British Educational Research Journal

Vol. X, No. X, Month 200X, pp. 000-000

The Assessment Revolution that has passed England by: Rasch Measurement

Panayiotis Panayides[*]*, Colin Robinson*[#] *& Peter Tymms*[*]

[*]*Durham University, UK*

[#] *Freelance*

Corresponding Author:

Peter Tymms

CEM Centre

University of Durham

Mountjoy Research Centre 4

Stockton Road, Durham DH1 3UZ

UK

Email: Peter.Tymms@cem.dur.ac.uk

The Assessment Revolution that has passed England by:

Rasch Measurement

Abstract

Keywords:

Assessment has been dominated by Classical Test Theory for the last half century although the radically different approach known as Rasch measurement briefly blossomed in England during the 1960s and 70s. Its open development was stopped dead in the 80s whilst some work has continued almost surreptitiously. Elsewhere Rasch has assumed dominance. The purpose of this article is to discuss the major criticisms of the Rasch model, which led to its rejection by some, and to give responses to these criticisms whilst encouraging social scientists to appreciate its strengths. The original breakthrough by Georg Rasch in 1960 has been developed and extended to address every reasonable observational situation in the social sciences.

Introduction

This paper starts with an historical perspective of assessment developments in England in the 70s and 80s. This outlines how traditional approaches to the analysis of test data were shown to be inadequate for the purposes at hand and how a new methodology was adopted and extended. This new approach was stopped abruptly but was continued elsewhere. The paper then outlines the theoretical basis of the Classical and revolutionary approaches and goes on to examine the criticisms which led to the abandonment of the Rasch approach to measurement.

Developments in the 70s and 80s

In the 1970s in the UK, there was great interest in the evaluation of the effectiveness of the education system and particularly trends of performance over time. Since the 1940s, reading had been assessed on a regular basis using unchanging standardised reading tests. Although for many years the pupils' scores on the test seemed to rise, in the early part of the decade they appeared to have declined. An investigation by Start and Wells (1972) suggested that the change might be caused by the test becoming dated and therefore results were no longer comparable. The use of such tests to show trends over long periods of time was called into question and new ways of monitoring the system were needed.

The response was the establishment of the Assessment of Performance Unit (APU) which argued that, in order to discover what was being taught and how effectively it was being learnt – "a broad balanced picture of pupils' performance" (APU 1979), it would be necessary to have extensive assessments. Even within a single curriculum area, these would need to cover a range of content – e.g. science would need to span at least biology, chemistry, and physics. Similarly a wide range of different assessment types would be necessary, including

written, oral and practical tests. It was estimated that in order to cover the full range of the science curriculum alone it would be necessary to have 36 hours of assessment. Obviously this was impossible. What was needed was a system of assessments which could cover the curriculum adequately but in which an individual student would take only a relatively small subset. If these assessments could then be put together on a single scale, it would be possible to draw conclusions about what topics were being taught well and which were not.

Experience in the USA was considered and the National Assessment of Educational Progress (NAEP), which had been set up in the late 60s, proved to be of particular interest. Clare Burstall, Deputy Director at the National Foundation for Educational Research (NFER), and Brian Kay, Head of the APU reported:

> *"Since NAEP had been given the task of assessing changes in the educational achievement of pupils and young adults in four different age groups which, together, made up a population of about 37 million, there was really never any possibility that a 'blanket' approach to assessment could have been adopted. In addition, it had been agreed that no student in the 'in-school' samples should be asked to give more than one class period of his time to the assessment programme. This meant, in effect, that no student could be given more than one package of the exercises prepared for use in any given cycle of assessment." (Burstall and Kay 1978 p 35).*

The approach favoured by NAEP was matrix sampling – small groups of randomly selected pupils taking small groups of test items.

The problem was how to deal with the data. Classical Test Theory could only compare items if they were all taken by the same group – or very closely matched groups. To design equivalent tests across such a wide area would be almost impossible.

The Examinations and Tests Research Unit (ETRU) was set up at the NFER by the Schools Council in 1964 and, in 1966, a pilot study was commissioned into "the feasibility of establishing banks or libraries of examination questions or items suitable for measuring the achievement of 16-year-olds taking examinations in various subjects". (Wood & Skurnik 1969 p 1). The principal focus of these item banks were to be the delivery of school-based assessment as part of the newly devised Certificate of Secondary Education (CSE) and the approach proposed was based upon the procedures then used for Mode 3 examinations in which teachers devised both the syllabus and its assessment.

