International Education Journal, 2007, 8(2), 249-259.
ISSN 1443-1475 © 2007 *Shannon Research Press.*
http://iej.com.au

X-ray your data with Rasch

David D Curtis
Australian Council for Educational Research *curtis@acer.edu.au*

Peter Boman
School of Education, James Cook University *peter.boman@jcu.edu.au*

By using the Rasch model, much detailed diagnostic information is available to developers of survey and assessment instruments and to the researchers who use them. We outline an approach to the analysis of data obtained from the administration of survey instruments that can enable researchers to recognise and diagnose difficulties with those instruments and then to suggest remedial actions that can improve the measurement properties of the scales included in questionnaires. We illustrate the approach using examples drawn from recent research and demonstrate how the approach can be used to generate figures that make the results of Rasch analyses accessible to non-specialists.

Rasch, partial credit model, reliability, threshold analysis, differential item function

We have used the physical science/medical analogy of the x-ray deliberately to indicate that the application of the Rasch model can reveal otherwise hidden aspects of data. We show that it is advantageous to view data at macro-, meso- and micro-levels in order to generate a complete understanding of how well an instrument is working and to identify modifications that might improve the measurement properties of the instrument. These levels respectively refer to the scale as a whole, to items within scales, and to thresholds within items. Other facets of measurement scales must also be considered and we refer to person fit and to systematic bias, for example gender bias, in instruments.

Our analyses are not restricted to the Rasch model. Compliance with the three pillars of sound data analysis, namely the structural, distributional and measurement assumptions (Rowe, 2002) implicit in any analysis, must be confirmed. The primary concern of ensuring appropriate structural properties of the data is with the design of the sample of respondents. Where samples are clustered, design effects may require multilevel analytic techniques. A review of the distributions of responses will indicate if responses are skewed excessively and if responses adequately reflect the range of views of respondents. The three levels of analysis that we describe contribute to a demonstration that the scale does indeed conform to the requirements of measurement (Michell, 1997).

Application of the Rasch model through software such as Quest (Adams & Khoo, 1999) provides estimates of person and threshold locations on the latent variable scale. The software also yields indices of item and person fit to show that the requirement of uni-dimensionality is met. Other approaches, notably confirmatory factor analysis, can also be used to provide alternative indicators of the coherence of items and of their conformity to the requirement of uni-dimensionality. Bringing multiple perspectives to bear on a data analysis problem can give greater confidence in interpretations arising from the analyses. Our main focus is on revealing the threshold structure of items, so we report summary results of confirmatory factor analyses but do not discuss them in detail.

MEASUREMENT AND THE RASCH MODEL

The data sets that we analyse come from surveys comprising polytomous items. The raw data derived from these instruments are ordinal and do not directly yield **measures** of the constructs

that the instruments are designed to assess (Harwell & Gatti, 2001; Wright, 1993). However, provided items in the scale comply with certain axioms of measurement, there is sufficient information in the ordered responses to enable item thresholds and person locations to be mapped stochastically onto a latent interval variable. Several requirements must be satisfied about individuals' responses to items in tests and survey instruments. Weiss and Yoes (1991) stated four requirements of measurement which may be paraphrased as (i) individuals respond honestly to item prompts; (ii) items are indicators of a uni-dimensional latent trait; (iii) items are locally independent; and (iv) item responses can be modelled using a monotonic function. Threats to each of these requirements are known.

In the assessment of attitudes, respondents may exhibit a range of behaviours such as acquiescence to perceived assessor expectations (Anderson, 1997). Departures from uni-dimensionality can be checked using item fit indices from Rasch software and by testing alternative structures (uni- and multidimensional ones) using confirmatory factor modelling. The confirmation of a uni-dimensional structure provides evidence of internal consistency. A departure from local independence is observed, for example, when common stems are used for several items. Linacre (1997) has shown that this can be detected by examining item covariances, but he described this is a "third order" problem. The fourth requirement articulated by Weiss and Yoes (1991) is a key subject of Michell's (1997) challenge to measurement in the social sciences. Michell asserted "constructs that are thought to be quantitative must be shown to be so empirically" (p. 355). The literal implementation of (part of) Stevens' (1951, paraphrasing Campbell) dictum that "measurement is the assignment of numerals to objects or events according to rules" has been adopted in much psychological research without establishing the required correspondence between the numbers assigned to responses and the trait that is being measured (Michell, 2002). Testing this hypothesised quantitative relationship is what the Rasch method does, and poor fit at the macro-level or poor precision of item and threshold parameters at the meso- or micro-levels indicate a failure of measurement in Michell's terms.

