 = SDB^2 - MSEB$

$$SDB^2 = \sum_{n}^{N} B_n^2 / N - \left(\sum_{n}^{N} B_n / N \right)^2$$

$$RMSEB^2 = MSEB = \sum_{n}^{N} SEB_n^2 / N$$

B_n = logit measure of person n

SEB_n = standard error of B_n

so $G^2 = R / (1 - R)$ and $R = G^2 / \left(1 + G^2\right)$

and $R = 1 - \left(MSEB / SDB^2\right)$ is

$\approx 1 - (VR / VS) = [(L - 1) / L]KR20$

with VR and VS as defined in Figure 19.1

(see Wright and Masters, 1982, Figure 1, pp. 105-106)

note: $MSEB = C / L$ in which $4 < C < 9$
and $C = 5$ or 6 is typical.

(See Wright and Stone, 1979, pp. 134-136)

20. VALIDITY

According to the Standards for Educational and Psychological Testing (1985), validity is "the most important consideration in test evaluation" (p. 9). Validity deals with the meaning of inferences drawn from test scores. The Standards emphasize that it is the inferences that are validated and not the test. The idea is that no test is valid or invalid in itself. Only its use in some application merits a designation of validity.

This primer discusses how validity is addressed in Rasch measurement. We explain how the types of validity discussed in the Standards are handled in a Rasch analysis of item response data.

TRADITIONAL VALIDITY

The three types of validity discussed in the Standards are (1) content related, (2) criterion related and (3) construct related. Validity itself, however, is held to be unitary. The Standards advance these types as related facets of a single problem. The types must be combined to validate the information obtained from the application of a test.

It is easy to become confused as to what is meant by validity because the three types are different in meaning and method. While the virtue of a single term "validity" is agreed upon by everyone, how to connect that term to the analysis of data is not. There are substantial and puzzling questions as to what is referred to, how it can be implemented in practice and what the results of implementation mean.

Since it is not clear what additional data are required to determine validity, that is what criteria are relevant, it is easy to become confused about what should be done. There is no unique or objective way to determine what the right criterion would be. There are always many possibilities. How do we establish which criteria are necessary, which are optional, which are decisive, which are only advisory? When criteria differ, how do we decide which one to use. Attempts to base validity on external criteria have raised more problems than they have solved. Many articles lament this dilemma (Bechtoldt, 1959; Beck, 1950; Campbell & Fiske, 1959; Cronbach & Meehl, 1955).

REAPPRAISING VALIDITY

The only way to escape this fruitless muddle is to focus on the data that are available, namely the actual responses of individuals to items and then to ask: What is there in these data that could answer validity questions?

What can we get from analyzing the data we have that could tell us about validity? When we look at responses to test items, two, and only two, types of data relevant to validity emerge. The first type concerns the ordering and spacing of items and persons which are produced by the analysis of item responses. The actualization of this kind of validity depends on prior knowledge of item content and person characteristics and, most of all, on clear intentions concerning what variable is to be defined and measured.

ORDER VALIDITY

The relation between item content and the empirical difficulty order of the items produced by the way persons respond to them either verifies, improves on or contradicts the intended definition and hence meaningfulness of the variable which the items are intended to implement. We expect one digit integer addition to be easier (have less difficulty) than division involving decimals and both to be easier than any problem involving a quadratic equation. It is almost impossible to write mathematics items without knowing in advance their intended and expected difficulty ordering.

For a spelling variable, we anticipate sequences of increasing difficulty like "cat", "wagon", "friendship", "meretricious" as defining a spelling variable that could be extended by adding easier and harder words as well as enriched by adding words at intermediate levels of difficulty between these four.

The way to begin this kind of thinking out of a variable is to write or select an initial item and then to write or find another item which we expect, according to our theory, to be easier or harder - continuing in this way, item after item, extending and filling in, step-by-step, until a detailed definition of the intended variable is laid out.

This simple beginning can lay out an orderly and meaningful item definition of an entire variable. We need only to apply a theory of what we are trying to do and to know our measurement intentions in order to think out an expected difficulty order for the items we plan to use. Then, when we use our items with persons, we can compare this intended and expected conceptual order with the empirical order actually provided by the data to see how well our expectations are confirmed. *Item order validity* operationalizes two of the Standard's three types of validity: content validity and construct validity.

Should we discover, however, that we are unable to imagine any canonical order for our items, then we are forced to admit that we do not understand the variable we are trying to define or how our items are supposed to implement its definition. We are forced to realize that we still have more work to do on our variable, by thinking it through more carefully before our purpose will become clear enough to us for useful action. Even in the earliest stages of variable construction we must have some idea of how to write items in an orderly fashion or else our measurement project cannot thrive. We must know ahead of time the difference between an easy and a hard item. We must know our purpose.

The difficulty order of items defines the variable's meaning and hence its content and construct validity. The ability order of persons that is produced by their performance on a test specifies the consequences of measuring on the variable and so determines the variable's utility. Relevant concomitant person orders such as those produced by age, school grade, civil service rating, or any other characteristics which ought to correlate with our intended measure, can help us to learn about its utility and so might be referred to as background criteria for our variable. But the variety of possibilities guarantees that no single criterion can be decisive. Nevertheless, to the extent that there is a part to be played by "criterion validity" in the evaluation of the utility of our variable, it is to be found in *person order validity* - the way persons are ordered by their measures.

CRITERION VALIDITY IN VARIABLE MAPS

The criterion validity of the Standards presupposes the existence of an external criterion sufficiently well established to serve as the base against which the test can be compared. The correlation coefficient is usually used as the index by which this comparison is evaluated. Two strategies are usually employed.

1. A test is designed to predict some already known criterion and the correlation with this criterion is taken to indicate the degree of criterion validity.

2. One test form is correlated with another test form to indicate their degree of consistency with one another.

The apparent simplicity of these approaches is flawed by the problem of the criterion. Is the criterion valid? Can it serve as a stable base? Does the correlation between two test forms address any substantive question about validity?

Criterion validity is better addressed by building an item map of the variable and then augmenting this map with the values of whatever concomitant criteria can be gathered along with the test data. All criteria can be located on this variable map together with the item calibrations and person measures.

When collecting test data we can record the associated person information of gender, age, school and scholastic level. The levels of these criteria can be plotted along with item calibrations and person measures on the variable map to show how these criteria relate to persons measures and also to item content.

We can formulate hypotheses about any criteria that we imagine might be relevant to these item calibrations and person measures and determine from the relative locations of these criteria on the map exactly how they are, or are not, related to the item calibrations and person measures.

The variable map is the best way to assemble and picture relevant criteria together with item calibrations and person measures. The map gives us a definitive and detailed picture whereas correlations only indicate the presence of some general relationships.

FIT VALIDITY

The second type of validity has to do with response pattern consistency for items and also for persons. This kind of validity comes from the fit of the observed person-item responses to a useful definition of measurement and hence to the estimated values of item calibrations and person measures. Although the necessity of this kind of response performance validity for persons and items was explained and satisfied by L. L. Thurstone in the 1920's (i.e., Thurstone & Chave, 1929), it is not mentioned in the Standards.

Item and person fit statistics are always necessary. The absence of fit statistics implies the absence of a model for what we expect - a lack of awareness of what we are trying to do. If we do not know what to expect, we cannot hope to explain what happened or know how to use the results.

Point-biserial coefficients (conventional item discrimination) have been used as item fit statistics for decades, though few practitioners have much idea as to what the statistical model for point biserials might be or what that signifies for the interpretation of their data. No one knows what size coefficients to seek or to act on.

When we take as the working requirement, however, that item responses shall be summarized by right answer counts i.e. raw scores, then we can deduce from that ubiquitous practice the necessary and sufficient measuring model. That necessary model is the Rasch model in which person raw scores or percent corrects can be used as the sufficient statistics for estimating person abilities and item p-values can be used as the sufficient statistics for estimating item difficulties (Andersen, 1977).

The mathematical form of the measurement model is deduced from the canonical requirements for measurement (Wright & Stone, *Deducing the Measurement Model*, Chapter 4). The measurement model specifies how to apply these measurement requirements to the data. It specifies what kind of relationship must be approximated between the observed data and the estimated measures in order for valid calibrations and measures to result.

ITEM FIT

When a Rasch analysis is made of item response data, it follows naturally to analyze the extent to which each person's response to each item fits the Rasch model expectation. An item fit statistic is calculated for each item. This summarizes the extent to which the sample's pattern of response to that item is consistent with the way these people have responded to the other items. This gives us "consistency" fit statistics for each item and for each person and also for any subsets of items and persons which might interest us (see Identifying Item Bias, Chapter 8).

The conventional approach to item fit has been the point-biserial correlation coefficient. Item misfit is thought to be indicated by a low point-biserial. It is equally true, however, that a high point-biserial coefficient can also indicate item misfit. This dilemma was identified by Loevinger (1947) as the attenuation paradox (Tucker, 1953; Andrich, 1982).

PERSON FIT

Although hardly anyone computes a person point biserial, the motivation to do so is even greater than the motivation to compute item point biserials. Of course, the attenuation paradox applies to person, as well as, item data. In a Rasch analysis a person fit statistic is calculated for each person. This fit statistic summarizes the extent to which that person's pattern of performance on the test is consistent (or inconsistent) with the way these test items are usually used by people responding to them.

When a person does some lucky guessing and so manifests some unexpected right answers on items that ought to be too hard for that person, we may doubt the validity of their performance and hence question the meaning of their score and measure. How much of the score tells us what they know and how much tells us that they are lucky guessers? When we examine the particular items on which they have failed, we may conclude that their score contains some lucky guesses and is thus misleading and hence somewhat invalid. At this point, however, our attempts to measure these "lucky guessers" need not cease. Our measurement model enables us to know what we are doing. We can use its fit statistics to identify the lucky guesses and its item-free estimation procedure to re-measure the "lucky guessers" on the basis of their answers to the items on which their right answers were not lucky guesses.

SUMMARY

Rasch measurement helps us to see that there are two, and only two, types of validity that can be evaluated from item response data: (1) the ordering of items and persons and (2) the fit of items and persons.

Order Validity

1.1 "Meaning" validity from the calibration order of items. This implements the *content* and *construct* validities of the Standards.

1.2 "Utility" validity from the measurement order of person characteristics. This implements the *criterion* validity of the Standards.

Fit Validity

2.1 "Response" validity determined from the discrepancy between a particular response and its expectation. This identifies individual observations the values of which contradict their use in the estimation of useful measures or calibrations.

2.2 "Item Function" validity determined by an analysis of the validities of the sample of responses to that item, i.e. item fit. This identifies for review and revision items which may not be working the way we intend them to.

2.3 "Person Performance" validity determined by an analysis of the validities of the responses of that person, i.e., person fit. This identifies for review and diagnosis person's who may not have taken this test in the way we expected them to.

21. SCORING MODELS

Social science data collection uses questionnaires composed of items constructed to elicit informative responses. The most common format asks respondents to select from among a set of alternative categories. The construction of useful alternative categories is essential to the success of the items. Once categories have been analyzed in order of increasing "strength" of response, a scoring model must be used to bring responses into a measurement construction process. The analysis of alternative scoring models is the focus of this primer.

Item formats for such instruments are of two general types:

1. Dichotomous responses such as

Agree	Disagree
1	0

and its variants.

2. Polytomous responses such as

Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree
4	3	2	1

and its variants.

These choices begin as discrete categories at a nominal level of measurement. Dichotomous choice of categories such as agree/disagree requires that such alternatives be mutually exclusive with no possibility that selecting one category overlap with another category.

The categories, when polytomous, usually have an intended order that ascends or descends across the alternatives. The Likert scale (Likert, 1932) is the most familiar form of instrument design.

Although polytomous responses begin as categories, they are usually intended as "ordinal" by the direction implied in the labeling of the alternatives.

The most informative ordering of categories, however, is not always as originally intended. Sometimes there is ambiguity about the order and position of the categories. Sometimes the categories are carelessly worded or implausible. This produces an experienced order which is different from what was intended. Consider this arrangement taken from U.S. News & World Report dated November 7, 1994:

How do you feel about Bill Clinton?