The statistical analyses were based upon Classical Test Theory, but in an appendix to the Wood and Skurnik report, Bruce Choppin described a method of arriving at sample-free estimates using pairwise comparisons of all the items in a test. (Choppin 1969 pp 134-140). This was based upon the model proposed in 1960 by the Danish mathematician, Georg Rasch, who had devised it originally for reading tests. A great deal of work on extending the model into the wider educational sector was being done in the USA by Ben Wright and his associates at the Measurement, Evaluation, Statistics and Assessment (MESA) unit at the University of Chicago. Bruce Choppin was one of a number of those who took on board the Rasch approach and disseminated it widely. Others included David Andrich and Geoff Masters who developed the procedures in Australia. They took the procedure much further than Rasch had envisaged: "*I do not expect this model to hold at all if applied to items belonging to different fields of mathematics*." (Rasch 1969 p 100)

The APU had two problems for which the Rasch model was seen as a potential solution. The first was to provide a means of comparing the difficulties of items used in different contexts and taken by different groups of pupils across a wide range of attainments. This would give the necessary information about the overall achievements of children across the curriculum as a whole. The second was to provide a metric that would allow changes in performance to be compared at different points in time. The first surveys were carried out in 1978 and were continued until 1988.

In 1978, the NFER decided to make use of the approaches developed for the APU to create a parallel bank of items that could be used by Local Education Authorities and schools to create custom built tests that matched their own curriculum but which could also be compared with national data to provide a check on comparative standards. The LEAs' and Schools' Item Bank (LEASIB) was seen as a potential replacement for the numerous standardised tests. The first Head of LEASIB was Alan Willmott, previously Principal Research Officer in the Examinations and Tests Research Unit. Both APU and LEASIB fell under Bruce Choppin's overall leadership.

The LEASIB process was simple and had been expounded in Choppin's appendix. A bank of items would be developed, extensively trialed and calibrated using the Rasch model. A user would be presented with a range of items in the appropriate curriculum area. Selection would be based on a wide variety of features, depending on what the purpose of the test might be. For example it might be appropriate to assess a particular aspect of mathematics one year and a different one the next. Other item characteristics, such as difficulty, could be used to design tailored tests. It was certainly not intended to be a random selection of items but would cover

areas that one might expect students to have covered at that stage in their career. The candidates would then take the test and their responses used to calculate an estimate of their abilities, but their data would also be used to refine the information stored in the bank.

The APU mathematics surveys used Rasch measurement for its first five consecutive years starting in 1977. However, Choppin saw that the use of Rasch measurement was already under attack: "*There are also growing doubts in my mind as to whether the APU is going to be allowed to monitor change except in one or two rather trivial aspects. APU activity in itself appears to be controversial even before we have any results. There are statisticians advising the DES that monitoring performance over time is impossible …*" (Choppin 1981). As a result of the criticisms, two seminars were convened by the DES, *The Rasch Model* in 1980 and *Monitoring over time* in 1981. Foremost amongst the critics was Harvey Goldstein (see for example Goldstein 1979).

One of Choppin's supporters, John M Linacre asserts "*Under Choppin's supervision British psychometrics could have led the world (to the great benefit of British students, teachers, and policy makers). Instead the entrenched interests condemned Britain to a 60 year regression.*" (Linacre 1995).

The rejection of Rasch measurement within NFER caused great concern amongst staff who wrote to the NFER Board complaining that insufficient attention was being given to its own staff – particularly Tony James, its chief statistician, and Alan Willmott. Nevertheless, both the NFER and the APU bowed under the pressure and Rasch was abandoned as a means of tracking changes over time although it did continue to be used to link data within years. Bruce Choppin resigned in 1981 and Alan Willmott the following year. LEASIB was discontinued,

though its bank of items continued to be used by the NFER-Nelson Publishing Company as the basis of custom made tests. Meanwhile Rasch measurement continued to thrive in other parts of the world, notably Australia and the USA, where its theoretical base was considerably expanded. The major international assessments (TIMSS, PIRLS and PISA) all use Rasch measurement or some Item Response Theory approach.

Such was the impact of the 1981 events that the British Educational Research Journal has not published a paper mentioning Rasch since Preece (1980).

The two testing theories

For much of the middle part of the twentieth century Classical Test Theory dominated the approach to testing across the world although it was well know that there were problems with it. A number of different approaches were developed. Amongst them Item Response Theory (IRT) was of major importance and in its simplest form, the so-called one parameter model equates to the approach taken by Rasch. But whilst Rasch purists think in terms of creating instruments for measurement others think in terms of modeling the data using IRT.

In the following sections the various approaches are outlined. This includes a description of the distinct approach taken by Rasch measurement and an outline of the criticisms leveled it together with responses to those criticisms.

