Controversy and the Rasch Model

A Characteristic of Incompatible Paradigms?

David Andrich, PhD

Abstract: The development of Rasch models in educational and psychologic measurement in the 1960s coincided with the introduction of other similar models, now described as models of item response theory (IRT). The application of IRT models has now extended to other social sciences, including health. Originally, there was substantial controversy between those who saw Rasch models as simply special cases of IRT models and those who saw them as essentially different. Because these different perspectives continue to manifest themselves in various ways, it seems relevant to understand the source of the original controversy. This paper attempts to do so by invoking Kuhn's studies in the history and philosophy of science at 3 levels. First, it suggests that the 2 perspectives reflect Kuhn's concept of legitimate, incompatible paradigms in which controversy is a typical manifestation. Second, because Kuhn recognizes individual histories in the development of paradigms, Rasch's own shift in perspective is summarized. Third, because proponents of the Rasch models emphasize the models' compatibility with fundamental measurement found in physical science, an analogy is made between how Kuhn explains the role of measurement in the physical sciences and how proponents of Rasch models explain the role of these models in the social sciences. In particular, these roles cannot be gleaned from textbooks in science and statistics, respectively.

Key Words: Rasch models, paradigm shift, Rasch controversy, Item Response Theory, fundamental measurement.

(*Med Care* 2004;42: I-7–I-16)

Two independent developments in test theory appeared in the 1960s, one articulated by Rasch,[1] the other by Lord and Novick[2] and Birnbaum.[3] This theory is now generally referred to as item response theory (IRT). Applications of these models have more recently been extended into medical and health care assessment.[4–7]

From the School of Education, Murdoch University, Murdoch, Western Australia, Australia.

Reprints: David Andrich, School of Education, Murdoch University, Murdoch, Western Australia 6150, Australia. E-mail:andrich@murdoch.edu.au.

ISSN: 0025-7079/04/4200-0007

DOI: 10.1097/01.mlr.0000103528.48582.7c

In education, as now in other fields, these models were seen as especially useful in assessing performance and achievement across groups using conformable items in which not all persons needed to respond to all items. However, 2 perspectives emerged. In one, Rasch models are only special cases of IRT models; in the other, they are essentially different. These perspectives, still evident in the literature,[7] originally generated controversy. Proponents of the 2 perspectives are now more likely to agree to disagree, but the perspectives continue to manifest themselves in reviews of articles, which in turn reflect what, with what emphasis, and where various studies are published, and in presentations of articles from one perspective without presentations from the other. New researchers reading the literature and applying IRT to new fields can find these differences puzzling. It seems relevant, therefore, to understand the 2 perspectives. This paper attempts to do so by considering the original controversy. The approach taken is to compare the 2 perspectives using the framework of the physicist and historian of science, Thomas Kuhn[8] at 3 levels: first, his concept of research paradigm within scientific revolutions; second, his recognition of roles of individual histories of scientists; and third, his explanation of the role of measurement in physical science.

Revolutionary Episodes

Certain episodes in science are revolutionary because they challenge assumed, taken-for-granted understandings of a field. To summarize a collection of mutually reinforcing understandings underpinning a science, Kuhn[8] used the term *paradigm*. Elaborations of theories that take a paradigm for granted are the province of *normal science*. Sometimes, a revolution which generates its own paradigm, in principle incompatible with the previous one, emerges. Different paradigms in ostensibly the same field produce distinctive research agendas, with one manifestation being controversy. To attempt to explain the controversy with the Rasch models in terms of paradigms, in particular between what in this paper are termed the *traditional* and *Rasch* paradigms, it is instructive to anticipate briefly the main feature that distinguishes them.

Before doing so, I note 3 preliminary points. First, there are many IRT and Rasch models relevant for different circumstances, but to help make some issues concrete, I use just 2 models for dichotomously scored items of performance or achievement assessment. Second, there is substantial overlap between the 2 paradigms, estimation of parameters in models, construction of tests of fit, and so on, which tends to mask the fundamental difference and its consequences. This overlap is also considered in the paper. Third, the starting point is that the traditional paradigm is defensible and subscribed to by the majority of researchers in IRT. Here I am trying to analyze why, given that the Rasch models are algebraically special cases of general IRT models, proponents of Rasch models appear to work within a different paradigm, how that paradigm was reached, and on what grounds this paradigm, too, might be defensible. Paradigms go beyond simply the specification of models.

The Traditional Paradigm

Briefly, in the traditional paradigm, the case for choosing one model over another is that *it accounts better for the data*. The data are given. In general, the model with a greater number of parameters accounts better for the data. If, according to available statistical checks, it does not, the model with fewer parameters is favored.

This perspective is generally simply taken for granted; it is part of the paradigm. Thus, statistics texts explain model fitting and using formal statistical tests between competing models. Invariably, it is models that are accepted or rejected, given the data. For example, this perspective appears essentially incidentally in a discussion on the application of IRT in health sciences[9]: "Normally, an assumption is made when fitting an IRT model to a set of data . . ." (p. 71) and "First, there is the difficulty of finding a model that fits the available data and estimating model parameters" (p. 73). Sometimes this position in psychometrics is explicit.[10] ". . . [W]hen the criterion indicates nonrandomness, an examination of residuals may suggest how the model should be modified to improve fit" (p. 5). From the traditional paradigm, *these quotes are not jarring*.