		Code A	Code B
Hopeful	30%	5	2
Disappointed	19%	2	0
Disgusted	16%	1	0
Uncertain	14%	3	1
Neutral	9%	4	1
Enthusiastic	7%	6	2
Angry	5%	0	0

There is no single, obvious order to this sequence of seven categories. From "neutral", the sequence moves downwards to "enthusiastic" and finally to "angry" or upwards to "disgusted", "disappointed" and then "hopeful"! If this is the sequence that occurred in the respondents' questionnaire, it is quite possible that some respondents became confused. If they did not look carefully at all categories, they might not consider all the alternatives available.

Codes A and B suggest two possible "scorings" for these categories. The use of either of them could give a sequential order better than the arrangement presented.

It is always necessary to determine the extent to which the categories were used by respondents in the way they were intended. Respondents often interpret questions and categories differently from the way in which they are intended. Confusion can also result from ambiguous directions.

Preliminary analysis must compare the intended order of alternatives and respondent behavior. The data cannot be accurately interpreted until a useful scoring model has been determined. This is the reason pilot studies are recommended. The best pilot study investigates the category ordering in the scale before it is used to gather the main data. In any case, a scoring model must be found before the main analysis can be undertaken. Some researchers immediately proceed to code the categories and analyze the responses before ascertaining the most useful scoring model. This leads to misinterpretation of data.

A full scale analysis consists of examining every reasonable scoring model to determine which one produces the most useful results.

Consider the Fear Survey Schedule (FSS) by Wolpe & Lang (1969). The FSS has 108 items. Each item is to be answered on a five-point scale:

0	1	2	3	4
Not At All	A Little	A Fair Amount	Much	Very Much

There are fourteen additional scoring models to be considered.

	FSS CATEGORIES				
Original	0	1	2	3	4
Alternative 1	0	0	0	0	1
Alternative 2	0	0	0	1	1
Alternative 3	0	0	0	1	2
Alternative 4	0	0	1	1	1
Alternative 5	0	0	1	1	2
Alternative 6	0	0	1	2	2
Alternative 7	0	0	1	2	3
Alternative 8	0	1	1	1	1
Alternative 9	0	1	1	1	2
Alternative 10	0	1	1	2	2
Alternative 11	0	1	1	2	3
Alternative 12	0	1	2	2	2
Alternative 13	0	1	2	2	3
Alternative 14	0	1	2	3	3

With 15 different scoring models, the question is, "Which one works best?"

We must study how the people used these categories. This can be done by examining the results for each of the alternative scoring models. Statistics from Rasch analysis provide explicit information about which scoring model is most useful with these data.

The Rasch statistics for the analysis of scoring models are:

1. The item and person separation statistics which are ratio scale equivalents to person and test reliability.

2. The item and person unbiased or "adjusted for error" standard deviations.

3. The fit statistics computed for item, person, and scoring category.

4. Finally, if we have supplemental information, we can go beyond the data and evaluate which scoring model best separates known groups of persons.

We evaluate the differences according to their standard errors and determine whether these differences are significant. But we need to use the model standard error as modified by misfit to accomplish this task.

When a scoring model shows considerable person misfit, this must be taken into account so as to produce a measure error that is increased by the amount of misfit. Any scoring model that increases the person separation statistic or the adjusted standard deviation is more efficient. Increase in either of these statistics gives us an index by which to judge the efficacy of scoring alternatives.

These statistical tools provide explicit ways to evaluate which of the alternate scoring models is most useful.

Category reduction, from more numerous categories to less, frequently provides a more efficient scoring model. Many researchers plan more categories than are used by respondents or useful to define the variable. Too many options are the result of idealized expectations rather than real experience. Fewer categories are often more efficient. Multiple categories are often "nonexistent" and can be modeled more effectively by dichotomous scoring.

We analyze all 15 scoring models of the FSS, using output from *BIGSTEPS* (Wright & Linacre, 1992) to determine which model is most useful.

Table 21.1 gives the data for the 15 scoring models. Column 1 gives the model code. Column 2 indicates the steps in the model. The alternative models are arranged by step. Column 3 gives the number of iterations (UCON) to a converged solution. Columns 4 and 5 give the standard errors for items and persons. Columns 6 and 7 give the person and items infit statistics. Column 8 shows the number of items identified beyond a standardized misfit statistic of 2.0. Columns 9 and 10 give the person and item separation statistics. Columns 11 and 12 give the item and person quotients resulting from the ratios of item and person errors to their standard deviations (Column 4/Column 6 and Column 5/Column 7).

Scoring Model #8 has been highlighted to identify it as the most efficient one. Columns 11 and 12 show this model to have produced the highest values for the adjusted ratio previously described. The person and item reliabilities for this model are 0.97 and 0.98 which are as good as for any other model. The number of misfitting items, while not the lowest, is less than for 11 of the other models.

This model contains only one step, thus indicating that the FSS can be scored efficiently as dichotomous. The more detailed ratings supplied by the authors do not correspond with how the respondents view the scale. Simple identification of fear is sufficient and attempts to discriminate further are unproductive. The FSS functions efficiently in this mode and provides a model that is consistent with the data.

Our example illustrates the need to investigate the possible scoring models in any scale before analyzing the data further. Failure to take into account the influence of scoring model choice confuses subsequent data analysis.

Table 21.1

Wolpe Fear Scale

Scoring Models Analysis

By Mark Stone

1	2	3	4	5	6	7	8	9	10	11	12
SCORING MODEL	STEPS IN MODEL	UCON # ITERATIONS	ISEP	PSEP	IINSD	PINSD	# Items Out	ISEPR	PSEPR	ISEP/ IINSD	PSEP/ PINSD
		[T.0.2]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[C4/C6]	[C5/C7]
1 SM 00001	1	4	1.9	1.6	0.13	0.07	4	0.78	0.71	14.62	22.86
2 SM 00011	1	4	3.3	2.6	0.12	0.12	8	0.92	0.87	27.50	21.67
4 SM 00111	1	3	5.4	4.0	0.11	0.16	11	0.97	0.94	49.09	25.00
8 SM 01111	1	4	7.4	5.3	0.09	0.17	9	0.98	0.97	82.22	31.18
3 SM 00012	2	8	3.2	2.3	0.17	0.27	8	0.91	0.84	18.82	8.52
5 SM 00112	2	3	5.2	3.8	0.16	0.30	11	0.96	0.94	32.50	12.67
6 SM 00122	2	8	5.2	3.8	0.13	0.25	10	0.96	0.93	40.00	15.20
9 SM 01112	2	20	7.4	5.5	0.18	0.40	15	0.98	0.97	41.11	13.75
10 SM 01122	2	9	7.5	5.8	0.18	0.35	20	0.98	0.97	41.67	16.57
12 SM 01222	2	7	7.9	6.0	0.14	0.30	21	0.98	0.97	56.43	20.00
7 SM 00123	3	14	5.0	3.6	0.19	0.36	14	0.96	0.93	26.32	10.00
11 SM 01123	3	10	7.2	5.4	0.25	0.51	26	0.98	0.97	28.80	10.59
13 SM 01223	3	4	7.7	5.7	0.19	0.44	21	0.98	0.97	40.53	12.95
14 SM 01233	3	10	7.6	5.7	0.19	0.40	22	0.98	0.97	40.00	14.25
15 SM 01234	4	15	7.4	5.5	0.25	0.51	22	0.98	0.97	29.60	10.78

22. DIMENSIONALITY

Developing measures from constructs is difficult because the properties of what we propose to measure are complex and interrelated. But these difficulties must be faced. How to proceed?

The first task is to consider how to deal with behavior. While we acknowledge that behavior is complex, we also recognize that we cannot advance the process of measurement by simultaneous attention to all aspects of complex behavior. There is no useful way to make measures of more than one variable at a time.

We have to isolate a single variable (dimension) and then develop it to the best level possible. When our thinking is muddled by considering two or more aspects of a variable simultaneously, then we only become confused.

We may recognize the multidimensionality of experience, but this multidimensionality cannot be addressed as a whole. For knowledge to develop, complex behavior has to be decomposed into single dimensions.

We begin by specifying a dimension as a variable. A variable is a single, unidimensional concept abstracted from the complexity of human behavior. Its successful abstraction results from a dialogue conducted between the abstract idea isolated and formed into a single concept and the wealth of sensory experiences that constitute the "real world." The former is conceptual and consists of the abstract idea we have gleaned from our experiences. The latter are elements of experience that have substance and constitute reality. The need for dialogue is to make our abstractions relate meaningfully to the real world. Our variable cannot be so abstract that it is devoid of reality. Nor can we clutter our thinking by such a bombardment of experiences that we cannot abstract a singular essence that stands out and separates the variable from all other experience.

This dialogue is not something which is done once and for all. The process continues indefinitely, inspiring a progression of further refinements.

As we satisfy the need for a single concept, it is necessary to recognize that the variable is posited as the unifying element between idea and experience. The variable is the focal point between experience taken from the real world and the abstraction of an idea. The idea of a variable embodies the projection of a line or arrow indicating the direction of "more" along the variable. The variable, i.e. line, is the representation of the experience abstracted and conceptualized as on a line. The experiences themselves are illustrations or examples taken selectively to arrange in a systematic fashion. The order in which they are arranged on the variable is the correspondence between our idea and our experience.

Figure 22.1

The variable.

Figure 22.1 shows the idea of the variable as an arrow. What we locate along the variable are illustrations from experience that embody the unidimensional concept we have in mind.

Consider simple experiences from life like length and weight. A variable of length can be "seen" by using what we now know as a ruler. It begins with an implied zero at one end, the numerals along the ruler (variable) signifying equal units marked out as we proceed to the right. If we consider weight, then the units signify whatever units we select, say pounds. Each unit indicates an increase of one pound as we move across the variable.

The ruling idea of the variable needs to be fully impressed upon our mind even though the examples may seem trivial. When we proceed to construct more sophisticated variables, especially those that we do not fully understand, it is important to have our methodology clearly understood so that we will not become confused by experience or overwhelmed by complexity.

The idea of a variable is always an abstraction, a simplification of what we experience as reality. If we could think about all of reality at once, there would be no need to abstract. But when we want to determine its essence, then it is important to decide exactly what we want to "see" in the experience, what is useful for us to think about. It is also important to note that we must disregard all other aspects in the pursuit of this goal. We do not disregard these other considerations because they are not important, but because they are divisive to the task at hand.

Suppose, in going through the checkout line of a large supermarket, we place groceries on the checkout conveyor. The people in front of us are also checking out, and at the front of the line a man is paying for his purchases. If the customer or clerk departs from the process of checking, receiving and giving change, the entire operation comes to a standstill. Suppose in giving change, a clerk notices a rare two-dollar bill. "Do you have more of these?" she asks. The process of checkout breaks down as the process of "collecting rare two-dollar bills" takes precedence. How will you feel during this time while the "collectors" engage in their discussion? What will the manager do?

Suppose the clerk is more interested in checking all the cash and silver for rare currency and coins for his collection than in making change and rendering service. Does this mean that coin collecting is wrong or not useful. Of course not, but in a busy supermarket you cannot mix the two processes without bringing the system to a standstill. Likewise with measurement. If you cannot identify a variable and focus upon it, then it will be impossible to achieve success because you will become distracted by all of the additional aspects that are possible to study.

When we identify one aspect for study, it is not because we believe that the other aspects are unimportant, it is because we cannot focus upon one aspect unless we treat the others "as if" they were not relevant. We know that they have impact, but we cannot consider their input relevant at this time. When we try to make measures by addressing all matters at once, it becomes impossible to sort out what is occurring.

It is important to distinguish between the procedures for developing measures and that of studying the relationship between measures. The former is a task of measurement, i.e. of building a variable. The latter is a task of statistical analysis, i.e. determining the relationship between variables. The statistical process can overwhelm the measurement process, if we do not pay attention to what problem we are addressing and inadvertently get the cart before the horse.

MULTIDIMENSIONALITY

The way to proceed in understanding multidimensionality is first to construct unidimensional variables upon which to make measures and then to evaluate the relationships among the measures generated from these singular variables.

It is often said, in explicating the idea of "multidimensionality," that the variables must embody all of the behavior that can be expected to be observed in the experience. This is not the way to proceed. The problem is not to see how complex we can make the experience, but instead, how simply we can abstract from an experience an essence of what we want to make observations on and measure. What is a single dimension that we want to construct as a variable that will allow us to focus successfully on that one aspect of experience? When we realize that there are other aspects left over from building a variable, then the next task is to get started on a new adventure - building another variable. The process goes on and on.

It is not the goal of measurement to be multidimensional. Measurement can only address one aspect of experience at a time. If we do the job of building a variable well and can quantify it usefully, then we can proceed to do it again and again with the construction of additional variables.