Classical Test Theory (CTT)

CTT starts with the model, **X = T + e**, where X is the observed score of an examinee on the test, T the true score (which is conceptualized as the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument) and e the error.

The model has the following assumptions:

(i) T is constant, changes in X are due to error

(ii) Errors are random and they do not correlate with T or with each other.

These assumptions together with the theoretical definition that: reliability is the proportion of variation in observed scores attributable to true scores (i.e. r_{xx} = variance of true scores/variance of observed scores) have led to the formulae for the reliability and the standard error of measurement:

$r_{xx} = 1 - \frac{S_\varepsilon{}^2}{S_x{}^2}$ and SEM = $S_x\sqrt{1-r_{xx}}$, where S_x^2 is the variance of the group's observed scores, S_ε^2 is the error variance and SEM is the standard error of measurement.

In item analysis, psychometricians employing CTT use two basic indices, item difficulty and item discrimination. Item difficulty is calculated by dividing the mean score of the item by the maximum possible score. If items have only one correct answer, which is worth one mark, then this index represents the percentage of examinees responding correctly.

The index of discrimination (D) for any item is the difference of the averages of two groups of examinees (the high and the low scorers) for the specific item, divided by the maximum

possible score on the item. The precise composition of the two groups varies from study to study but the basic definition remains.

The item-total correlation coefficient can also be used as an index of discrimination. The higher the correlation between the scores on a particular item and the total score on all other items, the better discriminator the item is.

Hambleton, Swaminathan and Rogers (1991) identify the following limitations of CTT:

- Ability scores of individuals are item dependent (i.e. they depend on the item difficulties).
- The item statistics (difficulty, discrimination, reliability) are examinee dependent. Discrimination indices as well as reliability estimates tend to be higher in heterogeneous examinee groups than in homogeneous ones.
- No information is available about how examinees of specific abilities might perform on a certain test item.
- Equal measurement error is assumed for all examinees (this measurement error is item dependent).
- Classical item indices are not invariant across subpopulations (i.e. different subgroups of the sample of examinees give different item statistics).

Further, as Anastasi and Urbina (1997) note:

> "*Item difficulty clearly depends on the ability of the group of test takers. This affects also the distribution of scores. In high ability groups the distribution is negatively skewed whereas in low ability groups it is positively skewed. It is preferable to add/revise or delete items so that the score distribution in the target group is approximately Normal"*. (pp 177-178)

Item Response Theory (IRT)

IRT provides alternative models to CTT with the following desirable features:

- Item characteristics are not group dependent.
- Scores describing examinees' abilities are not test dependent.
- A measure of precision for each ability score is produced.
- The probability that an examinee of any ability will answer items of any difficulty correctly is estimated.

As noted earlier the simplest form of IRT corresponds to the approach taken by Rasch (1960) and is sometimes referred to as the one-parameter IRT model. It deals with all the issues set out by Hambleton et al. (1991), and has some important advantages over other IRT models, which are discussed later.

The Rasch approach

A pupil may be given a test item which he or she could easily solve, and yet get it wrong. Similarly, a pupil may be given a test item which is too difficult and yet solve it. Rasch (1960) saw that "*We can never know with certainty how a pupil will react to a problem, but we may say whether he has a good or a poor chance of solving it*" (Rasch, 1960, p 11). This realisation led him to shift from deterministic models to probabilistic models; ones in which "*the possible behaviour of a pupil is described by means of a probability that he solves the task*" (Rasch, 1960, p 11). He also saw that, the probability for a right answer must only be governed by the candidate's ability (θ) and the item's difficulty (b).

> "*The ability of the person and the difficulty of the item must be considered to be joined or conjoint in all analyses of responses and a principle of relativity with*

respect to the item must underlie the task of measurement. This principle overcomes the problems that were raised in earlier decades and that claimed that measurement was not possible in the social and behavioral sciences."

(Keeves and Alagumalai, 1999, p 25)

Rasch set out the following formula for dichotomously scored performances:

$$\ln\left(\frac{\text{Pr}\,obability\ of\ \ success}{\text{Pr}\,obability\ of\ \ failure}\right) = \ Ability - Difficulty$$

Then with simple mathematical steps he deduced the formula for a person's probability of scoring 1 rather than 0 on item i:

$$P_i(\theta) = \frac{\exp(\theta\ - b_i)}{1 + \exp(\theta\ - b_i)}$$

where θ is the ability of the person and b_i the difficulty of item i.

Rasch based his model on three key assumptions: unidimensionality, local independence and invariance which are discussed later.