The application of Rasch models has led to some rethinking of the construction of tests and survey instruments. One of the "new rules" of measurement proposed by Embretson (1999), following the application of item response theories rather than classical test theory, suggests "shorter tests can be more reliable than longer tests". We argue in this paper that this new rule needs to be qualified. Short tests can be reliable, provided their items and specifically their item thresholds do cover adequately the trait range of the sample of respondents. We set out to show how item thresholds can provide information that adds to item- and scale-level information and how those thresholds can be mapped to provide a visual representation that communicates item structures to non-specialists.

Rasch modelling provides information (a) about scales through, for example, item and scale reliability indices – the macro-level, (b) about individual items though location parameters and item fit indices – the meso-level, and (c) about individual item thresholds through their locations and standard errors – the micro-level. Using information at each of the three levels can assist in diagnosing the sources of scale deficiencies.

We demonstrate a systematic approach to the analysis of survey data using three analytic levels, namely the macro (scale) level, the meso (item) level and the micro (response category or threshold) level. In addition, we undertake preliminary analyses to check the data before commencing the Rasch analyses, and we report confirmatory factor modelling as a means of checking the results of the Rasch findings.[1]

[1] Bentler (1996) observed that confirmatory factor modelling was equivalent to the two-parameter IRT model. This differs from the Rasch approach which assumes consistency of discrimination across items and models a single, difficulty, parameter for each item.

METHODS AND DATA

The data sets used in this study are drawn from research conducted into students' anger (Boman, 2002; Boman *et al.*, 2006; Boman *et al.*, 2003) and from some scales included in the Longitudinal Surveys of Australian Youth (LSAY) program.[2] Scales were selected because they illustrated some of the features that were made apparent by using Rasch based approaches to analysis and not because the scales were either particularly good or deficient. Almost any data set arising from the administration of survey instruments could be used.

The analyses reported in this paper were conducted using the program Quest (Adams & Khoo, 1999). Quest provided various indicators of the adequacy of scales and items and had very flexible output. The graphs were generated by exporting the Quest output to a standard spreadsheet program. Other programs would provide comparable scale and item statistics and most of the analyses shown below could be generated from many other Rasch-based programs. Preliminary screening of data was conducted using SPSS (SPSS Inc., 2003) and the confirmatory factor modelling was undertaken using M-plus (Muthen & Muthen, 2006).

A RASCH BASED APPROACH TO SCALE ANALYSIS

Preliminary Analysis

The main purpose of the preliminary analysis was to check the distributional properties of responses to items. The scales used in this investigation had relatively few response categories, and it was feasible to examine the responses using the item analysis (*itanal*) command in Quest. Where scales have many response categories, perhaps as many as 11, it would be feasible to regard the responses as coming from a pseudo-continuous variable and to examine the variance, skewness and kurtosis statistics produced by SPSS. That is, the distributional properties of the responses to items could and should be examined. The preliminary analysis was also a phase in which data can be screened and cleaned. Generating frequency tables for item responses would reveal any data entry or coding errors. This is a necessary process in very large data sets or secondary data analysis, especially where the analyst has not been involved in data collection and entry. We do not report these results here, as the focus is on the Rasch analysis, but no problems were encountered in this phase of our analysis of the scales.

The Macro Level: The Scale

Under classical item analysis, Cronbach's alpha is used as an indicator of the internal consistency of a set of items. An extensive body of literature has criticised the use of Cronbach's alpha as an index of scale reliability (see Rowe, 2002 for a summary of this literature).