In social science, the word "multidimensionality" usually implies that there are multiple dimensions, i.e. separate dimensions. But consider volume. We have three applications of the same dimension, not three "different" dimensions. Length, width, and height, measured in the same units for utility, give us the three "applications" of a single dimension needed to compute volume.

THE ORGANIZING PRINCIPLES OF VARIABLE CONSTRUCTION

Variables are constructed out of experience. From experience, initially encountered, we notice similarities and differences. The first step is to be able to segregate from the experience some single aspect that can be found in each instance of observation, some element that can be abstracted from each instance. This abstraction is the unifying idea from the experience. It is the first step in the construction of a variable. The unifying element is this single idea. The variable signifies focus upon a single aspect of experience, an experience elements of which are useful to focus upon and so become the basis for a variable.

The next step is to illustrate this aspect of experience by a scheme of graduated experiences. That is, have we found a variable that can be abstracted usefully from the experience?

We observe a large pack of horses. Observation suggests a way to group them. They can be arranged by height. Markings may also allow us to re-group them according to whatever characteristics are suggested. It can be by their behavior in running or according to some other skill. The same pack of horses can produce many variables to use in describing animals. And each abstraction of this complex experience is itself a simple and single representation of an idea that may be useful to subgroup the experience.

It frequently follows that the invention of a good variable eventually becomes useful to everyone. This is the hallmark of a good variable. Having become an extraction of experience, the newly constructed variable is recognized as so self-evident that everyone begins to use it. Some may even remark, "Why didn't I see that." The analogy is like an optical illusion, when the alternatives are pointed out, they can usually be seen, but before that we tend not to observe the alternatives. A good variable operates in a similar way. The variables, so identified, become recognized and useful. This is the history of good variable construction.

Application of the Rasch model and associated fit statistics can be used to identify items that define a single dimension. Scale development proceeds by successive variable definition. Items that fit contribute to a single, interval, sample-free scale. Locations of items, persons, and related attributes produce a definitive mapping of the variable. See Variable Mapping, Chapter 14, page 119.

23. QUALITY CONTROL PLOTS FOR EVALUATING CO-RELATIONS

This chapter describes a simple graphical method for studying the relationships between pairs of person measures, pairs of item calibrations or any other pairs of values intended to display a co-relation. This method replaces the elusive numerals of correlation and regression coefficient(s) with easy to see pictures for describing relationships between pairs of values.

PROBLEMS WITH CORRELATION AND REGRESSION COEFFICIENTS

The traditional approach to investigate a relationship between pairs of values has been to calculate a correlation coefficient and stop there. The single number which results is used to describe the relation between the paired values. But how can any single number do justice to a potentially complex relationship so completely and so fully that no further information is of interest?

The correlation coefficient attempts to summarize in one number all the information contained in all pairs of values under consideration. The single resulting number is only occasionally evaluated in terms of its standard error and even less often dis-attenuated for the measurement error inevitably contained in the estimation of the pairs of values compared.

Dis-attenuation for measurement error should always be done for every coefficient used to quantify a relation between pairs of values. Dis-attenuation for measurement error is called for in every regression analysis. And there are further problems to consider. The usual regression analysis assumes there is no error in the independent variable(s) and that error in the dependent variable away from the modeled relation is entirely random and the only error expected. But usually the independent variables themselves are estimates containing their own error of estimation.

Relationship analysis needs to identify, separate and separately consider:

(1) *modeled error*, the explicit stochastic part of the relational theory implemented by the regression analysis,

(2) *measurement error*, an unavoidable part of all values in the analysis which depend on a prior estimation procedure, and

(3) *model misfit error*, the discrepancy between the general theory modeled and the particular data which is being examined for the extent to which it constrains or contradicts the theory modeled.

These misunderstandings occur whenever simple correlation or regression coefficients are accepted as sufficient summaries of relationships. These single values give only the barest and most incomplete description of the situation. They are based on the presumption that nothing is happening in the data except a simple linear relationship between two exactly known variables which can be captured by one coefficient as a single value. To presume this condition is to specify in advance that all people or items whose pairs of values are used to compute the correlation are nothing more than random examples of a single, simple linear relationship.

To say this again: When we rely upon a correlation coefficient to convey all that is operating in the relationship between a set of data points, we are reducing all the people or items examined to the status of equivalent examples of whatever the single, simple linear relationship is determined to be. Every person or item is reduced to being exchangeable in demonstrating the one relationship presumed in the data.

This conceptual reduction is not only never true but also never useful. Reducing any relationship to a single number denies and conceals all of the interesting individual behavior occurring in the data. The reduction prevents any realization of the diagnostic capacity of the data. To routinely discard this rich potential is not good science.

INVESTIGATING CO-RELATIONSHIPS USING PLOTS

We give each pair of values its own identity when we plot their location. Instead of reducing the data to a single correlation coefficient, the paired instances of the two variables are plotted against each other so that every data point represents a relation between the paired values - a relation that invites further investigation before summarizing.

Every plotted point should be clearly labeled so that unusual points can be examined to determine their specificity. When a point represents two measures on two variables for a particular person, that point is specific to that person. If the person is better in spelling than in arithmetic, their data point is uniquely informative about that aspect of that person. This is quite beyond and far more interesting than any general correlation which may be observed between spelling and arithmetic.

The first step in addressing the problem of co-relation is overlooked when the plot of paired values is not drawn, not labeled and not carefully studied for the particular identities of unusual points.

Some people find it difficult to examine a plot by inspection. They do not derive benefit from a simple examination because the plot is not set up in a way that tells them a story about what might be seen in their data.

LABELS FOR PLOTTED POINTS

In order to interpret plots we need to enable the plotted points to bring out the purpose of the plot and to make the story contained in the points immediately visual. Careful attention to labeling enables us to make visible the idiosyncratic and diagnostic possibilities in the data.

It is essential to label each point with a label that identifies in each plot what each point stands for i.e. male (M) or female (F), black (B) or white (W), married (m) or not (n). We can't investigate whether points are as expected or discover a pattern, if we cannot see what the points stand for. If we discover clusters of points we need to see on the graph what characteristics the clustered points share and do not share. This means we will replot the same points but with differing label sets to bring out the dominant patterns.

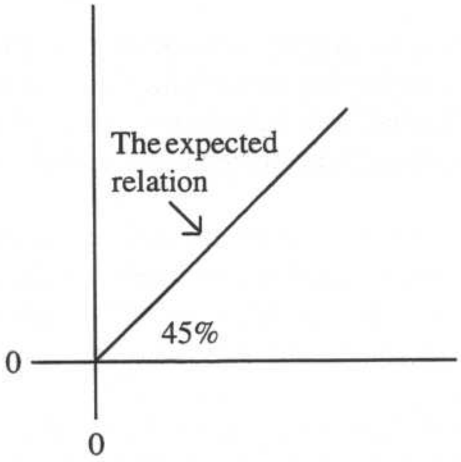
We need ways to label points so that any organization they manifest will be immediately apparent, so that we can see what the points indicate. The labeling of points must be as comprehensive and as versatile as possible. When labeling clutters up the plot or becomes too extreme to show on the plot itself, then a code number can be given each point and an accompanying legend (located next to or on the plot) constructed so that the points can be quickly identified and their pattern understood.

The more comprehensive the legend, the more assistance it will provide in investigating the nature of the plotted points. Graphical notations printed in position on the plot, however, communicate more quickly than text in a legend. Thus it is useful to develop versatility in successively altering point identification and replotting the same data so as to bring out the main patterns contained in the data by making them visible on the graph itself.

IDENTITY LINES

Labeling data points and providing a legend are not enough. We must go further and draw into the plot a line (or curve) that represents the main question to be investigated by these data - the main question which the plot is intended to answer. This "identity (of the question asked) line" should be a smooth, preferably (with data transformed so that it becomes) straight line drawn so that it marks the hypothesized path of the presumed relationship between the two variables.

To expedite the visual interpretation of any plot, it is important to adjust the scales of the horizontal and vertical axes so that the resulting "space" revealed is square. When this adjustment achieves a complete equation, the simplest version of this identity line goes through the origin with a slope of one, proceeding at a 45 degree angle across the plot from lower left to upper right and indicating a positive relationship between two variables on the same scale.



This simplest identity line specifies that the two sets of values are intended to measure the "same" thing from the same origin on the same scale: inches-to-inches, pounds-to-pounds or the logits-to-logits of commonly calibrated items.

If, in a study of item bias, we co-calibrate items to a common scale, we can plot pairs of item calibrations and use the identity line to model "no bias" between the two calibrations. The line shows which item points do not fit the "no bias" hypothesis represented by the identity line and hence which items require further investigation.

Usually the two values plotted originate on somewhat different scales. For pairs of measures, origins and scales are usually expected to be different. Then, a useful representation of the hypothesized relationship may be a different kind of identity line that passes through the means of the two sets of values with a slope equal to the ratio of their standard deviations. Again the best choice of horizontal and vertical scales is one that makes the resulting plot fill out a square.

An appealing and seemingly equivalent approach is to standardize the values of each variable by subtracting a mean *M* from each value *X* and dividing the difference by a standard deviation *S*:

$$z = \frac{X - M}{S}$$

When these standardized *z* values are plotted, the hypothesis bearing identity line once again goes through the origin with a slope of one.

The shortcoming of this standardization is that it draws our attention away from the metric(s) of the original variables. It is seldom useful to forget what the original metrics stand for. That metric information can be a key to understanding the data plot.

Thus it is usually more informative to retain the original metrics of the variables and not to standardize. That places the hypothesis bearing identity line through the intersection of the means with slope determined by the ratio of the standard deviations.

THE HYPOTHESIS REPRESENTED BY THE IDENTITY LINE

The identity line represents the hypothesis of a perfect relationship. The utility of the identity line is that it guides the eye in examining the data points with respect to the hypothesis. We can see which data points are close to the identity line and which points are far from it and thence indicative of a particular and identifiable digression from the perfect relation hypothesized.

The deviations are the exceptions, the unexpected digressions from the perfect idea indicated by the identity line. The identity line also guides the eye to locations where no data points exist. The data points which follow the identity line confirm our expectations. The data points that deviate contradict our expectations. The data points that are missing show us where we are uninformed.

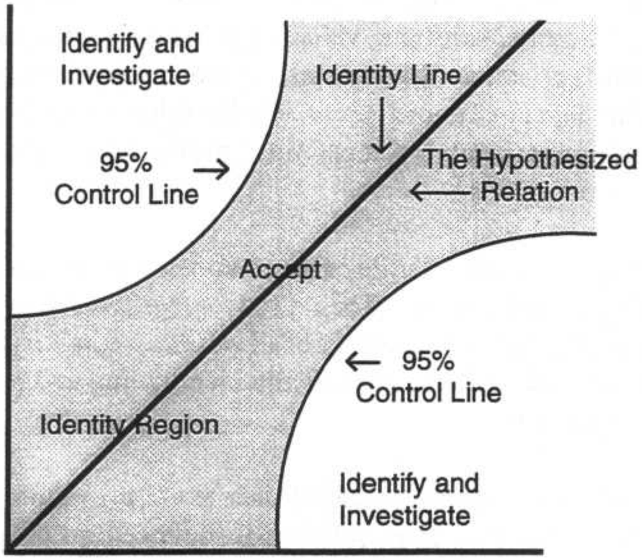
The statistical model used with most correlations is a null hypothesis of zero correlation between the two variables. But when we model a relationship, the relevant null hypothesis is seldom zero but rather a perfect relation as close to one as measurement error allows. This more useful "null hypothesis" of perfect relation is the one relevant to measurement analysis.

CONSTRUCTING QUALITY CONTROL LINES

How can we show the extent of expected error in a plot? How can we make allowance for measurement error visible in the plot? How can we visualize error dis-attenuation? The answer is to draw in quality control lines to guide inspection of the data plot and to provide guidelines for seeing how close, statistically speaking, our estimated points are to the identity line, given their errors of measurement.

These error guidelines are constructed in the same way as the statistics used in industrial quality control. We draw two boundary lines, one above, one below the identity line, to guide inspection of data points. These lines make the statistical boundaries of our hypothesis visible.

We usually construct this pair of boundaries so that they enclose 95% of the data points which measurement error around a perfect relation would produce. These boundary lines enclose a region containing two standard errors of measurement around the identity line in each direction:



Quality control lines enable visual evaluation of the data points. They show us the identity "line" and the identity "region"; the area around the identity line in which it is reasonable for data points to occur, given the measurement error.