The basic model is applicable to tests with dichotomous items which can be marked as right (1) or wrong (0). But many tests and questionnaire involve items which are not simply right are wrong and the basic model has been extended to deal with such polytomous items. If the test has a single type of item with the same number of marks available then the Rating Scale Model (RSM) applies. This is widely used for the analyses of Likert scales, even though the original intention of Andrich (1978), who developed it, was to use it in the evaluation of written essays.

If the marks allocated to items vary, then the Partial Credit Model, developed by Masters (1982), is appropriate. Situations where the Partial Credit Model is applicable are discussed by Bode (2004).

More complex IRT models

IRT models are sometimes extended to take into account the differing discriminations of each item. To do this an additional parameter (α) is added to the basic equation making it a "two-parameter" (2-P) model. A further refinement produces the "three-parameter" (3-P) model in which a guessing parameter (c), called by Hambleton et al. (1991, p 17) a *pseudo-chance-level parameter,* is added.

Comparing the 2-P and 3-P models with Rasch measurement

Wright (1983) argues that fundamental measurement in the social sciences is obtainable only through the Rasch approach and, in comparing Rasch with the 2-P and 3-P models, states:

> "*If measurement is our aim, nothing can be gained by chasing extra item parameters like c and a. We must seek, instead, for items which can be managed by an observation process in which any potentially misleading disturbances which might be blamed on variation in possible c's and a's can be kept slight enough not to interfere with the maintenance of a scale stability sufficient for the measuring job at hand. ... Only the Rasch process can maintain units that support addition and so produce results that qualify as fundamental measurement.*"
>
> (Wright, 1983, p 7)

Furthermore, the Rasch approach is the only one which uses the raw score as the sufficient statistic for estimating item difficulty or person ability. That is, the sufficient statistic for estimating person ability is the sum or count of the correct responses for a person over all items. In the other two models the sufficient statistic for ability estimation includes other parameters that must be estimated simultaneously.

In comparing the 2-parameter and 3-parameter models with the Rasch approach it is important to distinguish between measurement and modeling. If the purpose is to construct a good measure then the items and the test should be constrained to the principles of measurement. If on the other hand the purpose is to model some test data then the model which fits the data best should be chosen. Rasch corresponds to the principles of measurement whereas other IRT models correspond to modeling. Fischer and Molenaar (1995) state that:

> "*They (the 2-p and 3-p models) make less stringent assumptions (than the Rasch model), and are therefore easier to use as a model for an existing test. On the other hand, they typically pose more problems during parameter estimation, fit assessment and interpretation of results. Whenever possible, it is thus recommended to find a set of items that satisfies the Rasch model rather than find an IRT model that fits an existing item set.*"
>
> (Fischer and Molenaar, 1995, p 5)

Linacre (1996) adds to the above that allowing or parameterising discrimination or guessing, which are sample dependent indices, limits the meaning of the measures to just that subset of items and persons producing these particular data. This prevents any general inferences over all possible items probing that construct among all possible relevant persons.

A further important difference is the sample sizes required for the calibrations. The use of the 2-P or 3-P models requires larger samples of persons for calibrations. Thissen and Wainer (1982) determined the number of persons with normally distributed abilities necessary to produce an item difficulty that was accurate to one decimal place (i.e. s.e = 0.05). For the Rasch model approximately 2500 persons were needed whereas, for the 2-P model approximately 7500 and for the 3-P model approximately 67000.

Applications of the models

Rasch measurement has been applied in very diverse situations and six examples are outlined below:

- Prieto, Roset and Badia (2001) have explored the Spanish version of the assessment of Growth hormone deficiency in adults and confirmed its unidimensionality and construct validity using the Rasch approach.
- Bond and Fox (2001) describe how data from Piagetian interviews have been analysed using the Rasch approach to give fresh insights.
- Massof and Fletcher (2001) have evaluated the validity of, and suggested improvements to, the visual functioning questionnaire which is designed to assess health-related quality of life of patients with visual impairment.
- Chen, Bezruczko and Ryan-Henry (2006), driven by the need of health and social agencies to have systematic means of describing mothers' effectiveness in caregiving for their adult children with intellectual disabilities, have used Rasch analyses.
- Myford and Wolfe (2002) examined a procedure for identifying and resolving discrepancies in examiners' ratings.

- Lamprianou (2006) investigated the stability of two marker characteristics across tests: (a) severity and (b) consistency of marking.

The above selection of applications of Rasch measurement shows the diversity of situations in which this approach can be used productively over and above the usual assessments of ability in educational tests, the positioning of persons on the latent trait line in psychological tests and the identification of aberrant response patterns in tests or psychometric scales.