Rasch analyses produce several indicators of the adequacy of scale measurement. The responses of individuals to item prompts provide information about both the items and the persons. The difficulties of items and the abilities of persons are placed on a common measurement scale for the construct being assessed. The consistency with which individuals provide information about the difficulties of items that form the scale is reflected in the item separation index. The consistency with which individuals are placed on the scale by the items in the instrument is reflected in the person separation index.

Andrich (1982) showed that the person separation reliability is almost identical to the KR-20 (for dichotomous items, which is a special case of Cronbach's alpha which applies to polytomous responses), although under some circumstances the person separation index deviates from Cronbach alpha. Under the Rasch model, responses to items are assumed to be stochastic and there is an expectation that there will be some variation from a deterministic pattern of responses. However, if there is too much departure from complete consistency in observed responses, the placement of items and persons on the scale will be imprecise, and the two reliability indices will

[2] See the LSAY web site at http://www.acer.edu.au/research/projects/lsay/overview.html

be low. The magnitude of the measurement error limits the number of discriminable performance levels on the attribute being assessed. The calculation of the number of performance levels is shown in Wright and Masters (1982, pp. 91-92), and these and other summary data for the illustrative scales are shown in **Table 1**. The sample indices of item and person reliability, which are generated in Quest output, are shown in the table. For the three scales used as examples, the reliability of item separation is quite high. With the very large numbers of cases, the standard errors of item estimates are quite small and so the reliability of item separations are inflated compared with analyses based on smaller numbers of cases. Whether or not these reliability indices are thought to be inflated by having larger than necessary numbers of items or persons, it is desirable to examine the person and threshold errors of measurement to ensure that the instrument provides the desired precision given the purposes to which the data will be put. The issue of threshold precision is discussed below. In the analyses of some scales very low values for item separation reliability indices have been found (Curtis, 2003), even when the reliability of person separation is acceptable. Very low values of the item separation index indicate that the scale reflects a very broad construct or perhaps several conflated constructs.

The sample reliability of person separation is of particular interest. Values of this statistic vary from 0.69 to 0.81 for these scales. This statistic is related to the number of performance levels or bands that can be distinguished in the respondent sample using the set of items that comprise the scales. It is important to know the number of discriminable performance bands when interpreting individual achievement scores. Considerable attention is paid to this issue in large-scale testing programs, where achievement bands are associated with performance descriptors. The PISA reading and mathematical literacy scales have well described bands that can be used in describing the achievements of population groups and sub-groups and individuals (Thomson *et al.*, 2004, pp. 42-44, 92-95). Although the number of discriminable performance bands can be estimated from the separation indices, the cut points between these bands are not identified from these statistics and must be based on substantive performance criteria defined when the underlying construct is described and the items generated.

Table 1: Scale summary statistics for three example scales

Scales	Anger Intensity	Life Satisfaction	Vocational Interest (Realistic)
Number of items (L)	13	12	4
Number of respondents (N)	1400	8660	9378
Response options	4	4	4
Cronbach alpha	0.80	0.85	0.80
Item separation index	0.98	0.95	0.99
Person separation index	0.80	0.81	0.69
Number of performance levels	2.4	3.9	2.5
Mean person score	0.45	2.68	0.32

The targeting of an instrument for a sample of respondents is indicated by the distance between the mean score for items and the mean score for persons. Since the item mean is set to zero by default, the mean person score indicates the degree of mistargeting of the instrument. In a series of simulations of various factors that influence the precision of measurement, it was shown that scales for which the person mean is within 0.5 logits of the origin provide good measurement (Curtis, 2003). When the person mean was more than 1.0 logits from the origin, measurement was compromised (Curtis & Boman, 2004). In the scales analysed in this study, the Anger Intensity and Vocational Interests scales were well targeted (person means of 0.45 and 0.32) but the Life Satisfaction scale was very poorly targeted, having a person mean score of 2.68 logits. It is common to find that the around 90 per cent of case scores for well-targeted instruments lie in a range from -3 to +3 logits. For the Life Satisfaction scale, the mean score is close to the effective upper limit of this common distribution pattern. It suggests that the set of response options requires revision and that the lowest two response categories might be combined and an additional upper option be generated. This matter is addressed further in examining the micro-level features (thresholds) of items.