Data points which fall within the control lines can be accepted as statistically equivalent to the identity line and hence to the hypothesized relation. These data points do not contradict the hypothesis represented by the identity line.

Data points which lie outside the control lines are, however, instances which contradict the identity line i.e. the hypothesis. Each outlying point is a visible contradiction to the hypothesis and consequently each outlying point needs to be identified and investigated in order to understand and explain what has been observed, in order to discover the meaning of the contradiction.

If, when studying a sample of people who have been measured on two variables, we find that their paired measures follow an identity line, then the paired measures are clearly on a single variable and the two initial variables are empirically co-dimensional, at least for these people.

MEANS AND STANDARD DEVIATIONS

Even when two variables are both conceptually and empirically co-dimensional, there will still be some individuals for whom the relationship does not hold, some exceptions. Inclusion of these deviant values in calculating means and standard deviations for these data, however, disturbs these two commonly used reference statistics.

We want to determine the extent to which the data follow the line which asserts and/or supports our intended hypothesis. To make this determination we begin by evaluating all of the data points in terms of our theory.

Without any theory to guide our observations we are only fishing for something we cannot yet describe. This is not research but blind groping. While there may be times when we find ourselves perplexed by a measurement problem, that confusion is neither optimal nor scientific.

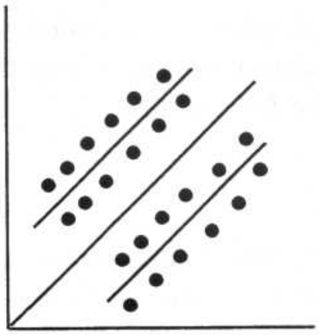
Outliers are contradictory data points. They become unusual in the light of our expectations and so in need of investigation. To include outlying values together with those data points that confirm the hypothesis in computing means and standard deviations is to remain confused by our own data. Means and standard deviations are vulnerable to extreme values. Outliers distort the conclusions we come to. We want robust statistics that are not unduly influenced in central location and dispersion by idiosyncratic, extreme values.

Statistics like the median and interquartile range are sometimes advocated as useful because these statistics are less influenced by extreme values. Their disadvantage is that they lack precision and power. What we want is not the mean and standard deviation of all values, exceptions included, but the mean and standard deviation of just those values which follow the identity line and hence do not contradict the hypothesis of a shared dimensionality.

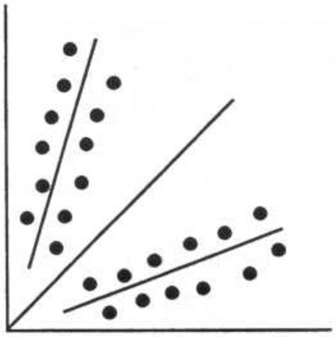
As we survey a plot we need a convenient and consistent way to exclude the outliers. When deviant data points are identified we want to recompute means and standard deviations without including these deviant points and then to re-draw the identity and control lines so that they represent only the points of the subsample of people who confirm the hypothesis of a general relationship and not those of the people who contradict it.

This does not mean that we throw the other data away. On the contrary, it is important to investigate all the data, and most especially the deviant data points. But, it is necessary to determine what criteria is to be used in making the decision concerning deviance. If our hypothesis is depicted by an identity line, then statistically significant deviations are those data points beyond the quality control boundary lines. These values, then, because they are different, do not belong when calculating the summary statistics used to locate the identity and control lines.

Usually when data plots are examined it is easy to see whether the points are following a line. Sometimes we see two groups of points that follow two lines:



or



No statistic can determine which of the two lines we should use as our expectation, our intended hypothesis, and which to consider as deviant, however. We must identify the data points involved and then engage in the hard work of thinking clearly about our intentions and how they emerge in the data plot. That is the only way to determine what reasonable hypothesis they support.

An average, even when successively adjusted for deviations is, nevertheless, only an estimate of central location. No automatic strategy can provide the detective work and speculative inspiration needed for creative analysis. No automatic technique can substitute for patient investigation, visual inspection and careful thought. No automatic process can examine the data points in such a way as to replace intelligent review. The evaluation of unexpected data points requires the combined efforts of data analyst and content specialist so that each can encourage the other to investigate all possible hunches concerning the patterns manifested in the plot.

Two other chapters provide examples of this: Chapter 8 (p.57), Identifying Item Bias and Chapter 9 (p.65), Control Lines for Item Plots.

24. GUESSING

What to do about guessing on multiple-choice (MCQ) test items has been a hot problem for 70 years. For some psychometricians the introduction of an extra item parameter for guessing is the way to settle the matter. We review their approach and show how guessing can be better dealt with - detected, diagnosed and managed - by the methods of Rasch measurement.

Webster says "to guess" is to form an opinion from *little or no evidence*. That suggests that when people guess on an MCQ test item, they decide on the basis of little or no ability how to answer. When they are lucky, they guess correctly. It follows that to count lucky guesses as manifestations of ability produces confusion, especially when these counts are combined with correct answers which are the outcome of applied ability. Haphazard combinations of accidental and informative outcomes are bound to be misleading. Webster's also says that "to guess" can mean to arrive at a correct solution by conjecture or intuition. Synonyms for "to guess" include; "to suppose," "to hypothesize." Another alternative definition, "stochastic" from the Greek stokastikos, suggests knowledge arrived at in a probabilistic manner.

Guessing is not done by items, but by persons. When a person, with no knowledge of the correct answer, guesses at random on a multiple-choice question with five alternatives, the probability of success might be as low as $P = .2$. If, however, ability enables the person to eliminate three alternatives as incorrect and hence to reduce the guess to one of two choices, then the probability of success might increase to $P = .5$.

These considerations leave some psychometricians content to deal with guessing as an item parameter. For us, however, the same considerations make clear the ultimate futility of attempting a psychometric solution based on test item characteristics. There are too many personal causes and consequences in guessing for any item guessing parameters to manage.

Guessing can only be addressed and managed by allowing for all of the factors, external and internal, which provoke a person to guess. The most important external factor is the intended use of test results. Internal factors include test administration directions, test format and time allowed. When a passing score allows one to acquire a license to practice a remunerative profession, but a failing score prevents this, a person's approach to a test is different than when the outcome offers no immediate advantage. The uses of test results influence examinee behavior. To expect a psychometric model to resolve these personal influences on the test item or person parameter level is unreasonable.

EXTERNALFACTORS

Guessing provoked by external factors can only be managed by addressing these factors in their own terms:

1. *Reduce the use of multiple-choice items.* Although this item format enables simplified answer sheet scanning, writers of multiple-choice items seldom overcome the excessive restriction this approach puts upon item construction. Multiple-choice items invite some persons to guess.

2. *Invent better methods of questioning* which eliminate guessing as an active possibility. The use of open-ended questions is one alternative. Providing long, rather than short, lists of possible answers discourages guessing. The versatility and capacity of modern scanners and computers can handle response patterns far more complex than the familiar simple rows of five choices.

3. *Qualify the use of test results* so that they do not force examinees to corrupt their test behavior in order to survive.

4. *Do not administer items that are so hard* that they provoke guessing as the only resort.

5. *Do not make speed a factor in testing.*

MISTAKING GUESSING AS AN ITEM PARAMETER

Psychometricians who deal with guessing as an item parameter argue that better measures result, but is this true? We know that the factors which influence test behavior produce responses to items that consist of idiosyncratic mixtures of ability and guessing. But, if they are combined in some responses but not in others, how can we untangle these components to determine which items have been answered by guessing, and by how much guessing and which have not?

An item guessing parameter assumes that it is the item that causes the guessing and that the effect is the same for every test-taker. Even though some items may sometimes seem to provoke more guessing than others, it is the person, not the item, who initiates guessing, whose momentary state of knowledge and urgency governs the possibility of a lucky guess. Even if some guessing could be handled by an item parameter, a person parameter for guessing behavior would also be needed. We know that some persons guess more than others, a few often, most rarely or never. We also know that no one guesses all of the time.

The item parameter approach to guessing raises the lower end of the item characteristics curve no matter who takes the item.

The asymptotic solution in Figure 24.1 forces a guessing penalty on every person who chooses not to guess. It does this (shaded area in Figure 24.1) by misestimating the item to be easier for non-guessers than it actually is.

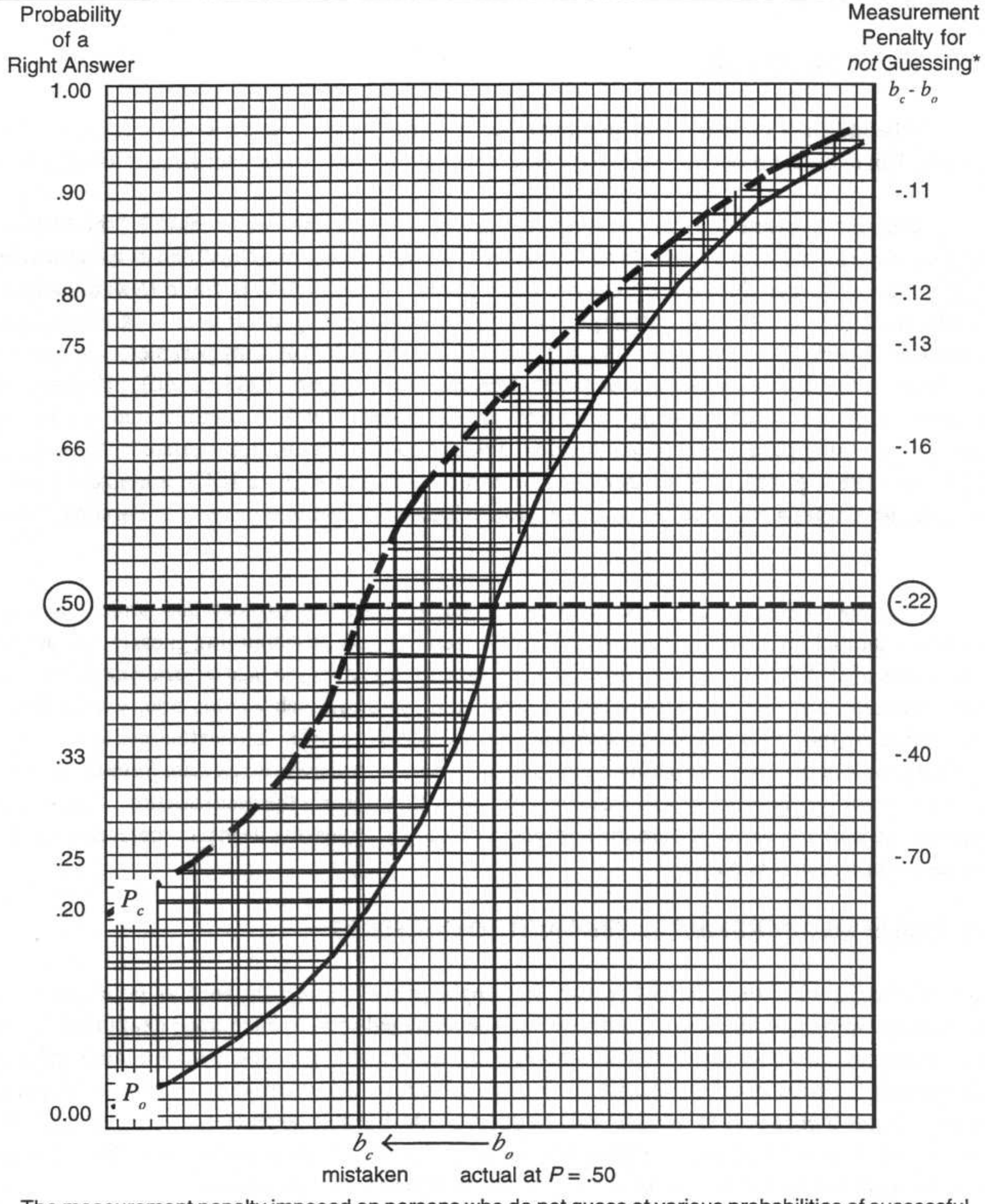
The measurement penalty for not guessing is the distance between b_c and b_o on the measuring variable b in Figure 24.1

$$\left[b_c = d + \log\frac{P_c - C}{1 - P_c}\right] \text{ but } \left[b_o = d + \log\frac{P_o}{1 - P_o}\right]$$

$$b_c - b_o = \log \left[\frac{(P_c - C)}{(1 - P_c)}\frac{(1 - P_o)}{P_o}\right] = \rangle \log \frac{P_o - C}{P_o}$$

when P_o is mistaken in the test score model for P_c .

Figure 24.1
Guessing as a lower asymptote.



* The measurement penalty imposed on persons who do not guess at various probabilities of successful performance.

Guessing, as a lower *asymptote*, modifies the whole curve.

$P_c = C+(1-C)\dfrac{\exp(b_c-d)}{1+\exp(b_c-d)}$ the guesser; $P_0 = \dfrac{\exp(b_o-d)}{1+\exp(b_o-d)}$ the non-guesser

An alternative is to use the lower boundary in Figure 24.2. In this approach there is still a penalty for not guessing, but it is only exacted from non-guessers with performance probabilities below the guessing level C.

ESTIMATION PROBLEMS

If the idea of a guessing item parameter were useful, its application would lead to successful practice. But even the most devoted advocates of a guessing item parameter lament its application.

Attempts to estimate item guessing parameters are uniformly *un*successful, "the likelihood function (of the model with a guessing parameter) may possess several maxima" and its value at infinite ability "may be larger than the maximum value found" when ability is finite (Swaminathan, in Hambleton, 1983, p. 30) and "attempts to estimate the guessing parameter ... are not usually successful" (Hulin, Drasgow & Parsons, 1983, p. 63). "40% of the guessing parameter estimates did not converge even with a sample size of 1593" (Ironson, in Hambleton, 1983, p. 160). "If a test is easy for the group (from which guessing parameters are estimated) and then administered to a less able group, the guessing parameters (from the more able group) may not be appropriate" (Wingersky, in Hambleton, 1983, p. 48). "When dealing with three parameter logistic ICCs, a nonzero guessing parameter *pre*cludes a convenient transformation to linearity" (Hulin, Drasgow & Parsons, 1983, p. 173).

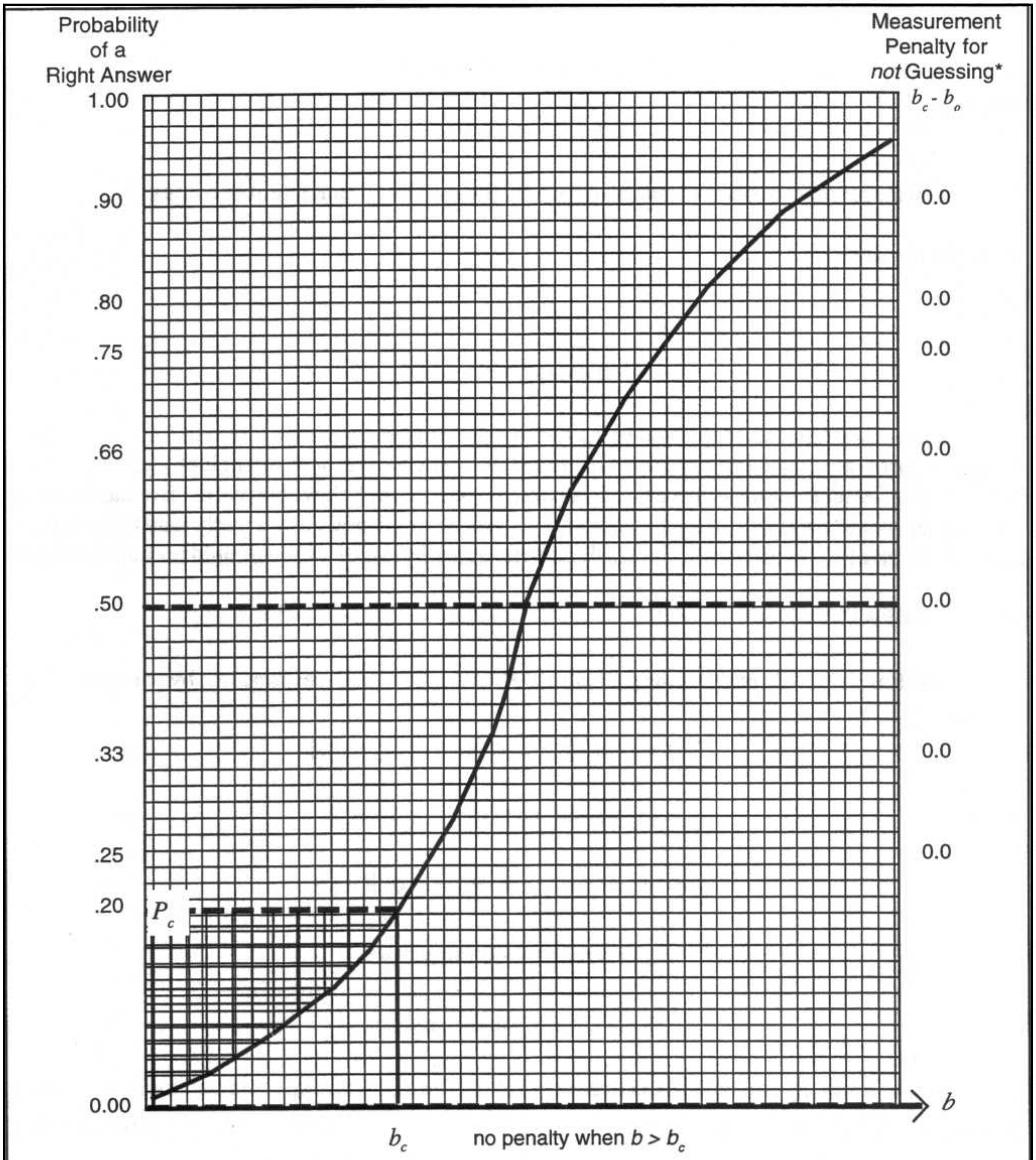
Stocking (1989, p. 41) reports in an extensive study "to explore and understand some apparently anomalous results in various LOGIST-based (a program estimating guessing parameters) applications of IRT that have been obtained from time to time over the past several years" that these same "anomalous results were obtained in simulation studies, such as this one, where data are generated to fit the 3PL (guessing parameter) model" (1989, p. 41). Thus attempts to resolve the guessing problem through estimating a guessing item parameter, *even when data have been created to fit that condition***,** have not been successful. Successful practice is the confirmation of theory. The ubiquitous inability to achieve a practical implementation of a guessing item parameter discredits the theories upon which it is based.

THE RASCH MEASUREMENT APPROACH TO GUESSING

To begin with, the external factors that might provoke guessing such as poor test format, abbreviated timing and threatening purpose must be managed so as to encourage examinees to make their responses as uncontaminated as possible by misleading guesses. Maintaining good test management requires constant attention. Failure to reduce the external provocations to guess is sloppy. The problem needs to be addressed by good test design and careful test administration. What must ***not*** be done is to default to a naive presumption that the problem of guessing can be "washed away" by a slick assumption that an item guessing parameter will do the trick.

Guessing is not avoided in Rasch measurement. Guessing is addressed directly by instituting quality control over all response patterns. Consider a score of five on a 10-item test with items positioned in order of increasing difficulty. Both the probabilistic nature of the model and our everyday

Figure 24.2
Guessing as a lower boundary.



* Guessing, as a lower *boundary* modifies the curve only for $P < c$. No penalty above b_c.

$$P_c - P_o = \frac{\exp(b-d)}{1+\exp(b-d)} \text{ for } P > .2$$

$$P_c = .2 \text{ for } P_O < .2$$

experience with typical response patterns lead us to expect patterns like

$$1\,1\,1\,1\,1\,0\,0\,0\,0\,0 = 5$$

$$1\,1\,1\,1\,0\,1\,0\,0\,0\,0 = 5$$

and, even

$$1\,1\,1\,0\,1\,0\,1\,0\,0\,0 = 5$$

The more improbable the pattern, however, the more questionable it becomes.

Consider the pattern

$$1\,1\,0\,0\,0\,0\,0\,1\,1\,1 = 5$$

or, worse,

$$0\,0\,0\,0\,0\,1\,1\,1\,1\,1 = 5.$$

Our surprise and our objection to the last two patterns are much greater than for the first three. We might speculate that the irregularities in the last two patterns are the result of lucky guessing on the hardest items. After five consecutive wrong answers, it become unbelievable that the five hardest items could be answered correctly on the basis of knowledge. We may not know exactly why this occurred. But we have identified a pattern that is clearly questionable in terms of what we could reasonably expect.

RESPONSE PATTERN ANALYSIS

The Rasch model specifies the probability P_{ni} of dochotomous response x_{ni} by person n to item i to be:

$$P_{ni} = \exp[x_{ni}(b_n - d_i) / [1 + \exp(b_n - d_i)]$$

where

b_n = the ability measure of person n

d_i = the difficulty calibration of item i

24.1

$x_{ni} = 0$ for an incorrect answer

$x_{ni} = 1$ for a correct answer.

Estimates of P_{ni} can be used as expected values for x_{ni}. The expected variance of x_{ni} can be estimated by $[P_{ni}(1 - P_{ni})]$. To estimate a standard residual z_{ni}, we subtract from the observed x_{ni} its expected value P_{ni} and divide by $[P_{ni}(1 - P_{ni})]^{1/2}$ its binomial standard deviation to get

$$z_{ni} = (x_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})]^{1/2} \,. \qquad 24.2$$

When the data approximate the measurement model we expect this estimated residual z_{ni} to be distributed symmetrically with a mean of 0 , and a variance of 1 .

As a rough, but useful, criterion for data fit, we examine the extent to which the distributions of these standard residuals approach

$$z_{ni} \sim N(0,1) \text{ normal}$$
$$z_{ni}^2 \sim \chi_1^2 \text{ chi-square.} \qquad 24.3$$

The reference value 0 for the mean and 1 for the standard deviation and the reference distributions of $N(0,1)$ and χ_1^2 help us to decide whether observed standard residuals deviate unreasonably from model expectations. This examination of residuals helps us to decide whether we can proceed to use these items to make measures and also whether particular persons have failed, at least in part, to respond to the test in a use
ful manner.

When a particular squared residual

$$z_{ni}^2 = (x_{ni} - P_{ni})^2 / P_{ni}(1 - P_{ni}) \qquad 24.4$$

becomes large, we suspect that something unexpected happened when that person n took that item *i*. A single unexpected response, however, is less indicative of trouble than a pattern of unexpectedly large z_{ni}^2. The accumulated impact of a pattern of large z_{ni}^2 values for a person [or an item] arouses concern for the utility of that person's measure [or that item's calibration].

Consider the responses patterns in Table 24.1.

The circles in Table 24.1 mark unexpected responses. To evaluate the improbability of these responses we replace each instance of an unexpected response by the difference between the ability measure for that person and the difficulty calibration for that item. For Person 1 on Item 4 the unexpected incorrect response associated with person ability $b = -1.2$ and item difficulty $d = -3.9$ produces a difference $(b - d) = (-1.2) - (-3.9) = +2.7$.

This difference 2.7 for Person 1 on Item 4 is placed at the location of that unexpected response in Table 24.2 where we have computed the differences for each instance of an unexpected response circled in Table 24.1.

Unexpected *incorrect* answers have been recorded as $(b - d)$, but unexpected *correct* answers have been recorded as $(d - b)$. We do this because, when the response is incorrect, and $X = 0$, then the index of unexpectedness is $[\exp(b - d)]$, but, when the response is correct, and $X = 1$, then the index becomes $[\exp(d - b)]$.

We record unexpectedness in Table 24.2 as a positive difference, whether from $(b - d)$ or $(d - b)$.

The corresponding values for

$$z^2 = P / (1 - P) = \exp(b - d) \text{ when } X = 0$$
$$\text{and } z^2 = (1 - P) / P = \exp(d - b) \text{ when } X = 1$$

Table 24.1

Some Unexpected Person-to-Item Responses (x)

PERSON	ITEM							NUMBER OF UNEXPECTED RESPONSES	PERSON ABILITY
	4	5	6	7	8	12	14		
1	⓪	1	1	1	1	0	0	1	-1.2
2	1	1	1	⓪	⓪	0	0	2	-1.2
3	1	⓪	1	1	1	0	0	1	-0.6
4	1	1	1	1	1	①	0	1	0.0
5	1	1	⓪	⓪	1	①	0	3	0.0
6	1	1	⓪	1	⓪	0	①	3	0.0
NUMBER OF UNEXPECTED RESPONSES	1	1	2	2	2	2	1	11	
ITEM DIFFICULTY	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8		
"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY						"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY			

can then be evaluated for the improbability of the response. These z^2 values, which are taken from Column 2 of Table 24.4 (Best Test Design, Wright & Stone, 1979, p. 73) have been entered in Table 24.3.