Rasch's different approach to the data-model relationship

Although the exponential models were known by the time Rasch worked with them he did not use them in the traditional way. As Andrich (2004) notes, the reason that Rasch's model turns the traditional data-model relationship upside down is that the model does not describe any data. "*The model renders in mathematical, and most importantly from a practical and applied prospective, testable form, the requirements of measurement*" (p 172). Andrich is referring to the requirements of invariant comparisons and quotes Rasch (1961) summarising those requirements:

> "*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 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 some other occasion.*" (Andrich, 2004, p 173)

Andrich (2004) argues that this is fundamentally a different approach to the data-model relationship. He equates the new approach to a paradigm shift of the type identified by Kuhn (1962) and draws parallels with other paradigm shifts and the criticisms that they drew from "experts" at the time only to become orthodox later.

Criticisms of the Rasch model

Goldstein (1979) outlined several criticisms of the Rasch model as did Dickson and Kohler (1996) when commenting on the appropriateness of Rasch measurement being used for transforming the responses of patients to the Functional Independence Measure (FIM) items from the ordinal scale to an interval one. (FIM records the severity of disability of rehabilitation patients). Others including Divgi (1986, 1989) Whitely and Dawis (1974) and Whitely (1977) have also criticised the Rasch approach. But between them, Goldstein (1979) and Dickson and Kohler (1996) cover the majority of the points and it was primarily Goldstein's criticisms that led to a severe reduction in the use of Rasch in the UK.

The major criticisms are outlined below and discussed.

Criticism 1: Unidimensionality

Goldstein's (1979) first criticism, and probably the most frequently occurring one, refers to the assumption of unidimensionality and more precisely to the fact that in order to fit the Rasch model the items must "*relate only to one underlying dimension of ability*" (p 214). He differentiates the Rasch approach from factor analysis (as methods for detecting the dimensionality of data) noting that in factor analysis "*the dimensionality or number of factors*

is studied in the analysis itself" (p 214), implying the superiority of factor analysis. Dickson and Kohler (1996) also criticise the requirement of a one-dimensional latent space.

Response to criticism 1

Since Goldstein's article, many psychometricians (see for example Hambleton et al., 1991; Keeves and Masters, 1999; Smith, 2004; Wright and Linacre, 1989) have made it clear that unidimensionality does not implicitly mean only one factor or dimension but rather the presence of a dominant dimension with the possible presence of minor dimensions.

Hambleton (1993) writes "*the unidimensionality assumption cannot be strictly met because there are always other cognitive, personality and test-taking factors that affect test performance, at least to some extent*" (p 150). Possible factors include test motivation, test anxiety, speed of performance, test sophistication, reading proficiency and other cognitive skills.

Linacre (1998) concurs noting that the presence of more than one dimension in the data does not necessarily imply substantive multidimensionality. Extra dimensions may reflect different person response styles or different item content area. For example, items on subtraction may define a different dimension than items on addition in a simple mathematics test for young children. Multidimensionality can also be an artifact of test construction. For example, including the identical item several times in a test produces a subset of highly intercorrelated items which may define an extra dimension. On the other hand, the use of different response mechanisms across items (multiple-choice, constructed-response, rating scales) introduces unmodeled variation which can be attributed to a dimension of 'item type'.

Multidimensionality becomes a real concern when the response patterns indicate the presence of two or more dimensions so disparate that it is no longer clear what latent dimension the Rasch dimension operationalises.

As far as factor analysis is concerned, Linacre (1998) showed that Rasch analysis followed by principal components analysis of standardized residuals was always more effective at both constructing measures and identifying multidimensionality than direct factor analysis of the original response-level data.

Principal components analysis of the standardized residuals is based on the specification of 'local independence', which is an assumption of the Rasch model. This asserts that, after the contribution of the measures to the data has been removed, what is left is random, normally distributed noise. Therefore the standardized residuals are modeled to have unit normal distributions which are independent and so uncorrelated. This is testable. If the resulting common factors explain nothing more than random noise across items, then the data conform to the Rasch model. The existence of substantive common factors, however, would indicate departure from unidimensionality.

Criticism 2: The use of probabilities

Dickson and Kohler (1996), in listing the shortcomings of the Rasch model, claim that 'any system of measurement based on probabilities must necessarily be imprecise' (p 161).

Response to criticism 2

All measurement is made with error and an explicit acknowledgement that this is so can allow the researcher to express test success in probability terms. The Rasch model does not introduce probabilities or imprecision into the data, on the contrary, it capitalises on their presence in the data to construct a measurement system.

Criticism 3: The absence of distributional descriptions

Dickson and Kohler (1996) criticize also the fact that no description of the sample distribution exists in Rasch analysis.

Response to criticism 3

The Rasch model does not need to assume anything about the distribution of the sample. This is a strength and means that it can reveal the underlying distribution and is not dependent on assumptions about hypothesised distributions.

Criticism 4: Constancy of item difficulties

Goldstein (1979) refers to the fact that the relative difficulty of the items in a test is the same for all individuals. He states: "*Hence, even if we were satisfied that a test tapped only one dimension of ability, in order to use the Rasch model we would also require that, despite different experiences, learning sequences etc., the difficulty order of items was the same for every individual*" (p 214), implying that because of different experiences, learning sequences etc. the difficulty order could not be the same for everyone.

Dickson and Kohler (1996) also criticise the assumption.

Response to criticism 4

Both Goldstein and Dickson and Kohler are referring to the property of invariance. This basic principle of order (or invariance) is not only an assumption of the Rasch model, but also the fundamental requirement for measurement.

Linacre (1996) argues that this is a virtue and not a flaw of the model.

> "*Constant item parameters imply a constant construct. Different item parameters across samples of the relevant population imply that the construct has changed. Then measures cannot be compared across samples, and we are reduced to a vague notion of what we are measuring.*" (Linacre, 1996, p 513)

Rasch, was not the first to require the same kind of invariance in social measurement. L. L. Thurnstone and L. Guttman, two of the most significant people in this field, both articulated this requirement and according to Andrich (2004)

> "*This leads to another reason that the Rasch models can be subtle. Because the property of invariance is built into a mathematical model, it is possible to study the consequences of the requirements of invariance by mathematical derivations.*" (Andrich, 2004, p 174)

Although invariance is a requirement of Rasch models, and of measurement, it is not an assumption for an analysis, in that one can test its veracity.

Criticism 5: Local independence

A different criticism refers to the assumption of local independence, which according to Goldstein (1979, p 214) means that "*for any individual, the response to an item is completely*

independent of his or her response to any other item", again implying that this is not easy to find in practice.

Response to criticism 5

What the assumption essentially means is that previous items should not give hints, clues, insights or guidance for the solution of other items. Such an assumption is more like common sense, and can easily be met by experienced test constructors. Athanasou and Lamprianou (2002), give an example of an item with sub-questions in simple arithmetic calculations.

"*There are 18 flowers in John's garden.*

(a) If he plants 6 flowers more, how many flowers will there be in total? Answer …….

(b) If you need double the number of flowers, how many flowers will you need?

Answer ……."

These two parts of the item cannot be treated as different and independent. If a pupil is not in a position to find the answer to the first part, he/she will not find the answer to the second part even if he/she is able to double a number correctly. This would be a valid criticism of an item banking system in which items are randomly selected for the test. However, with test constructors involved in the development, this is one aspect that would be checked.

Criticism 6: Symmetry between items difficulties and individual abilities

Goldstein (1979) also notes that the Rasch model "*seems to imply a symmetry between item difficulties and individual abilities … In reality, however, this is not the case*" (p 215)

Response to criticism 6

This appears to be a misunderstanding by Goldstein. The reference is presumably to the graphical representation known as the item-person map which often appears to be symmetrical. But the Rasch approach does not require such symmetry.

Criticism 7: Items need to be equally discriminating

Dickson and Kohler (1996) refer to the assumption that the Rasch model requires items to have equal discriminating power. An extension to that is Goldstein's (1979) argument that introducing a discrimination parameter makes the model more flexible and it is no longer necessary to have a constant relative difficulty between items. Although he acknowledges the increase in '*technical problems*' he states that "*Because of its greater flexibility we can expect the model to have a better chance than model (3) (the Rasch model) of fitting a set of test scores.*" (Goldstein, 1979, p 215)

Response to criticism 7

As noted earlier the aim of measurement should not be to accommodate the test data but to satisfy the requirements of measurement. The aim is to measure, not to model. The 2-P model, which introduces a discrimination parameter, seeks to fit a model to the data not vice versa.

Rasch measurement needs items to have discriminations that are equal enough to be regarded as the same. In practice, according to Linacre (1996), unequal discrimination is diagnostic of various types of item malfunction and misinformation. Allowing or parameterising discrimination, which is a sample-dependent index, limits the meaning of the measures to just that subset of items and persons producing this particular set of data. This prevents any general inferences over all possible items probing that construct among all possible relevant persons.

Criticism 8: The model is not perfect

Dickson and Kohler (1996) criticise the Rasch model in that no item fits the model exactly.

Response to criticism 8

The idea that the world is not perfect is not new. We use circles to approximate all sorts of round shapes and straight lines to describe objects that are not perfectly straight. If we were to stop investigations when things were not perfect we would do nothing.

A nice way of viewing the criticism comes from Andrich's (2004) paper where he argues that the Rasch approach, instead of simply describing data, provide the opportunity to understand data by the exposure of anomalies which is the prime function of measurement. The reason why the approach can be used this way is that it formalizes conditions of invariance, which lead to properties of measurement. Thus, when the data deviate from the Rasch it deviates from the requirements of measurement.

Similarly Linacre (1996) does not see non-fitting data as a criticism of Rasch measurement but of the data. He concludes (p 512) that "*usually, if the data have any meaning at all, they can be segmented into meaningful subsets that do fit the Rasch model and do support inferences*", implying that even if the data are not unidimensional, when grouped appropriately (separating the dimensions) they will separately fit the Rasch model.

Criticism 9: All people do not fit the model

With regard to the persons' response patterns and whether meaningful inferences can be made from these response patterns, Dickson and Kohler (1996) comment that they have seen people

who could climb stairs (success on a difficult item) but not being able to swallow (failing an easy item). The implied question in their argument is 'how can one make a meaningful inference from such a performance?'

Response to criticism 9

Again, when data do not fit the model they provide interesting anomalies to be investigated and to challenge the supposed scale. These anomalies are predicted by the Rasch approach to occur occasionally.

Concluding remarks

The Rasch approach has turned the traditional relationship between data and analysis upside down. To consider blaming the data rather than the model when there is a mismatch between them is a considerable shift from the traditional, statistical way of thinking. Most of the criticisms of the model have originated from this new approach to the data-model relationship.

Wright and Mok (2004) state that in order to construct inferences from observation a model with certain characteristics should be used. It must:

- Produce linear measures
- Overcome missing data
- Give estimates of precision
- Have devices of detecting misfit, and
- The parameters of the object being measured and of the measurement instrument must be separable.

Only the family of Rasch measurement models does this.

Finally we quote, as does Linacre (1996), from a New York Times Editorial writing about a theory of corporate finance:

> *"That is the true test of a brilliant theory*, says a member of the Nobel Economics committee. *What first is thought to be wrong is later shown to be obvious.* People see the world as they are trained to see it, and resist contrary explanations. That's what makes innovation unwelcome and discovery almost impossible.
>
> *An important scientific innovation rarely makes its way by gradually winning over and converting its opponents. ... What does happen is that its opponents gradually die out and that the growing generation is familiarised with the (new) idea from the beginning.* No wonder that the most profound discoveries are often made by the young or the outsider, neither of whom has yet learned to ignore the obvious or live with the accepted wisdom."
>
> "Naked Orthodoxy" (October 17, 1985)

References

Anastasi, A. & Urbina, S. (1997) *Psychological Testing*, *7th ed.* (New Jersey, Prentice Hall).

Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika*, 43, 561 – 573.

Andrich, D. (2004) Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation, in: E. V. Smith, Jr and R. M. Smith (Eds) *Introduction to Rasch measurement* (Minnesota, JAM Press).

APU (1979) *Science Progress Report 1977-78.* Assessment of Performance Unit, London.

Athanasou, J. & Lamprianou, I. (2002) *A teacher's guide to assessment* (Sydney, Social Science Press).

Bode, R. K. (2004) Partial Credit Model and Pivot Anchoring, in: Smith, E. V and Smith, R. M. (Eds) *Introduction to Rasch Measurement* (pp. 279-295) (Minnesota: JAM Press).

Bond, T.G. & Fox, C. M. (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (New Jersey, Lawrence Erlbaum Associates).

Burstall, C. and Kay, B. (1978) *Assessment – the American experience*. Assessment of Performance Unit (Assessment of Performance Unit, London)

Chen, S. P. C., Bezruczko, N. & Ryan-Henry (2006) Rasch analysis of a new construct: Functional caregiving for adult children with intellectual disabilities. *Journal of Applied Measurement*, 7(2), 141 – 159.

Choppin, B. H. (1969) An Item bank Using Sample-free Calibration, in: R. Wood, and L. S. Skurnik, (Eds) *Item Banking*. (Slough, NFER)

Choppin, B. (1981) "Is Education Getting Better?" BERA Presidential Address, University College, Cardiff, 1980, *British Educational Research Journal* 7(1), 1981

Dickson, H. G. & Kohler, F. (1996) The multi-dimensionality of the FIM motor items precludes an interval scaling using Rasch analysis. *Scandinavian Journal of Rehabilitation Medicine.* 26, 159 – 162.

Divgi, D. R. (1986) Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*. 23 (4), 283 – 298.

