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:

$$y = x - Ex = x - P \tag*{1.7}$$

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

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

$$Ex = P \tag*{1.9}$$

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:

$$Ey \approx (b_1 - b_0)\sum(dy/db) \tag*{1.10}$$

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

$$dy/db = dP/db = P(1-P) = q \tag*{1.11}$$

$$Ey \approx (b_1 - b_0)\Sigma q \tag*{1.12}$$

and

$$(b_1 - b_0) \approx \sum y / \sum q = g. \tag*{1.13}$$

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.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware