Chapter 19
Comparisons and Contrasts Between Item Response Theory and Rasch Measurement Theory

This chapter is a review, and an extension of ideas, functions and approaches to measurement in education and the social sciences that leads into the remaining chapters.

Approaches to Measurement and the Data-Model Relationship in Measurement

There has been a substantial degree of controversy in the approaches to social measurement. Because you are likely to come across this controversy in some guise or another, often not explicit, we consider it in this chapter. However, the controversy centres on two paradigms involved in the relationship between models and data (Andrich, 2004, 2011).

The controversy rests on the approach taken to *measurement* when statistical models are applied in the way they might be applied to data where the measurements already exist, and not in the many details of estimation, tests of fit, and so on. The technical details are generally common, although different models lend themselves to different considerations in the tests of fit. In both approaches, it is taken for granted that the raw data may need to be transformed if the data are intended to provide any generality across time, across instruments and across locations.

In constructing measurements of some construct, there are two simultaneous goals:

 (i) to better understand the construct or variable of measurement, and to modify the instruments in order to improve their operationalization and measurement of the construct;
(ii) to assess and formally measure the locations of objects of measurement, in our case often proficiencies or attitudes of people, on the construct or variable.

The two approaches to the data model relationship in constructing instruments for measurement are briefly summarized below.

Approach 1

Best efforts are made to construct test instruments which have content validity. Then an effort is made to construct plausible models that will characterize the data. The models have both the item (test) parameters, which help clarify the instrument and operationalize the construct, and person parameters. There is no particular a priori restriction on the class of models, and the parameters in these models, that might be used. Instead, the main criterion is whether the model fits the data. If the chosen model does not fit the data, another model of the same kind but with more parameters is tried. The model with more parameters will generally account for the data better than one with less parameters. This is the approach of IRT.

Approach 2

Best efforts are made to construct test instruments which have content validity. Then an effort is made to identify models that might characterize the data and which also subscribe to certain criteria of measurement. The models have both the item (test) parameters, which help clarify the instrument and operationalize the construct, and person parameters. There is a particular a priori restriction on the class of models and the parameters in these models that might be used. The case for these models is independent of any data sets, and data should be valid in content and also conform as close as possible to the models. We have seen in Chap. 7 that the case for the models rests on a certain kind of invariance that the responses should have in order that meaningful comparisons can be made as a result of the measurements.

It is this approach that is applied with RMT. Because it involves approaches that are essentially incompatible, we call the difference between the IRT and RMT approaches a *paradigm* difference.

The approach fostered in this book is the second, RMT, approach. This approach has traditionally been more unusual, but not uncommon, and is becoming more common. Equivalent criteria have been articulated by L. L. Thurstone (1920s), L. Guttman (1940s), and G. Rasch (1960s). Their criteria are consistent with each other, and consistent with the philosophy of Kuhn (1961) regarding the function of measurement in physical science. Kuhn and Thurstone are considered briefly at the end of this chapter. Guttman is considered separately, both earlier in this book and in the next chapter.

In this section, we build up a particular kind of invariance and function of measurement by considering quotes from key people. This criterion and function of measurement is also consistent with the approach taken in this book.

The Function of Measurement in Quantitative Research in the Natural Sciences: Thomas Kuhn

Thomas Kuhn was a physicist who turned to the history and philosophy of science and introduced the term *paradigm* in the philosophical discourse of the history of science and measurement. He introduced the idea that, in addition to traditional, cumulative science, there are episodes in the history of science in which the thinking is revolutionary. These intellectual revolutions can take centuries to be completed. Kuhn's key publications appeared in the 1960s.

What Do Text Books Teach Is the Function of Measurement in Science?

> In textbooks the numbers that result from measurements usually appear as the archetypes of the 'irreducible and stubborn facts' to which the scientist must, by struggle, make his theories conform… But in scientific practice, as seen through the journal literature, the scientist often seems rather to be struggling with facts, trying to *force* them to conformity with a theory he does not doubt. Quantitative facts cease to seem simply the 'given'. They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon. Often scientists cannot get numbers that compare well with theory until they know what numbers they should be making nature yield (Emphasis added) (Kuhn, 1961/1977, p. 193).

What Does Kuhn Say Is the Function of Measurement in Scientific Research?

> Only a miniscule fraction of even the best and most creative measurements undertaken by natural scientists are motivated by a desire to discover new laws and to confirm old ones (Kuhn, 1961, p. 187).

> …new laws of nature are very seldom discovered simply by inspecting the results of measurement made without advance knowledge of those laws. …because nature itself needs to be *forced* to yield the appropriate results, the route from theory or law to measurement can almost never be traveled backwards (Emphasis added) (Kuhn, 1961/1977, p. 197).

Is There a Role for Qualitative Study in Quantitative Scientific Research?

> …that large amounts of *qualitative* work have usually been prerequisite to fruitful quantification in the physical sciences (Emphasis added) (Kuhn, 1961, p. 180).

If discovering new laws or confirming existing ones is not the function of measurement, then

What Is the Function and Role of Measurement in Science?

> To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying serious anomaly, they tell scientists when and where to look for a new *qualitative* phenomenon. To the nature of that phenomenon, they usually provide no clues (Emphasis added) (Kuhn, 1961/1977, p. 205).

In summary, the function of measurement in physical science is the search for *anomalies*.

The Properties Required of Measurement in the Social Sciences: L. L. Thurstone

Thurstone was an engineer, who worked for some period with Thomas Edison, but then turned to psychology and was Professor of Psychology at The University of Chicago. His work on measurement in the social sciences was strongly influenced by his engineering and scientific background. Thurstone's key publications appeared in the 1920s.

Social Variables—What Is Distinctive About Variables of Measurement in the Social Sciences and What Are the Limits to Such Variables?

> One of the main requirements of a truly subjective metric is that it shall be entirely independent of all physical phenomena. In freeing ourselves completely from physical measurement, we are also free to experiment with aesthetic objects and with many other types of stimuli to which there does not correspond any known physical measurement (Thurstone, 1959, p. 182–83).

Thus They Must Be Independent of Physical Variables—What Else?

> The various opinions cannot be completely described merely as *more* or *less*. They scatter in many dimensions, but 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 (Thurstone, 1959, p. 218–19).

Why Do You Think We Have Quantification in the Social Sciences?

In practice, we have the following examples: marks for proficiency, performance and achievements, marks on national tests of educational progress and in attitude measurement. Clearly, in attempts to measure, the construct must include the idea of *more or less, greater or lesser, stronger or weaker*, and so on.

A Requirement for Measuring Instruments

> If a scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale (Thurstone, 1959, p. 228).

> 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, 1928, p. 416).

Thus in measurement, it is necessary for the instrument to operate the same way (invariantly) across groups. Do these requirements seem reasonable?

Georg Rasch

Rasch was a Danish mathematician and statistician, who was asked to help monitor the progress of students in reading and, in the process, developed a class of models for measurement in the social sciences. He then carried out statistical consulting to earn a living between the first and second world wars. Through a scholarship, he studied with Ronald Fisher for a year in 1934.

Rasch completed his career as Professor of Statistics as Applied to the Social Sciences at The University of Copenhagen. His consulting included, when it was formed, work for the Danish Institute for Educational Research and this is where his innovative work first took shape. He also had strong links with the Departments of Statistics and Education at The University of Chicago in the 1960s and 1970s. His last official appointment was as Visiting Professor in the Departments of Mathematics and Education at The University of Western Australia in 1974. Rasch's key publications appeared between 1960 and 1977.

On reporting on the work he did with models for reading and other kinds of data, he wrote the following regarding the relationship between a model and data.

> It is tempting, therefore, in the case with deviations of one sort or other to ask *whether it is the model or the test that has gone wrong.* In one sense this of course turns the question upside down, but in another sense the question is meaningful. For one thing, it is not easy to believe that several cases of accordance between model and observations should be isolated occurrences (Emphasis in original) (Rasch, 1960, p. 51).

What kind of model is a Rasch model for measurement? The model arises from the following requirement.

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, p. 332).

Compare this statement with that of Thurstone's regarding the property of an instrument. This invariance seems very important, not just for science and generality, but where humans are concerned, for social justice and for accurate diagnoses.

For example

 (i) if two markers grade student performance, we would require that the grades are independent of which marker is grading;
(ii) if two radiologists are studying X-rays, we would require that the interpretations are independent of which radiologist is reading the X-rays.

The Criterion of Invariance

The requirements of invariance articulated by Thurstone and Rasch are not descriptions of any data set. They are requirements that data need to meet if they are to be used in measurement.

The distinctive part of Rasch's models is that the criterion of invariance is built into the model, and the models are innately probabilistic. Then the check on invariance involves checking if the responses conform to the model. There are many ways in which the responses may not conform to the model. These are discussed in some detail in the rest of this book. We can consider such responses as, in some sense, *anomalies*.

It is a challenge in many situations (physical and social science) to meet this requirement. Meeting this requirement brings, as noted by Kuhn in physics, an integration of qualitative and quantitative considerations. The RUMM2030 program is consistent with this philosophy. It enables the examination of data from many perspectives with the researcher in control.

With more parameters than the Rasch models, other models are likely to absorb some of this lack of invariance. By not absorbing features of the data that models with more parameters would, the Rasch models are more likely to reveal anomalies in the data.

Fit with Respect to the Model and Fit with Respect to Measurement

Another perspective in the distinction between the IRT and Rasch measurement theory (RMT) paradigms is to contrast the concept of misfit. It helps to set the IRT paradigm in context.

The data modelling paradigm that arose from the natural sciences has been adopted in the social sciences and in IRT. However, in the natural sciences, it is taken for granted that the data are already measurements. For example,

> "Laws of error," i.e., probability distributions assumed to describe the distribution of the errors arising in repeated measurement of a fixed quantity by the same procedure under constant conditions, were introduced in the latter half of the eighteenth century to demonstrate the utility of taking the arithmetic mean of a number of measurements or observed values of the same quantity as a good choice for the value of the magnitude of this quantity on the basis of the measurements or observations in hand (Eisenhart, 1983, p. 1).

> Other discrete laws of error were proposed and studied by Lagrange; continuous laws of error by Simpson, Lambert, Laplace, Lagrange, and D. Bernoulli culminating in the quadratic exponential law of Gauss $f_x(x) = (h/\sqrt{\pi}) \exp(-h^2 x^2)$, upon which Gauss based his first formulation of the method of least squares, which became almost universally regarded in the nineteenth century as "*the* law of error" (Eisenhart, 1983, p. 1).

Notice that the distribution pertains to random errors of *measurement*. The Gaussian is the basis of the t, χ^2, and F distributions in which assessment is made as to whether or not the model has accounted for all the systematic variance. If it has not, then the distribution, given the model and its parameter estimates, will not be a random error distribution. Then a model with more parameters, which may account for the systematic factor or factors not accounted for by the simpler model, is sought. However, all these distributions assume that the data analysed are measurements.

In RMT, the task is to demonstrate that the instruments are producing numbers which are as close to measurement as can be obtained. The criterion is that of a relevant Rasch model which has properties of measurement and is not chosen to describe any particular data set. In RMT, the misfit of concern is relative to *measurement*. If one then uses a more complex model, misfit from measurement is absorbed into the model. The model will fit better because it has additional parameters such as discrimination parameters for items. However, this better fit may be hiding deviations from measurement, for example lack of invariance with respect to different groups and across the continuum, which the Rasch model highlights.

From the above perspective, when one is satisfied that the data show adequate measurement properties, then modelling such as the application of hierarchical linear

models to make group comparisons, and so on, becomes appropriate. Then the use of modelling data is analogous to the way data are modelled in the natural sciences.

The Linear Continuum as an Idealization

In concluding this chapter, we note again Thurstone's comment above: "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" (Thurstone, 1959, p. 218–19). The *linear scale* implies the mapping of the magnitude of the property onto a line. However, it must be appreciated that the line is an idealized abstraction and that there is no real line in nature. Thus, the property of measurement does not itself have to appear linear. For example, an electric wire that connects a power point to a computer might be bent in many places, but the strength of the electric current going through it, or the resistance of the wire, can be measured by the mapping of their magnitudes onto this idealized line.

Exercises

1. Summarize, in no more than 200 words, the idea of a *paradigm* in research and in scientific research.
2. Summarize, in no more than 300 words, the distinction between the two approaches to measurement outlined in this chapter. Relate this distinction to the idea of paradigms of research and to how models are used in these two paradigms.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Eisenhart, C. (1983). Law of error I: Development of the concept. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 530–547). Toronto: Wiley.

Kuhn, T. S. (1961/1977). The function of measurement in modern physical science. *ISIS, 52*(2), 161–193. Reproduced in Kuhn, T. S. (1977). *The essential tension*. Chicago: The University Chicago Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by Wright, B. D. (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Proba-*

bility (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in Bartholomew, D. J. (Ed.). (2006). *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334). London: Sage Publications.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*(4), 529–554.

Thurstone, L. L. (1959). *The measurement of values*. Chicago: University of Chicago Press.

Further Reading

Andrich, D. (1989). Distinction between assumptions and requirements in the social sciences. In J. A. Keats, R. Taft, & S. H. Lovibond (Eds.), *Proceedings of the XXIVth International Congress of Psychology, Mathematical and Theoretical Systems* (Vol. 4, pp. 7–16). B.V. North Holland: Elsevier Science Publications.

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease: Stakes and future challenge* (Chapter 7, pp. 66–91). Taylor and Francis: CRCS Press.

Kuhn, T. S. (1957). *The Copernican revolution*. Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific resolutions* (2nd Enlarged ed.). Chicago: The University Chicago Press.

Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*(1), 47–58.