The Rasch Paradigm

The case for the Rasch[11] model is its property of invariance; briefly, that the comparison between any 2 persons should be independent of which items from a class of items is used, and vice versa. The case is *not that it describes data*. From the traditional paradigm, this case for a model, independent of data, is *jarring*. This is the case that is elaborated in the paper as giving rise to a different paradigm.

For dichotomously scored items, the Rasch model resulting from the condition of invariance is

$$\Pr\{\text{positive response}\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \qquad (1)$$

where β_n and δ_i respectively characterize the locations of person n and item i. From the perspective of the traditional paradigm, when this model does not fit the data, the more general model,

$$\Pr\{\text{positive response}\} = \frac{e^{\alpha_i(\beta_n - \delta_i)}}{1 + e^{\alpha_i(\beta_n - \delta_i)}}, \qquad (2)$$

where α_i characterizes the discrimination of item i, is considered. From this perspective,[7,12] equation (1) is simply a special case of equation (2) with all $\alpha_i = 1$. Rasch's term, the simple logistic model (SLM), is used for equation (1), and the common term *two-parameter logistic* (2PL) for equation (2). There are generalizations of the SLM for ordered category formats that are Rasch models and generalizations of the 2PL that are not Rasch models.[7]

Rasch arrived at his criterion of invariance of comparisons after successfully modeling a set of data with the Poisson distribution. Rasch's change in perspective came slowly to him. It therefore seems relevant to describe Rasch's shift in perspective and to see if it helps explain the sense in which the SLM is not simply a special case of the 2PL. I present this shift, with suggested parallels to paradigm shifts in general, which are characterized, among other features, by controversy, relative simplicity of models, proponents talking through each other, discovering a problem after it is solved, newcomer to a field precipitating the new paradigm, solving different problems with different agendas, and recourse to philosophy.

Controversy

Controversy appears in a field because of incompatibility of paradigms and goals among researchers. The quotes below which appeared early in the debate are in the public domain, but many more can be found in the more informal literature, in debates in meetings that are not recorded, on internet forums, and as indicated earlier, in reviews of papers not normally public. First, the SLM is rejected outright.

It is argued that the Rasch model, and item banks based upon that model, constitute inappropriate tools for use in educational assessment. The paper discusses the dangers to the educational system which should result should this model be used for routine educational monitoring.[13]

". . . [T]he Rasch model . . . includes only one free item parameter, that for difficulty. . . . items that fit a one-parameter model all have the same discrimination parameter . . ."

These assumptions about items fly in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years.[14]

Second, the SLM is endorsed enthusiastically.

"Rasch[1] has devised a truly new approach to psychometric problems. . . . Rasch must be credited with an outstanding contribution to one of the 2 central psychometric problems, the achievement of nonarbitrary measures."[15]

"The Rasch model defines a new level of aspiration and sets a new challenge for social measurement . . . It follows that the case for the model does not stand or fall with the success or failure in applying it to any particular substantive area. . . ."[16]

Characteristically, these 2 sets of quotes do not address each other.

Relative Simplicity of New Models

Generally, an appeal of new paradigms is that their models, at least initially, are relatively simple.[8] The SLM is simpler than the 2PL. Figure 1 shows the curves of the SLM for 4 items with different locations, and Figure 2 shows the curves of the 2PL for the same locations but with different discriminations ranging from 0.5 to 2.5. In the former, they are parallel, in the latter they intersect.

In the SLM, a person's total score is sufficient for the parameter β_n; there is no information in the pattern of responses, though because different patterns have different probabilities, they can disclose misfit. This simplicity appeals to proponents of Rasch models. In contrast, in the 2PL, each pattern of responses provides a different location estimate,[17] though further misfit can still be identified.

Technically, sufficiency in the SLM permits writing the equation

Pr{response to i is positive and to j negative, given only one

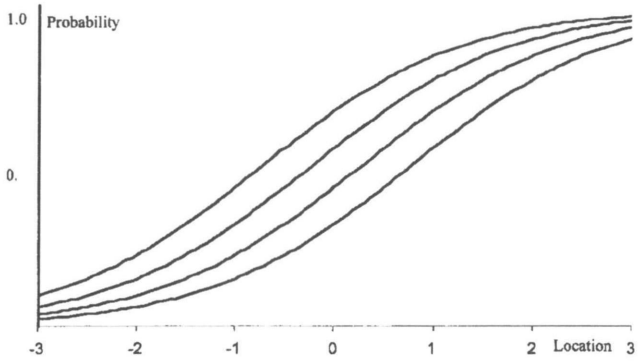

FIGURE 1. Parallel ICCs for the Rasch SLM.

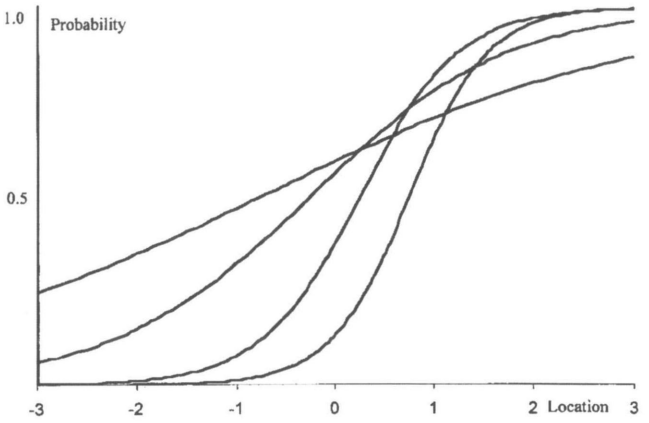

FIGURE 2. Intersecting ICCs for the 2-parameter logistic model.

$$\text{is positive}\} = \frac{e^{\alpha_j - \alpha_i}}{1 + e^{\delta_j - \delta_i}} \quad (3)$$

which does not contain the person parameter and characterizes comparisons of items which are invariant relative to the locations of persons.

Talking Through Each Other

Proponents of the SLM and the 2PL agree that the latter model is more likely to account for the data. However, for proponents of the SLM, the 2PL can also absorb features that would violate the simpler SLM and hide potential problems with the data relative to the SLM.[18] To proponents of the 2PL, the SLM is often *too* simple for data.[12] These 2 positions are irreconcilable in principle, and those taking each seem to do so from the perspective of their own paradigm.

"To the extent that 2 scientific schools disagree about what is a problem and what a solution, they will inevitably talk through each other when debating the relative merits of their respective paradigms. In the partially circular arguments that regularly result, each paradigm will be shown to satisfy more or less the criteria that it dictates for itself and to fall short of a few of those dictated by its opponent."[8]

Discovering the Problem After It Is Solved

A striking coincidence between Kuhn's analysis and the evolution of Rasch's models is the surprising inversion between a problem and its solution: in intellectual revolutions, the solution often comes first through an insight! Using Dalton's work illustratively, Kuhn writes:

". . . all 3 of Dalton's incompatible accounts of the development of his chemical atomism make it appear that he was interested from an early date in just those chemical problems of combining proportions that he was later famous for having solved. Actually, those problems seemed to have

occurred to him with their solutions, and then not until his own creative work was nearly complete.[8]

". . . often no . . . structure is consciously seen in advance. Instead, the new paradigm, or a sufficient hint to permit later articulation, emerges all at once, sometimes in the middle of the night, in the mind of a man deeply immersed in crisis."[8]

From Rasch's reflections:

"The outcome of the reading test experiment was beyond expectations: a statistically very satisfactory analysis on the basis of a new model which represented a genuine innovation in statistical techniques!

"But the understanding of what the model entails tarried several years."[1]

Rasch continues to describe how he showed Ragnar Frisch, a Norwegian economist and later Nobel Prize winner, an equation like equation (3) with no person parameter.

". . . until this point Frisch had only listened politely . . . On seeing [the elimination of parameters] Frisch opened his eyes widely and exclaimed: 'It . . . was eliminated, that is most interesting!' And this he repeated several times during our further conversation. To which I of course agreed every time—while I continued reporting the main results of the investigation and some of my other work.

"Only some days later I all of a sudden realized what in my exposition had caused this reaction from Ragner Frisch. . . .

"What Frisch's astonishment had done was to point out to me that the possibility of separating 2 sets of parameters must be a fundamental property of a very important class of models."[1]

Unmistakably, Rasch's recollections are consistent with Kuhn's analysis of both the *inverted* relationship between a problem and a solution and the way the insight into the *problem that has been solved* emerges. Unmistakably also, Rasch's focus *shifted* from describing data sets to a class of models, which he explained had the same parametric structure as the laws of physics.[1] He used much of the rest of his life connecting these models to epistemology and the theory of measurement.

Newcomer Precipitates a Revolution

A mathematician, Rasch turned to biologic and medical statistics to earn a living. For a year in the 1930s, he studied with Ronald Fisher.[19] In 1952, he helped design a study to monitor progress of pupils with reading difficulties in Denmark. Rasch[1] approached the analyses of these data as he approached biologic and medical data. In particular, he focused on individuals, not samples or populations, as he might have done from the perspective of test theory at the time. Briefly, he parameterized locations of persons and texts as abilities and difficulties, respectively, and arrived at a Poisson model that accounted for the total error count each pupil made on each text. Second, he *deduced* from this model, the model for dichotomous responses that would operate when a person read a single word. *Rasch did not derive the model to describe data for dichotomously scored items*, though he subsequently did apply it to data of course.

The relevance of this distinctive derivation is that Rasch was a newcomer to the field.

"Almost always the men [sic] who achieve these fundamental inventions of a new paradigm have been either very young or very new to a field whose paradigm they change."(Kuhn, 1970, p. 90).[8]

Solving Different Problems

Different paradigms generate different problems.

". . . [S]ince no 2 paradigms leave all the same problems unsolved, paradigm debates always involve the question: Which problems is it more significant to have solved?" (Kuhn, 1970, p. 110).[8]

There seem to be many examples of this phenomenon between the traditional and Rasch paradigms. I take one as an example and touch on others. From the view of the 2PL, finding methods for estimating the discriminations α_i were important problems to be solved,[20] whereas from the view of the SLM, identifying qualitatively why discriminations might be different and how to control these factors qualitatively or experimentally are important problems to be solved.[1,21] These are distinctively different research agendas implied by the 2 paradigms.

Bock[22] summarizes the substantial challenge and work that has been carried out in estimating parameters in the 2PL. These include Bayes' estimation, application of the EM algorithm, and marginal maximum likelihood methods. Each of these involves ancillary assumptions about the distribution of persons (eg, normal) or constraints beyond those required in the SLM. Proponents of this model are convinced that the problems of estimating item discriminations have been solved; proponents of the SLM[23] consider these are various ad hoc solutions to inherently intractable problems that cannot provide stability of estimates. Equation 3, which depends only on the item parameters, can be generalized to any number of items. Because of the elimination of person parameters, no ancillary assumptions need to be made, in particular about the distribution of the persons which is wholly an empirical matter. Rasch was aware that the separation and elimination of parameters may be considered only convenient.

". . . [S]ufficiency may appeal as nothing more than a surprising and singularly nice property, extremely handy when accessible, but, if not, then you just do without it. But to me sufficiency means much more than this. When a sufficient estimate exists, it extracts every bit of knowledge about the specified feature of the situation made available by the data as formalized by the chosen model. . . . .

"The realization of the concept of sufficiency, I think, is a substantial contribution to the theory of knowledge and the high mark of what Fisher did. . . ."[1]

The estimation of parameters in the Rasch model is not trivial, and it generates its own body of statistical work.[24] Estimation studies contribute to an apparent similarity of concerns in the Rasch and non-Rasch models. However, from the traditional paradigm, problems of conditional estimation are not central; they may be convenient if the Rasch model is being applied. In the Rasch paradigm, they are part of a large body of knowledge. For a combination of reasons (eg, substantial degree of missing data and therefore difficulty with conditional estimation), other methods of estimation may be used with the SLM (eg, the EM algorithm, and marginal maximum likelihood).

Recourse to Philosophy

The final criterion of a change of paradigms considered is Kuhn's[8] observation that revolutionary episodes result in recourse to philosophy. Rasch's last paper was published in the Danish Yearbook of Philosophy,[25] dealing with both probabilistic and deterministic models. His models have given impetus to further analysis in the philosophy of measurement.[16,26,27]

Measurement as a kind of mapping on a continuum partitioned into equal units from an origin is understood readily. Measurement is also an advanced concept, the achievement of which has been integral to the remarkable progress of physical science. Not surprisingly, therefore, mathematicians and social scientists have studied measurement in the physical sciences for its implications for measurement in general and in the social sciences in particular. They have articulated the field of *axiomatic, fundamental measurement*, known also as *additive conjoint measurement*.[28–30] Two approaches to fundamental measurement have occurred historically, the classic and representational.[32] Briefly, in the former, measurements are ratios of numbers characterizing properties of objects, where one is taken as the unit, and where these ratios maintain a relative meaning throughout the range of the variable. In the latter, numbers are assigned to objects in which operations on numbers correspond to analogous permissible operations on these objects; for example, where the addition of measurements corresponds to the concatenation of objects, as in classic laws of physics. For their proponents, not only are Rasch models compatible with fundamental measurement and the laws of physics, they can be applied to problems involving measurement.[31–34]

In the literature of fundamental measurement, Stevens's[35] definition that "measurement is the assignment of numerals to objects or events according to rule" is considered superficial and counterproductive to the development of fundamental measurement in social science.[26,32] A related set of problems within the Rasch paradigm, and not of concern in the traditional paradigm, is the mathematical proofs that the SLM is unique in its parameter separation for dichotomous items.[25,36,37] To purists in fundamental measurement, the Rasch model itself may be an example of only necessary, but not sufficient, conditions, to demonstrate that quantitative constructs have been identified and operationalized.[38]

The properties of the Rasch model have given rise to terms such as *item-free* and *person-free*. These can appear to overstate the facilities of the model which are expressions of the requirement of invariance. Within a *conformable* class of persons and items, different subsets of persons (eg, males and females, or persons at different points along the continuum), will give equivalent item parameter estimates, and different subsets of items will give equivalent person parameter estimates, but whether or not they are conformable, as evidenced by tests of fit, is an empirical question. Similarly, data do not provide fundamental measurement just by being analyzed with the SLM; fundamental measurement is a property of the SLM, and therefore whether or not data fit the SLM is also an empirical question.[39] I come later to what follows from the Rasch paradigm when data do not fit the SLM. However, note now the inversion of the relationship between the data and the model. In the Rasch paradigm, because the model provides an operational criterion for fundamental measurement, the data need to fit the SLM; in the traditional paradigm, the model which fits the data is chosen.

A Common Forerunner of IRT and Fundamental Measurement

Although the main distinction between the traditional and Rasch paradigms, and their evolution, has been summarized, this section develops further the substance of this change of paradigm by considering an earlier historical context of IRT and Rasch models. It shows that they can be seen to have a common forerunner, which can also contribute to confusion when considering their differences or similarities. Two historical overviews in the same issue of a journal by Bock[22] and Wright[23] testify to this observation.

Thurstone and Measurement

Bock[22] summarizes the seminal work of Thurstone in the 1920s and 1930s, which was overshadowed by classic test theory including factor analysis to which Thurstone himself turned, as anticipating IRT. Bock reproduces Thurstone's ICCs and estimates of the locations of items on a linear scale for an intelligence test and then describes the differences between Thurstone's and IRT models, emphasizing the challenge and solutions to problems of estimation in the latter. One apparent advance was to use the logistic model rather than the cumulative normal model used by Thurstone. The logistic was used because of its mathematical tractability. The primacy of the cumulative normal model, and the use of the

logistic for convenience given that it approximates the normal, is reflected in the traditional paradigm by writing the 2PL model in the form

$$\Pr\{\text{positive response}\} = \frac{e^{D\alpha_i(\beta_n - \delta_i)}}{1 + e^{D\alpha_i(\beta_n - \delta_i)}} \qquad (4)$$

where D = 1.7 is a scaling factor that makes the response probability virtually identical to the normal.[3,7] Researchers from the Rasch paradigm do not use this scaling factor.

Wright,[23] too, summarizes Thurstone's work. However, instead of describing any of Thurstone's empirical studies, he quotes Thurstone's articulation of the requirements of measurement, the concepts of unidimensionality, linearity, abstraction of measurement, and most importantly, the principle of invariance. He considers Thurstone's analysis of measurement as anticipating Rasch's models.

"A measuring instrument must not be seriously affected in its measuring function by the object of measurement . . . Within the range of objects intended, its function must be independent of the object of measurement."[40]

"It should be possible to omit several test questions at different levels of scale without affecting the individual score."[41]

The short reference to Thurstone here is made for 2 reasons. First, his work is acknowledged generally as seminal in psychometrics and by authors from both paradigms presented here. His use of the normal ogive model anticipates the traditional IRT perspective where the logistic replaces it only because it is more tractable, and his articulation of the requirement of invariance anticipates the criterion used by Rasch. Second, this shows that the criterion of invariance is not unique to Rasch. However, this also leads to a decisive step that Rasch took beyond Thurstone, which helps highlight the distinction between Rasch's work and Thurstone's as seen from the traditional paradigm.

Rasch and Measurement

Following his conversation with Frisch, Rasch presented his principles of invariance, summarized earlier, for making comparisons that lead to measurement.

"The comparison between 2 stimuli should be independent of which particular individuals were instrumental for the comparison . . .

"Symmetrically, a comparison between 2 individuals should be independent of which particular stimuli within the class considered were instrumental for comparison. . . ."[11]

The decisive step taken by Rasch beyond Thurstone is that Thurstone saw invariance essentially as a property required of *data*, while in addition, Rasch characterized it as a *property of a mathematical model*. This has meant that further derivations of its implications can be carried out mathematically, and these include surprising results[42]; for example, that adjacent categories for a response format with ordered categories cannot be collapsed arbitrarily, results which do not arise from the traditional paradigm. As viewed by Thurstone, and consistent with the traditional paradigm, different models can be used to characterize the data, and invariance can be checked by comparing parameter estimates in the different groups. Thus, there is no barrier to using the 2PL for this purpose. However, as viewed by Rasch, this same criterion of invariance generates a restricted class of models compatible with fundamental measurement, precluding consideration of other models.

Rasch wrote his conditions for invariance in a general equation, which, in the probabilistic case, takes the form

$$Pr\{(x_{1j},x_{2j});\beta_n,\delta_i,\delta_j|f(x_{1i},x_{1j})\} = \vartheta(x_{1i}x_{2j},\delta_i,\delta_j), \qquad (5)$$

where the right side of the equation (5) is independent of the person parameter β_n. For dichotomous responses, the solution gives the SLM. In this development, the logistic arises as a mathematical proof requiring the addition or subtraction of parameters; its simplicity, tractability, and separation of parameters are seen as an integral part of its justification for measurement. It is not used because of its tractability and its approximation to the normal ogive model as is the 2PL.

In summary, in the Rasch paradigm, the case for the SLM is neither that it describes any set of data nor that it has convenient properties of estimation, although of course both properties are exploited. The case is that it provides an operational criterion for fundamental measurement of the kind found in the physical sciences. The relevance of the fundamental measurement to proponents of SLM provokes a closer consideration of the functions of measurement in science, which is the third reason for considering the work of Kuhn.

The Role of Measurement

Kuhn[43] turns upside down the traditional view that the function of measurement in science is to discover theories from data; rather, he argues that its function is to disclose *anomalies* in data generated by theories that scientists do "not doubt." An analogous observation is made that because models in the Rasch paradigm are chosen on a criterion independent of data, their role, too, is to disclose anomalies in data relative to the chosen Rasch model that its proponents, too, do not doubt. This lack of doubt, independent of data, can contribute to misunderstanding between proponents of the 2 paradigms.

Measurement and Data Analysis in Physical Science. Kuhn[43] describes how textbooks give the traditional myth of the role of measurement.

"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."

Eventually scientists learn how to make the numbers conform to theory. They do so using measurement in a role that would not be gleaned readily from science textbooks.

"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 phenomenon. To the nature of that phenomenon, they usually provide no clues."

Rasch Measurement and Data Analysis in Social Science

The role of the Rasch models in the Rasch paradigm can be seen as analogous to Kuhn's role of measurement in science and therefore defensible. This role, which would not be gleaned readily from statistics textbooks, was set by Rasch in his first 2 analyses of data with the SLM.

Having deduced the SLM from the model for reading errors of whole texts, Rasch applied it out of curiosity to data that he had available from Raven's[44,45] nonverbal test of reasoning and from a Danish military intelligence test. By the standards he applied, the Raven's data fitted the SLM; the military test data did not—the data did not conform to his hypothesized lines that would imply equal discriminations among items.[1] Thus he had 1 success and 1 failure with his new model.

The next sequence of steps is crucially different from the ones that would have arisen from the traditional paradigm. Instead of modifying the model to account for the different discriminations, he demonstrated the statistical misfit as substantive anomalies; he showed that the test seemed to be composed of different kinds of items. Appreciating the point of this demonstration, the head of the military psychologists instructed a psychologist to construct 4 tests covering the required intellectual tasks but where each test on its own should conform to Rasch's new model. From these steps, the question arises as to which was seen to fail, the data from the original intelligence test or the SLM in Rasch's second application. In parallel to Kuhn's analysis, it seems the model gave clues to problems with the data, and new data were constructed to conform to the model.

When the new tests were analyzed,[1] the tests again did not conform perfectly, although they did "by and large." In some cases, items showed a greater discrimination than the majority. Again pursuing the data for clues, it appeared that the more discriminating items were towards the end of 1 test. This clue led to learning that the supervisor of the test had reduced the prescribed time of testing because of some other competing demand in the class, making it, against the intention, in part a speeded test. Subsequent to the publication of the book, the data were subdivided according to where students had reached, so that nonresponses to subsequent items were not treated as wrong responses. Now the items did conform to the model, showing that for students who worked at different speeds, the relative difficulty values were equivalent. Yen,[46] also noticed that items at the end of a test seemed to show consistently higher discriminations.

The key feature of the approach is that the SLM governed the generation of data and then was not abandoned in cases where the data again did not conform to the model but was used to understand the source of new anomalies disclosed by misfit to it. In this sense, identifying substantive anomalies from statistical misfit, resisting modification to the model, collecting new data guided by the model, is consistent with the role of measurement in physical science enunciated by Kuhn and thereby encourages proponents of the class of Rasch models.

Paradigms and Data Collection

An important aspect of the difference between the Rasch and the traditional paradigms is the sense in which data are sacrosanct. In the former, data are sacrosanct in providing information whether they fit the model or not. The information they provide in the case of misfit is reconsidered substantively and qualitatively, and there is a greater preparedness because of the measurement properties of the model and the conviction that seem to go with it, for new data to be collected to conform to the model.

"To construct measures, we require orderly, cooperating, noncrossing curves like the Rasch curves. . . .This means that we must take the trouble to collect and refine data so that they serve this clearly defined purpose. . . ."[23]

Otis Dudley Duncan, who made major contributions to quantitative social science, in particular introducing path analysis from genetics, eloquently articulated this aspect of the Rasch paradigm.

"We must be clear that Rasch does not pretend to supply the most general model one might want for the study of response structures . . . Most data sets that readily come to hand, whether they originate in testing enterprises or surveys, will exhibit complications . . . and will, therefore, be too 'messy' for the Rasch model to fit well. As long as we are content merely to find models sufficiently flexible to fit available data acceptably well, only by happenstance will we achieve what can be termed 'measurement' in any reasonably rigorous sense of the term. . . . Nor will we know whether

Rasch measurement can be carried out in a deliberate way in surveys until we try hard to do it."[16]

Conforming to the model, of course, is not the only criterion, something that Rasch proponents in their enthusiasm sometimes overlook.

"The Rasch model . . . does not revoke the criteria scientists normally cite in deciding whether right variables have been measured."[16]

In the traditional paradigm, data are more sacrosanct from the point of view to being modeled, and the data are less likely to be abandoned if a better fitting model can be found.

". . . [I]f the proportion of misfitted items is large, the reasonable solution is to discard the Rasch model and try a model that includes discrimination . . . in a systematic manner."[12]

Birnbaum,[3] who seems to have been the first to present the 2PL model in test theory, justified the discrimination parameter as parameter additional to the SLM because data showed empirically different discriminations. As with the application of the SLM, considerations of substantive issues can also arise from misfit to the 2PL, but these considerations are not in principle enforced by the traditional paradigm as they are by the Rasch paradigm.

An explanation of the 2PL in the traditional paradigm, when it fits better than the SLM, is that the discrimination reflects differences in the degree of error or random variations in responses to the items and that it therefore gives relevant statistical weight to the item according to this error; in the Rasch paradigm, these differences are seen as fundamental problems and that this precision among items must be equivalent. The analogy in measuring instruments is that the lines marking the units on the operational continuum are not of different widths! This difference in legitimately different perspectives is again irreconcilable, reflecting incompatible paradigms. For example, although comparing fit of the 2 models to data would resolve the choice of model within the traditional paradigm, because the choice of the SLM does not rest on it describing data, such a comparison does not resolve the choice within the Rasch paradigm.

Rasch was aware that his change of perspective embodied in his models turned the traditional question *upside down*.

"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. Furthermore the application of the model must have something to do with the construction of the test; at least, if a pair of tests showed results in accordance with our theory, this relationship could easily be destroyed by adding alien items to the tests. Anyhow, it may be worthwhile to know of conditions for the applicability of such relatively simple principles for evaluating test results."[1] (Emphasis in original.)

Rationalization of Models

The expression *paradigm shift* has become part of everyday language because people perceive apparent incompatibilities of perspective relatively often. Most of these are not general or important, and it may be overly dramatic to describe them as arising from a revolution. This may be the case for Rasch models. However, Rasch proponents draw on further evidence from studies in fundamental measurement to lend support to the idea of a significant paradigm shift, and possibly a revolution.

In an essay review of Krantz et al's[30] *Foundations of Measurement*, Ramsay,[47] not a particular proponent of the Rasch models, summarized the last chapter of the book in the following terms:

"Also somewhat outside the concerns of the rest of the book, this chapter . . . deals with the fact that virtually all the laws of physics can be expressed numerically as multiplications or divisions of measurements."

and

"The most challenging chapter in my mind is the last; it confronts the remarkable fact that throughout the gigantic range of physical knowledge numerical laws assume a remarkably simple form providing fundamental measurement has taken place. . . . [T]he extension to behavioral science is obvious: we may have to await fundamental measurement before we will see any progress in quantitative laws of behavior. In short, ordinal scales (even continuous ordinal scales) are perhaps not good enough and it may not be possible to live forever with a dozen different procedures for quantifying the same piece of behavior, each making strong but untestable and basically unlikely assumptions which result in nonlinear plots of one scale against another. . . . A rationalization of quantification may be necessary precondition to Psychology as a Quantitative Rational Science."[30]

Rasch proponents see Rasch models providing such a rationalization and consider that these models have the potential to contribute to the social sciences what fundamental measurement has contributed to the physical sciences. If they did so, then the suggestion of a revolution may not be overly dramatic. This perspective does not have to deny that useful instruments and quantification are being carried out from the traditional paradigm; it does, however, change the perspective from providing useful instruments and quantification to the more ambitious challenge of generating data that provide fundamental measurement. This focus on data generation to meet requirements seems to arise more directly from the perspective of physical sciences than statistical modeling of extant data in the social sciences. The conviction of propo-

nents of Rasch models in pursuing this challenge, and often their implicit and explicit argument that other scientists involved in quantification should pursue the same challenge, exemplifies a paradigm difference which can be a source of irreconcilable disagreement.

Summary

This paper suggests that the controversy associated with the Rasch and IRT models when they were introduced in education continues to manifest itself in new fields of application and that to better understand these manifestations, it is useful to understand the original controversy. It seeks to explain this controversy using Kuhn's explication of incompatible paradigms.

In what is termed the traditional paradigm, which arises understandably and defensibly from statistics, when the simpler Rasch models do not account for the data, then more complex models that account better for the data are entertained. Properties of the Rasch models are found convenient but not inherently significant to give them primacy if they do not account for the data. In suggesting and explaining a new paradigm, the paper describes Rasch's own *shift in perspective from modeling sets of data to studying and formalizing a class of models* that turn out to be compatible with fundamental measurement found in the physical sciences. In what is termed the Rasch paradigm, these models become operational renditions of fundamental measurement, and this is the source of conviction for their proponents to want to generate data that conform to them while retaining their substantive validity.

The paper focuses illustratively on the choice between the Rasch SLM and the IRT 2PL models for dichotomous items. From the perspective of the traditional paradigm, the SLM is considered because it is (i) after an arbitrary scaling, virtually identical to the normal ogive; (ii) it is tractable; and (iii) it is the simplest of models for dichotomous responses. Thus, if it fits dichotomous data, it is the preferred model; if it does not, the 2PL with different item discriminations and which is an algebraic generalization of the SLM, is considered. From the Rasch paradigm, the SLM is used because it arises from a mathematical formalization of invariance which also turns out to be an operational criterion for fundamental measurement. From this perspective, the SLM, which is the unique model with this property for dichotomous responses, does not generalize to the 2PL and therefore the 2PL is not considered when the data do not fit the SLM.

When misfit between the data and the SLM is identified, the main challenge in the traditional paradigm is to those with expertise in statistics or data analysis to identify a model that accounts better for the given data, notwithstanding that they may find other problems in the data; the main challenge in the Rasch paradigm is for the those with expertise in the substantive field of the construct to understand the statistical misfit as substantive anomalies and, if possible or necessary, to generate new data that better conform to the model while also enhancing substantive validity of the variable. In only the second application of the SLM, in which the model did not account for the data, Rasch set such a precedent.

From the traditional paradigm, proponents of Rasch models within the social sciences may be considered unduly optimistic in pursuing fundamental measurement and quantitative laws of the kind found in the physical sciences. However, it is, apparently, because of this optimism that they give primacy to Rasch models over competing models when initially attempting quantification. New researchers applying the models to new fields can be legitimately puzzled when studies are reported from the perspectives of 1 of 2 paradigms which, while covering the same issues of statistical analysis of data, in principle have aspects that are irreconcilable.

ACKNOWLEDGMENTS

An earlier version of this paper was presented at of the International Conference on Objective Measurement: Focus on Health Care, Chicago, October 2001. The paper draws on a recorded interview with Rasch in 1979 in which he reflected on his work. Two unknown reviewers and Irene Styles provided valuable comments on an earlier version of the paper. This research was funded in part by the Australian Research Council.

REFERENCES

1. Rasch G. *Probabilistic Models for Some Intelligence and Achievement Tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition 1983. Chicago: MESA Press; 1960.
2. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley; 1968.
3. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley; 1968:397–545.
4. McHorney CA, Haley SM, Ware JEJ. Evaluation for the MOS SF-36 Physical Function Scale (PF-10), II: comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol*. 1997;50: 451–461.
5. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustration with functional status items. *Med Care*. 2000;38(supp II):II-43–II-59.
6. Cella D, Chang C-H. A discussion of item response theory and its applications in health status assessment. *Med Care*. 2000;38(suppl II):II-66–II-72.
7. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*. 2000;38(suppl II):II-28–II-42.
8. Kuhn TS. *The Structure of Scientific Revolutions*. 2nd ed. Chicago: The University of Chicago Press; 1970.
9. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. *Med Care*. 2000;38(suppl II):II-28–II-42.
10. Bock RD, Jones LV. *The Measurement and Prediction of Judgment and Choice*. San Francisco: Holden Day; 1968.
11. Rasch G. On general laws and the meaning of measurement in psychology. In: Neyman J, (ed). *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press; 1961:321–333.
12. Divgi DR. Does the Rasch model really work for multiple choice items? not if you look closely. *J Educ Measure*. 1986;23:283–298.

13. Goldstein H. Consequences of using the Rasch model for educational assessment. *Br Educ Res J.* 1979;5:211–220.
14. Traub RE. A priori considerations in choosing an item response model. In: Hambleton RK, ed. *Applications of Item Response Theory.* Vancouver, Canada: Educational Research Institute of British Columbia; 1983.
15. Loevinger J. Person and population as psychometric concepts. *Psychol Rev.* 1965;72:143–155.
16. Duncan OD. Rasch measurement: further examples and discussion. In: Turner CF, Martin E, eds. *Surveying Subjective Phenomena.* Vol. 2. New York: Russell Sage Foundation; 1984.
17. Andrich D. An elaboration of Guttman scaling with Rasch models for measurement. In: Brandon-Tuma N, ed. *Sociological Methodology.* San Francisco: Jossey-Bass; 1985:33–80.
18. Masters GN. Item discrimination: when more is worse. *J Educ Measure.* 1988;24:529–544.
19. Andersen EB, Olsen LW. The life of Georg Rasch as a mathematician and as a statistician. In: Boomsma A, van Duijn MAJ, Sniders TAB, eds. *Essays in Item Response Theory.* New York: Springer; 2000.
20. Wingersky M. SLOGIST: a program for computing maximum likelihood procedures for logistic test models. In: Hambleton RK, ed. *Applications of Item Response Theory.* Vancouver, Canada: Educational Research Institute of British Columbia; 1983.
21. Wright BD, Stone MH. *Best Test Design: Rasch Measurement.* Chicago: MESA Press; 1979.
22. Bock RD. A brief history of item response theory. *Educ Measure Issues Pract.* 1997;16:21–33.
23. Wright BD. A history of social science measurement. *Educ Measure Issues Pract.* 1997;16:33–45.
24. Andersen EB. The numerical solution of a set of conditional estimation equations. *J Royal Stat Soc.* 1972;3:42–54.
25. Rasch G. On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook Philos.* 1977;14:58–94.
26. Duncan OD. *Notes on Social Measurement.* New York: Russell Sage Foundation; 1984.
27. Fisher WPJr. Objectivity in measurement: a philosophical history of Rasch's separability theorem. In: Wilson M, ed. *Objective Measurement: Theory into Practice.* Vol. 1. Norwood, NJ: Ablex Publishing Corporation; 1992:29–58.
28. Roberts FS. *Measurement Theory.* Reading, Mass: Addison-Wesley; 1979.
29. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol.* 1964;1:1–27.
30. Krantz DH, Luce RD, Suppes P et al. *Foundations of Measurement.* Vol 1. New York: Academic Press; 1971.
31. Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. *Appl Psychol Measure.* 1979;3:237–256.
32. Michell J. Numbers as quantitative relations and the traditional theory of measurement. *Br J Philos Sci.* 1994;45:389–406.
33. Brogden HE. The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika.* 1977;42:631–634.
34. Wright BD. Campbell concatenation for mental testing. *Special Interest Group: Rasch Measurement.* 1988;2:3–4.
35. Stevens SS. On the theory of scales of measurement. *Science.* 1970;103: 677–680. [Reprint. In: Haber A, Runyon RP, Badia P, eds. *Readings in Statistics.* Reading, Mass: Addison-Wesley; 1946.]
36. Wright BD. Additivity in psychological measurement. In: Roskam EE, ed. *Measurement and Personality Assessment.* Selected papers, XXIII International Congress of Psychology, Vol. 8, (*pp.*). Amsterdam: North Holland; 1985:101–111.
37. Fischer GH. Derivations of the Rasch model. In: Fischer GH, Molenaar IW, eds. *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer; 1995:15–38.
38. Michell J. Rasch models and the prospect of measurement. *J Appl Measure.* 2001;2:202–204.
39. Andrich D. Relationships between the Thurstone and Rasch approaches to item scaling. *Appl Psychol Measure.* 1978;2:449–460.
40. Thurstone LL. Attitudes can be measured. *Am J Sociol.* 1928;33:529–554.
41. Thurstone LL. The scoring of individual performance. *J Educ Psychol.* 1926;17:446–457.
42. Andrich D. Models for measurement, precision and the non-dichotomization of graded responses. *Psychometrika.* 1995;60:7–26.
43. Kuhn TS. The function of measurement in modern physical science. *Isis.* 1961;52:161–190.
44. Raven JC. Matrix tests. *Ment Health (Lond).* 1940;1:10–18.
45. Raven J. The Raven progressive matrices: a review of national norming studies and ethnic and socioeconomic variation within the United States. *J Educ Measure.* 1989;26:1–16.
46. Yen WM. The extent, causes and importance of context effects on item parameters for two latent trait models. *J Educ Measure.* 1980;17:297–311.
47. Ramsay J. Review of *Foundations of Measurement,* Vol. I, by D. H. Krantz, R. D. Luce, P. Suppes and A. Tverskey. *Psychometrika.* 1975; 4O:257–262.