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?"

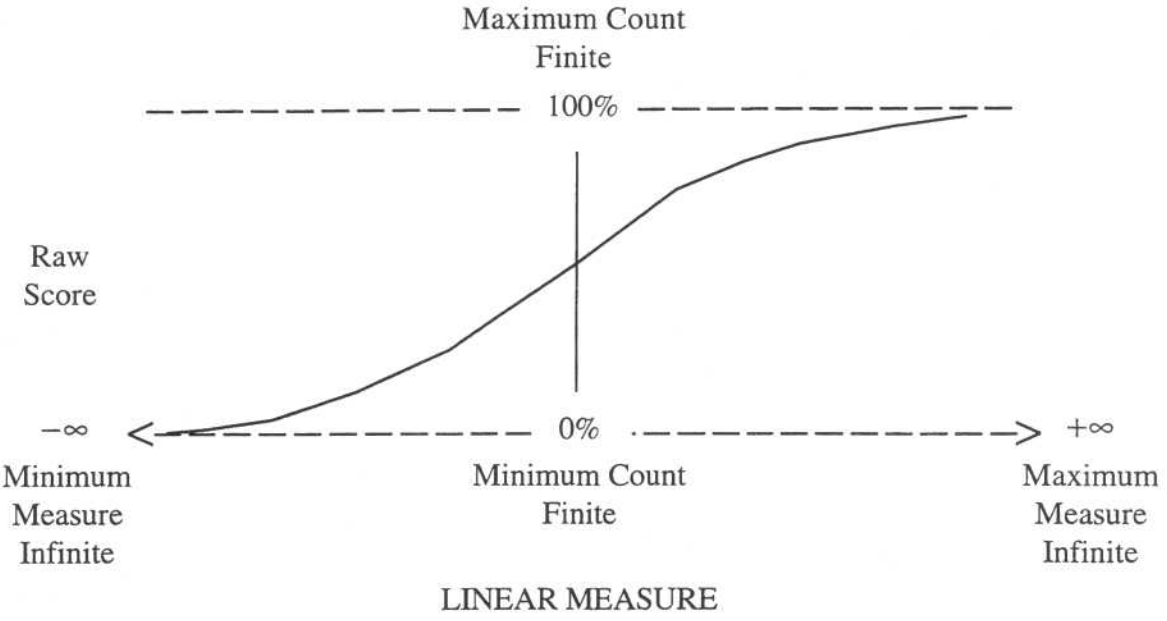
When data is complete and all data are used, then the relationship between concrete raw scores and the abstract measures they may imply becomes monotonic. This makes covariation analyses of raw scores and the measures they may imply appear similar.

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 measure-

ment 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.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware