IIIf. Rasch's Really Big Idea

There is no such thing as measurement absolute*; there is only measurement* relative. Jeanette Winterson.

The Case of the Missing Person Parameters

It was a cold and snowy night when, while trying to make a living as a famous statistical consultant, Rasch was summoned to the isolated laboratory of a renowned reading specialist to analyze data related to the effect of extra instruction for poor readers. There may be better ways to make a statistician feel a valued and respected member of the team than to ask for an analysis of data collected years earlier but Rasch took it on (Rasch, 1977, p. 63.)

If we could measure, in the strictest sense, reading proficiency, measurements could be made before the intervention, after the intervention, and perhaps several points along the way. Then the analysis is no different, in principle, than if we were investigating the optimal blend of feed for finishing hogs or concentration of platinum for re-forming petroleum.

In order to obtain evidence about reading proficiency, a parent or teacher might listen to the student read, commenting on errors and flow. There are many possibilities for evidence that might be collected, number of errors, number of words, perhaps having students retell the story in their own words, or respond to multiple choice items about main ideas, vocabulary from context, literary devices, sequence of events, use of imagery, topic sentences, etc.

There very well might be people for whom reading speed is not a reasonable indicator of proficiency. For advanced readers, for students with vision or hearing impairments, or students reading in a second language, reading faster might not imply reading better and pronunciation errors may imply aural or social impairment rather than lack of understanding. Or perhaps it works in Danish but not Mandarin.

We could begin to quantify the activity by counting the words read rather than just listening and commenting. In order to standardize, we could fix the amount of time allowed. Students reading from the same text for the same length of time could be ranked by the observed counts of words read or errors made. Alternatively, we could fix the number of words and count the number of seconds needed to read them. These counts are not measurements: without further refinement, the counts cannot be compared to other counts that were based on other texts or other time intervals.

The students in question had been given oral reading tests over a number of years. Texts of increasing difficulty were used on successive occasions, reasonably enough; but now, there were no connections between texts or between occasions. The only thing in common was the students and the goal was to measure how much they had changed. There seemed no sensible way to compare the scores from one text on one occasion to the scores from another text on a different occasion. That was the mystery Rasch needed to solve.

Rasch's background in mathematics and statistics and his understanding of Fisher's sufficiency brought to mind a simple expression[1], which, after some very clever mathematics, remarkably involved only the difficulty of the passages: the relationship between passages is estimated from

[1] It is portending of future debates that he avoided the cumulative normal distribution, which has almost the same shape but not the same properties for measuring as the one he choose.

the ratio of the counts of words read regardless of who is doing the reading. Rasch was able to collect data from totally new samples that connected all the texts and to use those estimates to evaluate the progress of the original sample. Mystery solved.

When Rasch described this work to Ragnar Frisch, Norwegian Nobel laureate in economics, in a casual conversation, Frisch remarked repeatedly, *the person parameter has completely vanished*. Rasch repeatedly responded, *yes, it has*, and continued to explain what they had concluded about remedial reading instruction. It took Rasch several days to appreciate what had struck Frisch immediately: **separating the two sets of parameters suggested an important new class of models with simple sufficient statistics.** (Rasch, 1977, p. 66) Rasch had not set out to change the world through *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960) but the solution to his practical problem yielded an expression with all the properties of fundamental measurement Thurstone was seeking.

The counts and probabilities of words read are person-dependent. Raw scores in any form are bound to the people who produced them. To deal with this, Rasch very cleverly devised a ratio of conditional probabilities that should be the same, within statistical limits, whether it came from a very fast reader, a very slow reader, from a male or female, fourth grader or eighth grader, an engineer or salesman, etc. But if these ratios should be invariant to the choice of the readers, then one must make sure they are. Control of the model, beyond blind acceptance of what comes out of the computer, means comparing the ratios based on different disaggregations of the group tested. This is what Rasch meant when he said "*the relationship should be found in several sets of data which differ materially in some relevant respects.*" (Rasch, 1960, p. 9.)

Specific Objectivity Redux

Rasch's Big Idea provided the solution to the problem posed by Thurstone's Big Idea:

> *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, 1928, p. 547.)*

And the analogous assertion that the measures for the objects must be independent of the measuring instrument. These are Thurstone's two conditions for *Fundamental Measurement*; he defined it and told us what it looked like but not how to get there. Rasch called his method for reaching Thurstone's measurement Nirvana[2] *Specific Objectivity* and described it as:

> *The "difficulties" of the tests of course have to be estimated from the body of data available, i.e., the results in two or more tests (or items) for each person of a certain collection. This collection, however, is not to taken to be a sample from any "population". On the contrary, the estimation procedure may be so conducted that the*

[2] This is a much reduced version of the Buddhist concept. We are only freeing ourselves from the confinement of sample-dependent statistics and population-based inferences, not the fires of desire, aversion, and delusion. Well, maybe delusion.

personal parameters—the "abilities"—and their possible distribution are eliminated. (Rasch, 1960, p. 3)

Thurstone laid down *principles*; Rasch described *procedures*. To avoid sounding too much like a philosopher or mathematician, all either of them is saying is that the calibration of the instrument should not depend on who has been tested and the measurement of a person should not depend on which form of a test was administered.