Divgi, D. R. (1989) Reply to Andrich and Henning. *Journal of Educational Measurement*. 26 (3), 295 – 299.

Fischer, G. H. & Molenaar, I. W (1995) Rasch Models. *Foundations, Recent developments and Applications.* (New York, Springer-Verlag New York Inc.).

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*. 5(2), 211 – 220.

Hambleton, R. K. (1993) Principles and selected applications of Item Response Theory, in: R.L Linn, (Ed), *Educational Measurement*. (*3rd ed.)* 13-104. (Phoenix, Oryx Press).

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991) *Fundamentals of Item Response Theory.* (California, SAGE Publications, Inc).

Keeves, J. P., & Alagumalai, S. (1999) New approaches to measurement, in: G. N. Masters and J.P. Keeves (Eds) *Advances in measurement in educational research and assessment* (Amsterdam, Pergamon).

Keeves, J. P., & Masters, G. N. (1999) Issues in educational measurement, in: G. N. Masters and J.P. Keeves (Eds) *Advances in measurement in educational research and assessment* (Amsterdam, Pergamon).

Kuhn, T (1962) *The Structure of Scientific Revolutions* University of Chicago Press

Lamprianou, I. (2006) The stability of marker characteristics across tests of the same subject and across subjects, *Journal of Applied Measurement*, 7(2), 195 – 205.

Linacre, J.M. (1995) *Bruce Choppin: Visionary* Available online at: http://www.rasch.org/rmt/rmt84e.htm (accessed 19 November 2007)

Linacre, J. M. (1996) The Rasch Model cannot be "Disproved". *Rasch Measurement Transactions*, 10(3) 512-514.

Linacre, J. M. (1998) Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2(3), 266-283.

Massof, R. W. & Fletcher, D. C. (2001) Evaluation of the NEI visual functioning questionnaire as an interval measure of visual ability in low vision, *Vision Research*, 41(3), 397 – 413.

Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Myford, C. M. & Wolfe, E. W. (2002) When raters disagree, then what: Examining a third rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings, *Journal of Applied Measurement*, 3(3), 300 – 324.

Preece, P. (1980). On rashly rejecting Rasch: a response to Goldstein. *British Educational Research Journal* **6**(209-211).

Prieto, L., Roset, M. & Badia, X. (2001) Rasch measurement in the Assessment of Growth hormone Deficiency in adult patients, *Journal of Applied Measurement*, 2(1), 48 – 64.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests.* (Reprinted in 1980 with a forward and afterward by Benjamin D. Wright) (Chicago, MESA Press).

Rasch, G (1969) personal communication quoted in: R. Wood and L. S. Skurnik, (1969) *Item Banking*. (Slough, NFER)

Smith, Jr., E. V. (2004) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals, in: E. V Smith Jr, and R. M. Smith (Eds) *Introduction to Rasch Measurement* (Minnesota, JAM Press).

Start, K.B. & Wells, B.K. (1972) *The Trend of Reading Standards,* NFER.

Thissen, D. & Wainer, H. (1982) Some standard errors in item response theory. *Psychometrika*, 47, 397 – 412.

Whitely, S. E. & Dawis, R. V. (1974) The nature of the objectivity with the Rasch model. *Journal of Educational Measurement.* 11(3), 163 – 178.

Whitely, S. E. (1977) Models, meaning and misunderstandings: some issues in applying Rasch's theory. *Journal of Educational Measurement*. 14(3), 227 – 235.

Willmott, A.S. & Fowles, D.E. (1974) *The objective interpretation of test performance the Rasch model applied*. (Windsor, NFER Publishing Co Ltd).

Wood, R. and Skurnik, LS. (1969) *Item Banking* A method for producing school-based examinations and nationally comparable grades. (NFER, Slough).

Wright, B. D. (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97 – 116.

Wright, B.D. (1983) *Fundamental measurement in social science and education*. MESA Psychometric Laboratory. Available online at http://www.rasch.org/memo33a.htm

(accessed 24 October 2006)

Wright, B. D. (1989) Rasch model for counting right answers: Raw scores as sufficient statistics. *Rasch Measurement Transactions*, 3(2), 62. Available online at http://www.rasch.org/rmt/rmt32e.htm (accessed 28 October 2006)

Wright, B. D. & Linacre, J. M. (1989) Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-860. Available online at http://www.rasch.org/memo44.htm (accessed 24 October 2006)

Wright, B. D. & Mok, M. M. C (2004) An Overview of the Family of Rasch Measurement Models. In E. V. Smith Jr. and R. M. Smith. *Introduction to Rasch Measurement* (Minnesota, JAM Press).