Table 24.4 gives values of $z^2 = \exp(b-d)$ for unexpected incorrect answers $x = 0$ or values of $z^2 = \exp(b-d)$ for unexpected correct answers $x = 1$.

The entry C_x in Column 1 of Table 24.4 is $C_0 = (b-d)$ when the response is incorrect and $x = 0$ and $C_1 = (d-b)$ when the response is correct and $x = 1$.

We locate the difference +2.7 for the $|b-d|$ of Person 11 on Item 4 in Column 1 of Table 24.4 and read the corresponding z^2 in Column 2 as 15. This value and all of the other values for the differences in Table 24.2 have been recorded in Table 24.3 which now contains the z^2 for each instance of unexpectedness that we have observed for the six persons and seven items. The margins of Table 24.3

Table 24.2

Differences *(b-d)* Between Person Ability and

Item Difficulty for Unexpected Responses

PERSON	ITEM							PERSON ABILITY
	4	5	6	7	8	12	14	
1	2.7							-1.2
2				1.7	0.8			-1.2
3		2.7						-0.6
4						1.7		0.0
5			3.3	2.9		1.7		0.0
6			3.3		2.0		2.8	0.0
ITEM DIFFICULTY	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8	
	"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY					"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY		

give the sums of these z^2 for each person and each item. These sums indicate how unexpected the *patterns* of person or item responses are.

Column 3 of Table 24.4 shows $p = 1/(1+z^2)$, the model improbability of each observed response. This value provides a significance level for the null hypothesis of acceptable fit for any particular response. With our example of *(b - d)* = 2.7 we find a significance level of .06 in the table, against the null hypothesis that the response of Person 1 to Item 4 is according to the model.

QUALITY CONTROL

The evaluation of response patterns is a quality control procedure. In Rasch measurement, quality control over response patterns is implemented by determining the fit of response patterns to modeled expectations. Fit, or response plausibility, is determined from the difference between the estimates of person ability b and item difficulty d for each person/item contact. When this difference is positive, the item should be easy for the person. The more positive the difference, the easier the item and hence the greater our expectation that the person will succeed. Similarly, as the difference between

Table 24.3

Fit Mean Squares (z^2) for Unexpected Responses

PERSON	ITEM							PERSON MISFIT TOTAL
	4	5	6	7	8	12	14	
1	15							15
2				6	2			8
3		15						15
4						6		6
5			27	18		6		51
6			27		7		17	51
ITEM MISFIT TOTAL	15	15	54	24	9	12	17	146
	"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY					"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY		

person ability and item difficulty becomes more negative, the item should be more difficult for the person, and our expectation of failure increases.

The response pattern produced by each person is evaluated for the amount of misfit occurring. The diagnosis of patterns is expedited by plotting to show each pattern's shape and by summarizing the misfit in that particular pattern. A summary fit statistic is computed for each person and each item.

Figure 24.3 shows a response pattern that suggests guessing with an initial ability measure of $b = 3.2$. Four easy items were answered correctly followed by five items of increasing difficulty answered incorrectly followed finally by *two quite difficult items answered correctly.* Our attention is attracted to these last two most difficult items with correct responses following five easier items answered incorrectly. These last two correct responses are implausible.

Table 24.5 shows the residual analysis of the original pattern of responses and of the corrected pattern. We can compute two ability estimates for this person. One, at $b = 3.2$, is based upon the original full pattern. The other, much lower, at $b' = 1.7$, is based on deleting the last two implausible items. We question whether the original ability estimate $b = 3.2$ is a good indicator of this person's position on the variable because the response pattern misfit is $t = 5.3$. The corrected pattern fit of $t' = -1.2$ is more

Table 24.4

Some Misfit Statistics

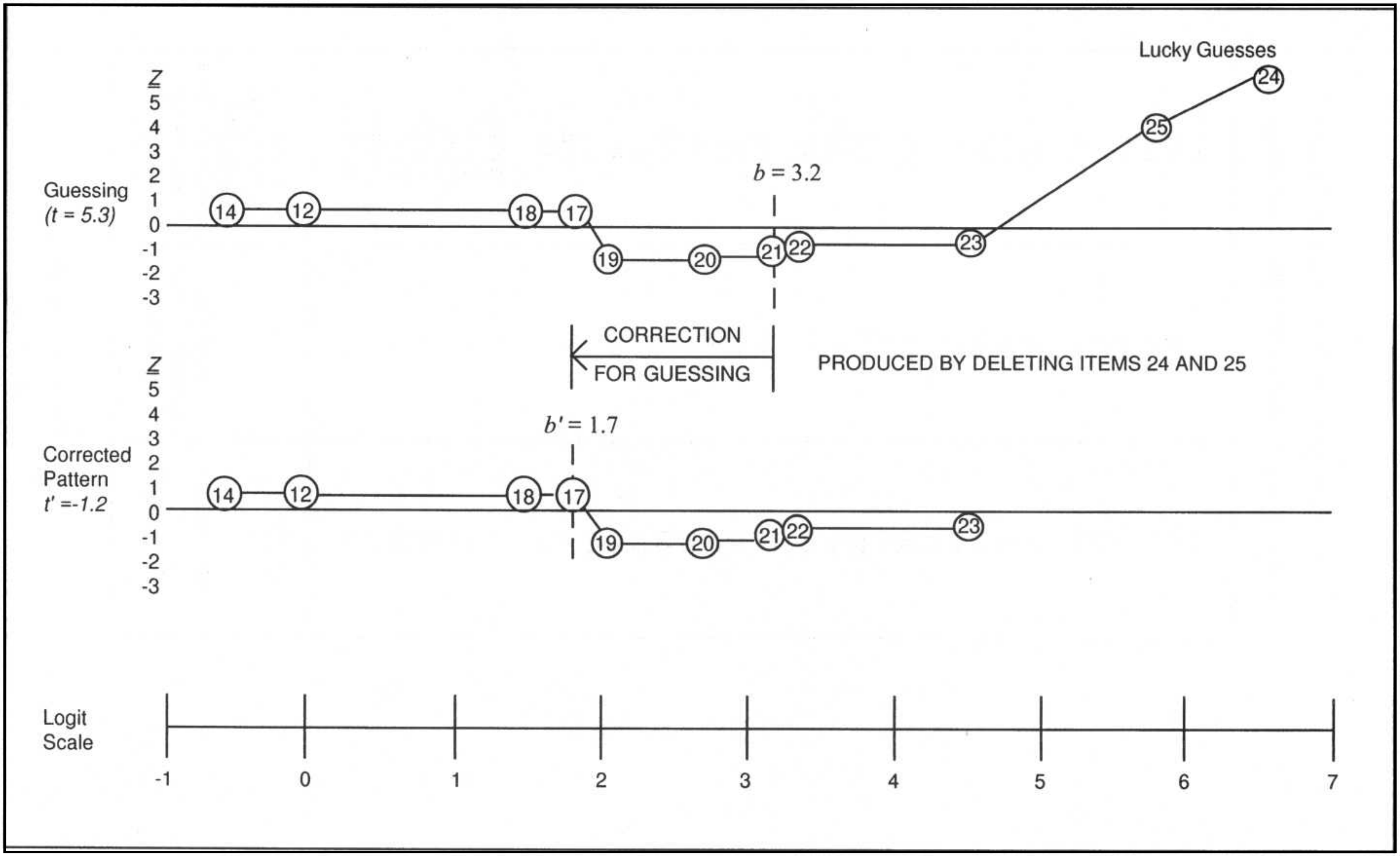
1. DIFFERENCE BETWEEN PERSON ABILITY AND ITEM DIFFICULTY C_x*	2. SQUARED STANDARDIZED RESIDUAL $z^2=expC$	3. IMPROBABILITY OF THE RESPONSE $P=1/(1+z^2)$
-0.6, 0.4	1	.50
0.5, 0.9	2	.33
1.0, 1.2	3	.25
1.3, 1.5	4	.20
1.6, 1.7	5	.17
1.8, 1.8	6	.14
1.9, 2.0	7	.12
2.1	8	.11
2.2	9	.10
2.3	10	.09
2.4	11	.08
2.5	12	.08
2.6	13	.07
2.7	15	.06
2.8	16	.06
2.9	18	.05
3.0	20	.05
3.1	22	.04
3.2	25	.04
3.3	27	.04
3.4	30	.03
3.5	33	.03
3.6	37	.03
3.7	40	.02
3.8	45	.02
3.9	49	.02
4.0	55	.02
4.1	60	.02
4.2	67	.02
4.3	74	.01
4.4	81	.01
4.5	90	.01
4.6	99	.01

* For incorrect responses when $x = 0$ then $C_o=(b-d)$.

For correct responses when $x = 1$ then $C_1=(d-b)$.

Figure 24.3

Correcting a guessing pattern.



acceptable. Which estimate we decide is more useful depends upon what we think about the responses of the person to these two items. If we think that these responses are implausible, that it is unlikely that this person would get these last two items correct after five failures, then we might take the corrected $b' = 1.7$ as the more useful estimate of this person's measure.

Statistical analysis alone cannot tell which estimate is more useful, but it can detect and arrange the available information into a concise and objective summary to use as part of our evaluation of the person. Persons who guess on multiple choice items may succeed on difficult items more often than their abilities predict. This could make them appear more able, especially if many items are too difficult for them. This is because their frequency of success would not decrease as item difficulty increased. A similar but opposite effect occurs when able persons become careless with easy items, making these persons appear less able.

Item responses affected by guessing express the simultaneous influence of more than one variable. There is the ability to be measured and, in addition, there is the tendency to guess. The "guessingness" of the item may or may not be a simple function of its difficulty on the main variable, or, if a multiple choice item, of its distractors. For the person being measured, at least, two quite different variables are involved. One is ability, the other is inclination to guess. The accurate measurement of either variable is threatened by the active presence of the other. In our empirical experience, when guessing does occur, it is dominated by the specific individuals who do the guessing and not by particular items, unless the items are poorly constructed.

When we detect a significant misfit in a response record, diagnose the response pattern and identify possible reasons for its occurrence, it is finally necessary to decide whether an improved measure can or should be determined. Whether a statistically "corrected" measure is "fair" for the person or "proper" for the testing authority cannot be settled by statistics. Nevertheless, knowing how a measure can be corrected objectively gives us a better understanding of the possible meaning in a person's performance.

For other examples of misfit patterns see Chapter 17 (p.143), Information and Misfit Analysis and Best Test Design (Wright & Stone, 1979, pages 165-190).

TAILORED MEASURING

In situations where we think that guessing may be influenced by test format as, for example, when we think a person may guess at random over m multiple-choice alternatives, we can use the guessing probability of $1/m$ as a threshold below which we suppose guessing might occur, as in Figure 24.2. To guard our measures against this kind of guessing we can delete all items from a person response record which have difficulty greater than $b + \log(m - 1)$ where b is the person's initial estimated ability. After these deletions we reestimate the person's ability from the remaining items attempted. When we do this, we are taking the position that when items are so difficult that a person can do better by random guessing than by actually trying, then, whatever the person's responses may be, such items should not be used to estimate the person's ability.

The procedure is:

1) When several unexpected responses are "correct" beyond some set fit statistic, say $t > 3$, suggesting the possibility of lucky guessing on the part of this particular person, delete all

Table 24.5

Correcting a Guessing Pattern

		ITEM NAME AND DIFFICULTY (IN DIFFICULTY ORDER)										
	ITEM NAME:	#14	#12	#18	#17	#19	#20	#21	#22	#23	#25	#24
	ITEM DIFFICULTY:	-0.5	-0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3

CASE DESCRIPTION	RESPONSE STATISTIC	RESPONSE PATTERN										
							b = 3.2				DELETE	
"Guessing" Pattern (b = 3.2)	x	1	1	1	1	0	0	0	0	0	1	1
	$(2x\text{-}1)(d\text{-}b)$	-3.7	-3.3	-1.8	1.3	1.2	0.3	-0.1	-0.1	-1.3	2.6	3.1
	z^2	0.0	0.0	0.2	0.3	3.3	1.4	0.9	0.9	0.3	13.5	22.2
	z	0.2	0.2	0.4	0.5	-1.8	-1.2	-1.0	-1.0	-0.5	3.7	4.7
					CORRECTION FOR GUESSING *							
				b' = 1.7								
Corrected Pattern (b' = 1.7)	x	1	1	1	1	0	0	0	0	0		
	$(2x\text{-}1)(d\text{-}b')$	-2.2	-1.8	-0.3	-0.2	-0.3	-1.2	-1.6	-1.6	-2.8		
	z^2	0.1	0.2	0.7	1.2	0.7	0.3	0.2	0.2	0.1		
	z	0.3	0.4	0.9	1.1	-0.9	-0.5	-0.4	-0.4	-0.2		

* *"Guessing" correction rule*: for m-choice items delede $d > [b + ln\ (m - 1)]$.
i.e., if $m = 5$, and $b = 3.2$, then delete any items with $d > 3.2\ ln\ (4) = 3.2 + 1.4 = 4.6$, i.e., items #25 and #24.

	MEASUREMENT						RESIDUAL ANALYSIS		
	Score r	Relative Score $f = r/L$	Relative Ability X_{fw}	Error Coefficient $C_{fw}^{1/2}$	Ability b	Error s	Sum of Squares $\sum z^2$	Mean Square v	Fit Statistic t
"Guessing" Pattern (b = 3.2)	6	.55	0.4	2.9	3.2	0.9	43.0	4.3	5.3 misfit
Corrected Pattern (b' = 1.7)	4	.44	-0.4	2.4	1.7	0.8	3.7	0.5	-1.2

Correction in measure: 1.7 - 3.2 = 1.5.

$v=\sum z^2/(L-1)$

$t=[\ln(v)+(v-1)][(L-1)/8]^{1/2}$

"too hard" items from this particular response pattern, that is, all items with $d > [b + \log (m - 1)]$ where m is the number of alternatives.

2) Compute a new ability estimate after the deletion of the too hard items and make another analysis of fit.

Steps 1 and 2 can be reiterated until successive values of *b* become stable and fit becomes acceptable. When this procedure is applied to response patterns generated entirely by guessing, the illegitimate responses are peeled away one at a time until the entire pattern is gone.

The use of a quality control process through misfit analysis of response patterns is the Rasch measurement way of dealing with guessing. In Rasch measurement we do not accept guessing as something to be tolerated when it can be avoided by external means, nor do we leave guessing to faulty estimation procedures produced by unworkable models. Instead, we arrange the testing experience so that guessing is least likely to occur and then use the quality control procedure of fit analysis to monitor all response patterns for any manifestations of whatever lucky guessing may occur.

25. EXPLAINING VALIDITY AND RELIABILITY

"Validity" and "reliability" are ubiquitous terms in social science measurement. They are prominent in the APA "Standards" (1985) and earn chapters in test theory texts. Yet they remain ambiguous and confusing in both theory and practice.

QUALITATIVE AND QUANTITATIVE ASPECTS OF VALIDITY

Validity has both qualitative and quantitative aspects. The qualitative aspects are conceptual. The quantitative aspects are numerical.

Fundamental to validity is the concept of a variable. A variable is intended to be a undimensional manifestation of one clear idea. It is the embodiment of an intention that is defined by the items written to implement the idea. Successful realization of a variable results from its concrete representation in items. This representation is then used to collect data from which the coherence and utility of the idea and its items is determined.

QUALITATIVE VALIDITY

The conceptual plan of a variable is its qualitative validity. Qualitative validity refers to the abstract idea of a variable and its content and illustrations that transform the abstraction from an idea to manifestation by items.

The qualitative aspects of a variable are its "content" and "construct" validity. These two forms of validity express the meaning of the variable. They explain the organization and construction of items and their use in eliciting manifestations of the variable. The items carry the intent of the variable and bring to concrete realization the abstract intention.

Content validity is demonstrated by items planned and written to bring the variable to life. But successful item writing must follow a plan. The plan requires thinking of the variable as a line with direction (an arrow) and arranging items according to their difficulty along this line. Such a plan follows a single dimension. The plan requires empirical confirmation. A dialogue results between the tasks of writing items according to the plan and gathering empirical evidence. Variable development requires this dialogue first to construct a useful variable and then successively to substantiate it. A variable by definition is undimensional. Its items operationally define the variable. This activity addresses the "content" validity question.

When such a plan and its resultant items are a useful representation of a variable, the items operationally define the variable and the variable is manifest by responses to the items. If the structure of the variable is supported by the item calibrations and if the characteristics of persons can be substantiated by their placement on the variable, we have construct validity.

QUANTITATIVE VALIDITY

The quantitative aspects of validity are numerical and statistical. They are content-free and devoid of text. They apply to any variable whatever its intention. They have no meaning by themselves, but are useful tools for crafting a variable and defining its numerical properties.

THE DIALECTIC BETWEEN QUALITATIVE AND QUANTITATIVE

The qualitative and quantitative interact in a dialogue between the idea of a variable and its quantitative manifestation. The idea provokes a plan and the writing of items. The calibration of items and measurement of persons provides the empirical realization. The process involves successive cycles modifying the plan of the variable and the items until perturbations between the two reduce to a point where further refinement is either unnecessary or not possible.

We might wish this dialectic to reach perfect agreement. But should we ever reach that level, the problem would become trivial or disappear altogether. Indeed, we do not expect these two aspects to fully reconcile. That would be the death of discovery. It would mark the end of improvement to the variable.

Instead, the result of this interaction produces something more than is possible by considering only the plan or only the results. Actually, we know there will and must always be a sense in which these two aspects of validity are never brought together. They can never achieve a perfect union, because it is precisely their *interaction* that engenders our motivation to press on.

INTEGRITY AND UTILITY VALIDITY

The two "validities," qualitative and quantitative, are often applied to test practice to determine concomitant and predictive utility. The Standards call these "criterion" and "predictive" validity.

Their designations as validity, however, are not apt because they do not bear on the integrity of the test. They only bear on the test's utility. It is useful to differentiate between integrity and utility.

By *integrity* we mean the sine qua non of the variable, the demonstration through construction of the variable intended, illustrated by items (qualitative validity) and supported by calculations (quantitative validity).

By *utility* we mean the application of the variable to whatever circumstances appear useful for investigating relationships between this variable and others.

Discussions about utility arise from applications of the variable to any number of circumstances. There is no limit to the number of questions and answers that might be aroused by applying the variable to different settings. There is no end to the investigation of utility, although not all situations merit equal attention.

Every conceivable application gives an answer to utility. Unless one application becomes and remains the main consideration, there is no way that utility can be interpreted as an essence of validity. Consequently, utility is not inherent to the integrity of the variable.

In some applications, the variable may have only one use. In such a narrowly defined circumstance, we might consider utility as relevant to the validity of the variable. The difficulty with this approach, however, is that inevitably the criterion changes.

THE STANDARD ERROR

The main consideration facing us in any measurement application is to determine what standard error must be achieved in a designated application. The question becomes one of measurement precision, a question about reliability. We want to know how many items we need in a given region of the variable to obtain measurement precise enough to meet a current need.

THE RELIABILITY COEFFICIENT

The reliability coefficient is commonly used to indicate how well the variable correlates with itself over independent applications as it were. The traditional reliability coefficient, however, is a confusing mixture of qualitative validity and measurement precision. It remains confusing until it is decomposed into its constituent elements.

The traditional reliability coefficient is a mixture of item fit and prediction because it is composed of point-biserial correlations between the item responses and total score. Unfortunately, it mixes these two issues in such a way that one cannot explicate the component parts. Furthermore, the two issues are quite different from one another.

Precision of measurement depends on nothing more than the number of items administered, the extent to which the items are on target and the degree to which the respondent uses the items coherently. It is a straightforward calculation that can be done for any example.

BEST TEST DESIGN

Best test design depends on relating the characteristics of test design T (H, W, L - height, width and length) to the characteristics of the target G (M, S, D - location, dispersion and distribution) so that the SEM is minimized in the region of the variable where the measurements are expected to take place.

A test design can be defined completely by three test characteristics; height, width and length:

H = the height of the test on the variable, the average difficulty of its selected items,
W = the width of the test in item difficulties, the range of its item difficulties, and
L = the length of the test in number of items.

A target specification states where on the variable we suppose the target to be:

M = our best determination as to target location,
S = our best determination as to target dispersion, and
D = our best determination as to target distribution.

A best test is one which measures best in the region where measurements are expected to occur. Measuring best means measuring most precisely. A best test design T (H ,W, L) is one with the smallest error of measurement SEM over the target G (M, S, D) for a given test length.

If we know what precision in measures we want and have enough items near where we think our targets will be found then the standard error of measurement in logits is almost always well approximated by 2.5 over the square root of the number of items administered. (For further details see Wright and Stone, 1979, pp. 129-140.)

There is nothing more to investigate and nothing else to compute. It is a straightforward calculation of measurement precision for the proposed test and application.

THE TEST PLAN

Qualitative considerations need to precede the collection of data because they direct the construction and development of items from the plan for the collection of data.

In item writing we can specify preliminary anchor item values determined by our theory underlying their construction. Then we can plot the resulting empirical values against the intention of the items to evaluate whether resulting values relate to the previous specification according to theory. This approach connects intention and realization.

Realization is never exact. But it can meet the guidelines and be supported by the collected values showing how on target the numerical values of the items are relative to the intended plan. We expect the plan to be substantiated to some degree, we also expect further refinement.

MISFIT ANALYSIS

When our intentions are supported by data we get construct encouragement. If there are some discrepancies they teach us something about the relationship. If there are misfits of certain items, these teach us something about what it is that we are working towards.

The item fit statistics are most important. But, unfortunately, they are overwhelmed by the reliability coefficient in the traditional approach.

In Rasch measurement, we isolate the misfit statistic so that we can see which items demonstrate quantitative validity and also where they appear in the hierarchy of qualitative validity. Fit statistics are diagnostic of validity. They guide the measurement process by detecting *lack of fit* and *too good fit*. The former identifies discrepancies between our intention and the results. The latter identifies circumstances too good to be true and hence, suspicious. Both need further investigation. (See Chapter 17 (p.143), Information and Misfit Analysis.)

The confrontation of qualitative and quantitative validity provides opportunities in data analysis as we learn more about our data and as we resolve the discrepancies appearing in the fit analysis. Then we take steps to see whether we can make them more in agreement.

CONCOMITANT AND PREDICTIVE VALIDITY

A second kind of validity, concomitant and predictive, is how the task arranges the people.

The motivation now is to measure differences between people. We began by addressing the items, but the ultimate purpose is measuring the people. Now our concern is how the people are spread out along the variable.

This is the relationship between the standard error of the test and the standard deviation of the people we are measuring. This can never be addressed in general. It depends upon what problem is posed, whether our attention is over a wide or narrow range of persons, over all people, or only 4 year old children.

The answer depends upon the application of the characteristics of the test. Whether the test is good enough for the circumstances. In general, we need enough precision for each level of growth that is being studied.

A good measurement strategy is to first use a pilot location test, followed by a specific test to target the discovered location with greater precision than can be achieved solely by the pilot test. The pilot test determines the general location on the variable and the second test more precisely targets that location. The combination of pilot and target tests is less wasteful of time, achieves the desired measurement precision, and generally uses fewer items.

However, what we are addressing now is the test's utility, not its validity. Questions about how people are spread out over the variable are secondary to questions about the integrity of the instrument, the idea of the variability and its development.

26. THE STEPS TO MEASUREMENT

Everyone who studies measurement encounters Stevens's levels (Stevens, 1957). A few authors critique his point of view, but most accept his propositions without deliberation. An enumeration, with some examples, suffices.

Stevens worked in the psychophysics of Weber and Fechner. But current authors on measurement see little connection between psychophysics and modern measurement practice. This primer shows a connection between psychophysics, Stevens's levels and Rasch measurement.

THE FUNDAMENTAL STEPS TO MEASUREMENT

When Stevens specifies four levels, he is identifying not four kinds of measurement but the four fundamental steps which lead to measurement. These steps are necessary to make measures. Each step must be satisfied to reach generalizable findings. The steps are: categorizing, ordering, constructing a unit, and setting an origin. The correspondence between the fundamental measurement steps to Stevens's levels is shown in Figure 26.1.

Figure 26.1

The steps of measurement.

FUNDAMENTAL STEPS	STEVENS' LEVELS
STEP AND FEATURE	LEVELS REACHED
1. Categorizing - deciding what to collect	1. Nominal
2. Ordering - defining what to count	2. Ordinal
3. Constructing an Abstract Unit - establishing a real number line	3. Interval
4. Setting an Origin - incorporating the logarithmic/ exponential connection between addition and multiplication	4. Ratio

On the left of Figure 26.1 are the steps to measurement: categorizing, ordering, constructing a unit, and setting an origin. These steps correspond to Stevens's nominal, ordinal, interval, and ratio levels on the right of Figure 26.1.

CATEGORIZING

Categorizing is the first step to measuring. We begin by determining what is to be noticed and collected, and also what is to be disregarded. In order to focus, we fix on single aspects isolated from the infinite variety of observations we might make. Until we can identify and maintain a single focus in our observations, i.e. categorize, we remain overwhelmed by the volume of possibilities that we perceive.

ORDERING

The second measurement step is to identify and isolate a useful line of inquiry along which elements can be ordered and comparisons made. This becomes the potential "variable," that guides us in collecting relevant observations which are relatively uncontaminated by irrelevant complications. It is not that complications cannot be informative, it is only that complications must become irrelevant in order to focus upon the one variable of interest. Sole focus evokes the clear determinations not possible when we admit a clutter of complications. Multivariate analysis is useful only after each variable has, itself, been realized as a single workable measure.

Stevens' nominal level concerns what is accomplished by categorizing.

CONSTRUCTING A UNIT: ADDITION

Constructing an abstract unit is the third step. We need to determine "how much," along the abstract variable, is indicated by observations of concrete counts.

Determining "how much" models the units as additive and hence linear. This results from item calibrations which define distances between observations categorized, ordered and counted.

SETTING AN ORIGIN: MULTIPLICATION

Fechner's postulation of a logarithmic relationship between stimuli and sensation enabled him to restate Weber's law to show that sensations are proportional to the logarithms of their exciting stimuli.

The third and fourth steps to measurement - units and origins are shown to merge by the logarithmic/exponential relationship between additive and multiplicative functions.

The logarithmic function (and its inverse, the exponential function) connect addition and multiplication. This connection is the tool we need to make measures.

Stevens calls the additive function "interval" measurement and the multiplicative function "ratio" measurement. However, the difference is merely the two sides of one characteristic, numerosity. The additive and multiplicative functions are dualities that bring out the joint necessity of determining a unit of measurement and setting an origin.

MAKING MEASURES

Raw counts of observations represent the additive function in its primitive concrete form. But counts are not measures. Measures are constructed from counts by transforming counts of concrete observations into abstract measurement. It is this transformation which constructs measures. How is this transformation made? What are the steps?

Georg Rasch addresses this in his *Probabilistic Models for Some Intelligence and Attainment Tests* (1960, 1993).

> [Rasch] makes use of none of the classical psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability or one item as the standard of difficulty, the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty... Indeed, these two properties were once suggested as criteria for absolute scaling (Loevinger, 1947); at that time proposed schemes for absolute scaling had not been shown to satisfy the criteria, nor does Guttman scaling do so. Thus Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of non-arbitrary measures (Loevinger, 1965, p. 151).

RASCH'S METHOD OF PARAMETER ESTIMATION

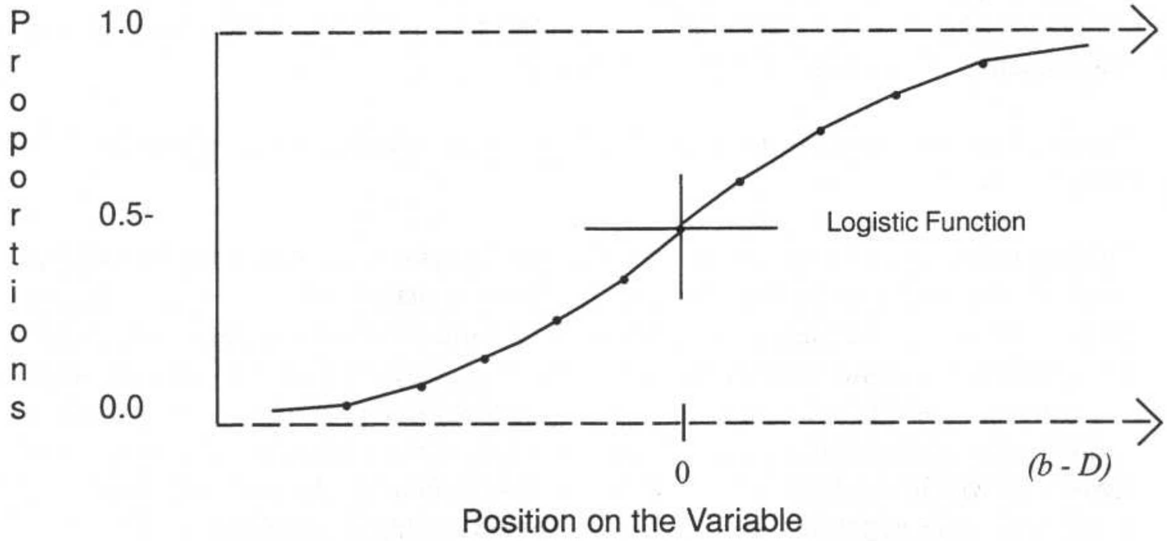
In *Probabilistic Models*, Rasch groups persons by their scores and takes the log ratio of successes to failures for each score group. He then fits these log odds to a two-way linear model to derive item-by-score group logits.

The log odds (logit) transformations of the original success-failure counts (when there are no interactions) produce parallel straight lines when item-by-score group logits are plotted against their averages over items or score groups. The logit transformation connects the two-factor multiplicative function to a one-dimensional additive function.

In Figure 26.2 we see how counts of success/failure on the ordinate can be transformed into measures on the abscissa by the logistic function. The bounded values of counts are transformed into unbounded measures. Given score groups large enough to give every item some successes and some failures, this logit transformation enables estimation of the simple linear structure that Rasch called objective measurement. Methods of parameter estimation are described in Wright & Douglas (1975), Choppin (1978), and Wright & Stone (1979).

Figure 26.2

Relating counts to measures.



Ordinate: counts expressed as $0 \leq$ proportion ≤ 1 (bounded).
Abscissa: Differences expressed as logits $-\infty < (B - D) < +\infty$ (unbounded).

REFERENCES

American Psychological Association (1985). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69-81.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR20 index and the Guttman scale response pattern. *Education Research and Perspectives, 9(1)*, 95-104.

Bechtoldt, H. (1959). Construct validity: A critique. *American Psychologist, 14,* 619-629.

Beck, L. (1950). Construction and inferred entities. In H. Feigl and M. Brodbeck (Eds.) *Readings in the Philosophy of Science, (1953), 262-287.* New York: Appleton-Century-Crofts.

Binet, A. & Simon, T. (1911). *A Method of Measuring the Development of the Intelligence of Young Children.* Lincoln, IL: Courier.

Binet, A. & Simon, T. (1916). The development of intelligence in children. (*Translations of articles in L'Annee Psychologique, 1905, 1908, and 1911.*) Vineland, NJ: Vineland Training School.

Boring, E. G. (1950). *A history of experimental psychology.* New York: Appleton-Century-Crofts.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika, 39,* 324-345.

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

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

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

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 42,* 631-634.

Choppin B. (1976). Recent developments in item banking. *Advances in Psychological and Educational Measurement.* New York: Wiley.

Choppin, B. (1978). Item banking and the monitoring of achievement. *(Research in progress Series No. 1.)* Slough, England: National Foundation for Educational Research.

Choppin, B. (1981). Educational measurement and the item bank model. In C. Lacey & D. Lawton (Eds.), *Issues in evaluation and accountability.* London.

Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Fischer, G. (1968). *Psychologische Test Theorie.* Bern: Huber.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A., ccxxii,* 309-368.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1934). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, xxii,* 700-725.

Fisher, R. A. (1935). *The Design of Experiments.* New York: Hafner.

Fisher, R. (1958). *Statistical Methods for Research Workers, 13th Edition.* New York: Hafner (First published in 1925).

Gould, J. G. (1981). *The Mismeasure of Man.* New York: Norton.

Guilford, J. P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

Guttman, L. L. (1944). A basis for scoring qualitative data. *American Sociological Review, 9*, 139-144.

Guttman, L. L. (1950). The basis for scalogram analysis. In Stouffer et al. (Eds.), *Measurement and Prediction.* New York: Wiley.

Hambleton, R. K. (1983). *Applications of Item Response Theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Hoyt, C. (1941). Test reliability by analysis of variance. *Psychometrica, 6(3)*, 153-160.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item Response Theory.* Homewood, IL: Dow Jones-Irvin.

Keats, J. A. (1967). Test theory. *Annual Review of Psychology, 18,* 217-238.

Keats, J. A. (1971). *An Introduction to Quantitative Psychology.* Sydney: John Wiley.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika, 29,* 1-27.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society (Series B), 27,* 251-263.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22,* No. 140.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability.

Psychological Monographs, 61.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72,* 143-155.

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,* 327-256.

Rasch, G. (1958). *On Applying a General Measuring Theory of Bridge Building Between Similar Psychological Tests.* Copenhagen: Danmarks Paedogogiske Institute.

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press. (Original work published 1960.)

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 321-333.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science.* 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.

Rasch, G. (1967). An informal report on the present state of a theory of objectivity in comparisons. In L. J. van der Kamp & C. A. J. Vieck (Eds.), *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof."* Leiden.

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction.* In Report from European Meeting on Statistics, Econometrics and Management Sciences: Amsterdam.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58-94.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press. (Original work published in 1960.)

Smith, R. M. (1982). *Detecting measurement disturbance with the Rasch model.* Unpublished doctoral dissertation, University of Chicago.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64(3)*, May.

Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties*. Research Report. Princeton, NJ: Educational Testing Service.

Thorndike, E. L. et al. (1926). *The Measurement of Intelligence.* New York: Teachers College Press.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17,* 445-457.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273-286.

Thurstone, L. L. (1927). A mental unit of measurement. *Psychological Review, 34,* 415-423.

Thurstone, L. L. (1928a). Attitudes can be measured. *American Journal of Sociology, 33,* 529-554.

Thurstone, L. L. (1928b). The measurement of opinion. *Journal of Abnormal and Social Psychology, 22,* 415-430.

Thurstone, L. L. (1928c). Theory of aptitude measurement. *Psychological Review, 22,* 415-430.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review, 36,* 222-241.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26,* 249-269.

Thurstone, L. L. & Chave, E. J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Thurstone, L. M. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17,* 446-457.

Thurstone, L. M. (1928). Theory of aptitude measurement. *Psychological Review, 22,* 415-430.

Tucker, L. (1953). Scales minimizing the importance of reference groups. *In Proceedings of the 1952 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Wilkinson, G. S. (1993). *WRAT3 Administration Manual.* Wilmington, DE: Wide Range, Inc.

Wolpe, J. & Lange, P. (1969). *Fear Survey Schedule.* San Diego, CA: Educational and Industrial Testing Service.

Wright, B. D. (1968). Sample-free test calibration and person measurement. I*n Proceedings of the 1967 Invitational Conference of Testing Problems.* Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2)*, 97-116.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3,* 281-288.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction. *Applied Psychological Measurement, 12,* 314-318.

Wright, B. D. & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21(4),* 331-345.

Wright, B. D. & Douglas, G. A. (1975). *Best Test Design and Self-Tailored Testing.* Research Memorandum No. 19. Statistical Laboratory, Department of Education: University of Chicago.

Wright, B. D. & Douglas, G. (1977a). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281-294.

Wright, B. D. & Douglas, G. (1977b). Conditional versus unconditional procedures for sample free analysis. *Educational and Psychological Measurement, 37,* 573-586.

Wright, B. D. & Linacre, J. M. (1991). *BIGSTEPS.* Chicago: MESA Press.

Wright, B. D. & Linacre, J. M. (1991). *SAMS.* Chicago: MESA Press.

Wright, B. D. & Linacre, J. M. (1992). *A User's Guide to BIGSTEPS.* Chicago: MESA Press.

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

Wright, B. D. & Mead, R. J. (1976). *BICAL.* Chicago: MESA Press.

Wright, B. D., Mead, R., & Bell, S. (1976). *BICAL: Calibrating items with the Rasch model.* Memorandum No. 23. Statistical Laboratory, Department of Education: University of Chicago.

Wright, B. D. & Panchapakesan, N. A. (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.