There are of course conditions that must be met before Rasch's procedures yield results that satisfy Thurstone's principles. A condition sufficient (not necessarily necessary) for specific objectivity is easy to state, not so easy to achieve:

All elements of the instrument must be equally valid and reliable.[3]

Taken together, validity and reliability as I understand them mean the agents have the same relationship to the aspect we are measuring and are not systematically related to any other aspect the objects may share. Then it makes sense to use as our summary (and sufficient) statistic for hardness the count of the number of materials scratched, for temperature the number of marks exceeded, for football prowess the number of points scored, or for math proficiency the number of items passed.

The Rasch Principle

Specific objectivity is the Rasch principle. *Objective* because it is not person or population-dependent; *specific* because the relationship may not be universal. When it holds, then any reasonable collection of people will provide the same estimate of the relative difficulty of any two relevant items and any relevant selection of items will provide the same relative abilities between any two appropriate people. The essential roles of separable parameters and sufficient statistics are to allow consistent estimators of the model parameters, and after the sufficient statistics have been extracted, all that remains should be noise. If it is, we have measurement; if not, we have a more interesting project than we thought.

The Rasch Method

The basic steps in Rasch's method is to:

A. devise agents to provoke valid responses consistent with our ideas about the aspect,
B. harvest the parameter estimates from the sufficient statistics, and
C. glean any structure from the residuals.

Often we rush through the first by just using the items as our definition of the aspect, shortchange the second by not trusting the sufficient statistics to tell us everything, and skip the third because we aren't sure we want to know the answer. If we get through these steps and there is no structure in the trash, we can proceed to making and analyzing measurements; in short, to doing science. If there is structure in the residuals, relating to, say, individuals, subgroups, item types or item content, we need to rethink the theory, reconsider the observations, review the instrument revising or discarding items, or restrict the domain of individuals for whom the instrument is appropriate. Reconsider anything and everything except the Model. Rasch was a very unreasonable man.

---

[3] For most of us most of the time, *instrument* means *test* and *element* means *item*.

Controlling the model to establish if, when, and where specific objectivity holds is the center piece of Rasch's method. It is generally appropriate, when estimating the model parameters, to use all the data one can get one's hands on, because larger samples mean smaller standard errors. However, for control, the total collection is partitioned and the results compared every way that is a potential threat: A partition is threatening if we suspect the relationships among agents or among objects might not hold on both sides of the partition and it matters if they don't: high vs. low performers, males vs. females, fourth grade vs. fifth grade, or computer vs. paper-and-pencil administration, selected response vs. constructed response. Having done all these checks with satisfactory results, we still do not know if the relationship is independent of visual acuity, mobility, computer experience, ethnicity, type of school or community, language spoken at home, eligibility for free or reduced lunch, region, mother's occupation, father's education, ad infinitum. Objectivity is specific to the threats eliminated.

Objectivity is specific to the threats eliminated.

Any of the possible subdivisions of the data can be investigated using likelihood ratio tests (Fischer & Molenaar, 1995), between group χ^2 (Wright & Panchapakesan, 1969; Wright & Stone, 1978), residual analyses (Mead, 1976), mean squares, weighted mean squares (Smith & Smith, 2004), or any number of other statistics[4]. Rasch often did it graphically. His work is filled with plots comparing the performance of groups of examinees, demonstrating the degree to which specific objectivity held, and identifying the instances where it did not.

No single fit statistic is either necessary or sufficient. David Andrich

Rasch's method, once we have observations from a well thought-out instrument, to get from the lowly level of counts to the lofty level of measures means we must:

I. Choose an appropriate mathematical form that:
   a. Has separable sets of parameters, and
   b. Describes the process that generated the counts.
II. Do the arithmetic:
   a. Eliminate nuisance parameters from the estimation equations using the model's sufficient statistics.
   b. Compute the parameter estimates.
III. Check that the sufficient statistics really are sufficient:
   a. Partition the total group every way possible, and
   b. Confirm that the relationships among the parameter estimates hold for all important subgroups.
      i. If yes, use the instrument to make the measurements needed to answer the questions we were interested in in the first place.
      ii. If not, rethink the theory, revise the items, revisit the instrument, and restrict the universe.

There is a growing family of mathematical forms to choose from and computers will do the arithmetic. The psychometrics is easy; constructing an appropriate instrument is not.

[4] If your data have something to tell you, your statistics won't stop it. *G. E. P. Box*.

The goal of this exercise is to develop measurement scales that are as well-defined and as useful in the classroom and clinic as the ones routinely expected in the physical sciences. Thurstone, and others, defined what measurement must be. Rasch provides the mechanism to achieve it and the framework to know when we have. It depends on the conscientious development of the agents (items) from a substantive theory and a rigorous verification of performance based on data. When accomplished, we will have measuring instruments that we can place alongside thermometers, rulers, radar guns, and scleroscopes, no apologies needed.