The Meso Level: Items

The meso- or item-level of analysis tends to be the one that receives most attention through an examination of fit statistics. Many analysts appear to pay considerable attention to item fit statistics, but few attend to person fit statistics. (See the review in Curtis & Boman, 2004).

The issue of item fit has been canvassed thoroughly in many texts and articles (see, for example, Bond & Fox, 2001; Linacre, 1995; Linacre *et al.*, 1994; Wright & Masters, 1982) and it will be covered only briefly in this paper.

Item Fit

Many indices of item misfit have been tested (Li & Olejnik, 1997; Linacre, 1998; R. M. Smith, 1996), but most common Rasch analysis software programs use the information weighted index (Infit Mean Square or IMS) and the unweighted index (the Outfit Mean Square or OMS). There has been some debate in the literature on whether to use the statistics themselves or their t-transformed variants. In most statistical procedures, a *t* (or similar) statistic is computed and its corresponding probability of being observed under normal sampling variation is used to decide whether the observation is likely (or not) to have arisen by chance. It is tempting, therefore, to apply the same logic to evaluating the fit statistics produced for items (and persons) in Rasch analysis. The consensus appears to favour the use of the fit statistics rather than their *t*-transformations (Bond & Fox, 2001; Linacre et al., 1994). Bond and Fox (2001, p. 179) provide advice on acceptable ranges of item fit using IMS values, depending upon the purpose of the measurement exercise. There is some value in using both IMS and OMS values. The OMS statistics are more sensitive to outliers, and if an item shows acceptable fit on one index, but marginal or poor fit on the other, the item should be investigated more closely. This might include a search for outlying cases, perhaps using person fit statistics, or could involve a detailed analysis of thresholds (micro-level analysis, see below) or differential item function. A decision to accept, modify (and if so, in what way) or to reject items should be made after an examination of all relevant evidence. Item fit statistics provide part of this evidence, but additional evidence may be found from an examination of person fit statistics and inspection of the micro-level (threshold) structure of items.

In the three scales used as examples in this paper, item fit statistics lie within the acceptable range from 0.7 to 1.4. Item fit statistics for the Life Satisfaction scale are shown in **Table 2**. The item asking about money earned fits the scale much less well than the others. Using the fit criteria suggested by Bond and Fox (2001, p. 179) this item could be retained. However, it could be argued that wages reflect a unique dimension of satisfaction with an individual's life situation, but a decision to remove or retain this item should be based on the substantive intent and meaning of the scale and not on the arbitrary application of a particular criterion value for a fit index (Linacre et al., 1994). 'Your life as a whole' summarises the views individuals had expressed through other items in the scale. It can be expected to be highly correlated with other items and to add little information that is not already conveyed by responses to other items. This redundancy is reflected in the low IMS value (0.75) for this item.

It is instructive to examine the *t* transformed fit statistics for these items. The first item (satisfaction with work) has an IMS value of 1.06, a figure that would be acceptable for even the most rigorous assessment purposes. However, its *t* statistic is 3.05, suggesting significant misfit. The LSAY Y03 2004 sample included 8690 cases and this large sample leads to low standard errors of estimates and to high *t* statistics for a given level of misfit.

Person Fit

Fit statistics are determined by the response vector of persons to a set of items or of items for a sample of persons. However, each response is stochastic in that a person with a certain trait may choose a particular response option on one occasion and alternative option on another occasion.

When there are few items with few response categories, selecting a different response category can lead to a substantial change in the person fit statistic. Thus, the power of fit statistics to detect aberrant response patterns depends on the number of items and the number of effective response choices for those items (Curtis & Boman, 2004). Typically, there are many more respondents to a survey instrument than there are items within it, so much more latitude has to be allowed in deciding whether to accept or reject a person as fitting. For an attitude scale, an upper limit on the IMS range for an item might be 1.4 (Bond & Fox, 2001, p. 179), but for a person this figure might be set at around 1.6. For scales with few items and few response choices, even greater latitude must be permitted. In simulations in which some responses that were entirely random, IMS values in excess of 4 were found. If respondents with these high values are found, it may be desirable to remove these cases for instrument calibration because their inclusion would lead to increased estimated standard errors of item estimates. A decision on whether to re-admit these individuals to the sample for other purposes, such as scoring, might be made separately. This decision would depend on what interpretation can be made of their scores. This situation is illustrated in **Table 3**, which shows parameters for selected cases on the Realistic Vocational Interest scale. Note that two individuals had the same trait estimate (1.24), but one had an IMS value of 1.02 indicating good fit to the model while the other has a misfit of 2.99, suggesting very poor fit. A question for the analyst is 'What substantive meaning can be imputed to these two cases?' In this instance, it seems that one individual has expressed preferences that accord with the order of preferences of most other respondents, but the other may have a unique pattern of interests or may simply have responded thoughtlessly or randomly to the items.

Table 2: Item parameters for the Life Satisfaction scale from the LSAY Y03 2004 questionnaire

Items	IMS	OMS	Infit t	Outfit t
Firstly, how happy are you with…				
The work you do, at school, at home or in a job	1.06	1.08	3.05	3.29
What you do in your spare time	0.99	0.97	-0.37	-1.32
How you get on with people in general	0.96	0.94	-2.35	-2.36
The money you get each week	1.39	1.46	21.74	20.27
Your social life	0.99	0.98	-0.70	-0.73
Your independence - being able to do what you want	1.09	1.11	5.14	4.78
Your career prospects	1.06	1.06	3.46	2.90
Your future	0.92	0.88	-4.95	-5.57
Your life at home	0.86	0.81	-8.44	-8.63
Your standard of living	0.84	0.80	-10.29	-8.39
Where you live	0.98	0.97	-1.48	-1.38
Your life as a whole	0.75	0.67	-18.02	-14.82

Notes: IMS = Infit Mean Square; OMS = Outfit Mean Square.

Precision of Person Estimates

In educational psychology, there has been considerable focus on scale properties – for example reliability as indicated by Cronbach alpha – with much less attention being paid to the precision of the estimated trait levels of individuals (Adams, 2005, p. 164 citing Weiss and Davison (1981)). In classical test theory, it is assumed that measurement error is constant across the trait range. In Rasch analysis, this is not so, and nor should it be expected. In the range where the instrument is targeted and where most thresholds are located, the precision of measurement will be greatest, but at the margins of the measured range where there are few fixed points on the measurement scale (threshold locations), estimates of individuals will be less precise. The precision of individual estimates is illustrated (**Table 3**) by some selected cases, reflecting the range of ability estimates on the Realistic Vocational Interests scale. Estimates near zero logits have the lowest standard errors (at 0.6 logits) while those at the margins of the scale have standard errors of more than one logit. Two questions arise. First, is the precision of the measurement in the most effective region of the scale adequate for the purposes to which they are to be put? Second, at what point do the errors become too great for the estimates to be useful for their intended purposes?

Table 3: Person estimates, standard errors and Infit Mean Square for selected cases from the Realistic Vocational Interest scale

ID	Raw score	Estimate	se	IMS
4366	11	2.75	1.10	1.34
1543	11	2.75	1.10	0.49
3453	10	1.85	0.84	0.57
5356	9	1.24	0.73	2.99
833	9	1.24	0.73	1.02
9918	8	0.75	0.67	0.15
1324	7	0.33	0.63	0.81
8977	7	0.33	0.63	0.05
9109	6	-0.06	0.62	5.73
4753	5	-0.44	0.62	1.77
6162	5	-0.44	0.62	0.59
10346	4	-0.83	0.64	1.37
2692	3	-1.27	0.69	0.57
6086	2	-1.81	0.79	0.30
6914	1	-2.62	1.05	0.93

The Micro Level: Thresholds

In this paper, delta, or Masters, thresholds are used. A delta threshold is the point on a scale at which a respondent has an equal probability of endorsing either of two adjacent response categories. The delta threshold estimates are reported relative to the scale origin. Thurstone thresholds are reported by default in some software packages. They are the points at which a person has an equal probability of selecting a response category or any of those above that level. Thurstone thresholds are necessarily ordered, whereas the delta thresholds may not be. Disordered thresholds may indicate a failure of correspondence between the latent trait and the assignment of scores to supposedly ordered response categories, but it is more likely that threshold reversals arise because of low response frequencies to one or more options. If reversed thresholds are detected, the cause should be investigated.

In this paper, the partial credit model, in which the steps or distances between thresholds are permitted to vary across items, has been used. By comparison, in the rating scale model (Andrich, 1997), steps between particular thresholds pairs are held constant across items. The penalty of the partial credit, compared with the rating scale, model is that additional parameters must be estimated, but the advantage is that if the thresholds of a particular item are either not discrete or, worse, disordered, the partial credit model will reveal that problem. Conquest (Wu *et al.*, 1998) provides a method of comparing the relative fit of these two models. The rating scale and partial credit models can be tested for a particular data set, and a deviance statistic is calculated for each. Invariably, the deviance statistic will be greater for the rating scale model, but if the difference between the two is not significant, the rating scale model can be used without loss of relevant information. However, in order to investigate the threshold structure of items, it is necessary to use the partial credit model and to allow the threshold steps to vary across items.

Thresholds for the Anger Intensity scale from the Multidimensional School Anger Inventory (D. C. Smith *et al.*, 1998) are shown in **Figure 1**. The mean score of persons on this scale was 0.45 and the standard deviation 0.86 logits. The target measurement range, that is the region of the scale where most respondents lie, is approximately from -1.0 to +2.0 logits, and it can be seen from **Figure 1** that this range is evenly and well populated with thresholds. Thus, the measurement of individuals in this range is based on many calibrated points on the scale. There are, however, thresholds outside this range. Most of the first series of thresholds (Delta 1 in **Figure 1**), those that separate the lowest two response categories, lie outside this range. The confidence intervals for these thresholds are rather large, indicating that these thresholds are not estimated precisely. This is to be expected, since there are few individuals in this region of the scale and so few responses endorsing the lowest two categories from which the locations of these

thresholds can be estimated. A similar situation was observed for the Life Satisfaction scale, with imprecisely estimated lower thresholds and poor separation of many of these thresholds. The problem with the Life Satisfaction scale can be attributed to its poor targeting, as indicated by a mean person score of 2.68 when the item mean is set at the origin for the scale.

For the Anger Intensity scale, of greater importance than the low precision of the lower threshold estimates, is the lack of well-separated thresholds for items 5, 9 and 10, circled in **Figure 1**. Items 5 and 10 have quite distinct lower thresholds, but the upper two that separate the "I'd be a little angry" from "I'd be quite angry" and "I'd be really angry" indicate that these response options are not working well for these two items. Item 9 reveals an even greater problem, in that all three thresholds are poorly separated. Item 9 has a slightly high but acceptable IMS value of 1.22. In effect, its three thresholds could be collapsed into a single one with little loss of information, and it appears that the item is working as a dichotomous one. The cause of this was diagnosed, at least in part, to a typographical error when the instrument was adapted for use in Australia and the word "special" was omitted from the item, which should have read "The teacher's pet gets to do all of the special errands in class" (Boman et al., 2006). Changing one word in an item can have a quite marked impact. This suggests, encouragingly, that apparently minor revisions can improve problematic items, but it should serve as a warning that making apparently cosmetic changes to items can have striking influences on their measurement properties.

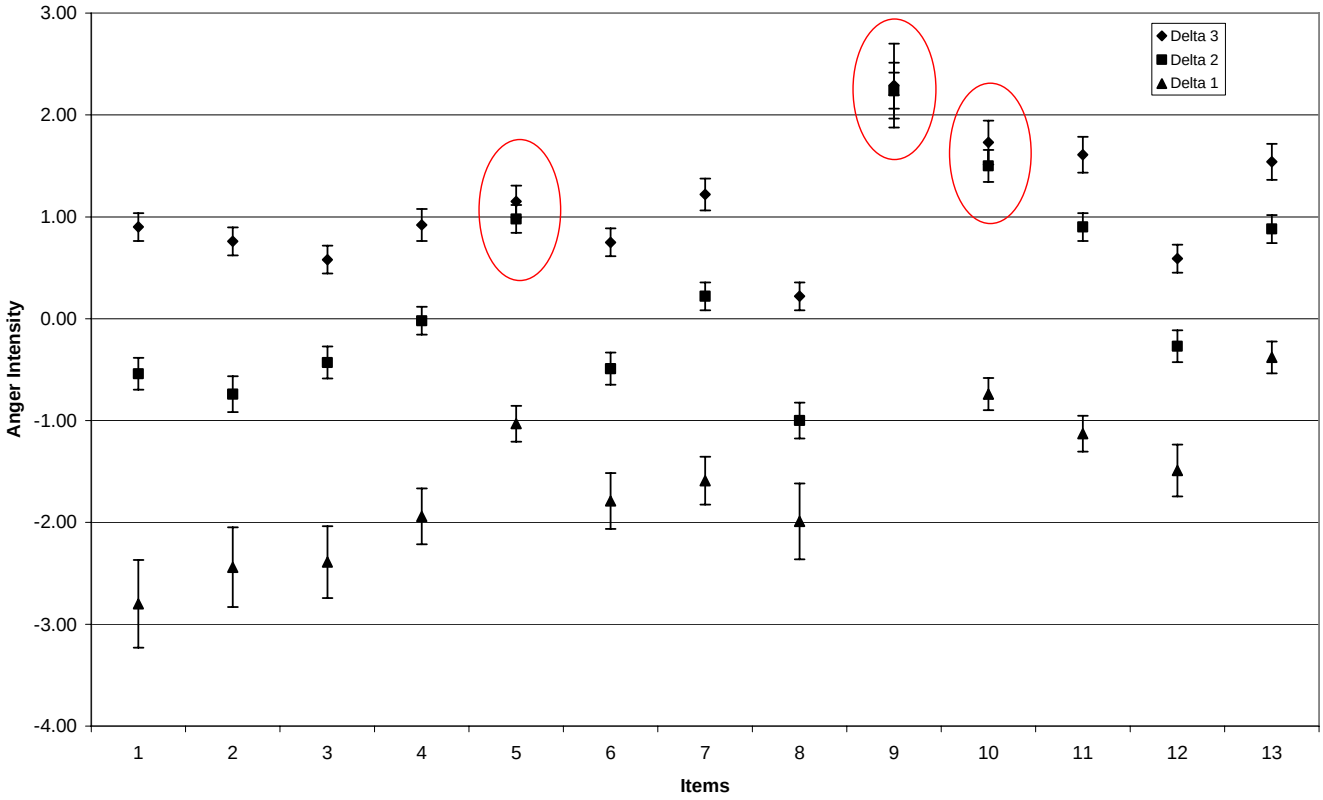

Figure 1: Thresholds for the items of the Anger Intensity scale

Differential Item Function

The Realistic Vocational Interest scale was well targeted and had ordered and well separated item thresholds. The person estimates were not very precise (standard errors > 0.6 logits) and appeared to be a good scale. There was concern, however, arising from relevant theory. The items had been selected to reflect the Holland (1985) theory of vocational preference. Gottfredson (2002) has shown that vocational choices are made through a process of comparing person attributes with perceived characteristics of occupations. One of the characteristics that is perceived early, and

accurately, in career decision-making is the gender-typing of jobs. Many of the realistic[3] vocational activities used in assessing vocational interests are strongly gender-typed, so differences between the ordering of preferences were sought by gender by undertaking a differential item analysis using Quest's *Compare* command. The thresholds of items for males and females are shown separately for the four items of the Realistic Vocational Interest scale in **Figure 2**.

Two items – 'drive trucks' and 'use tools' – show very substantial differences in the thresholds for males and females. These thresholds are highlighted in **Figure 2**.[4] Only males whose realistic vocational interest is much stronger than is the case for females endorse the 'like it a lot' option for this item. There are consistent and substantial differences in the level of realistic vocational interest between males and females on the 'use tools' item. These differences suggest that the structure of vocational interests differs between males and females. It may be necessary to calibrate these scales separately for males and females or even to consider developing different items that refer to different occupational activities for the two genders.

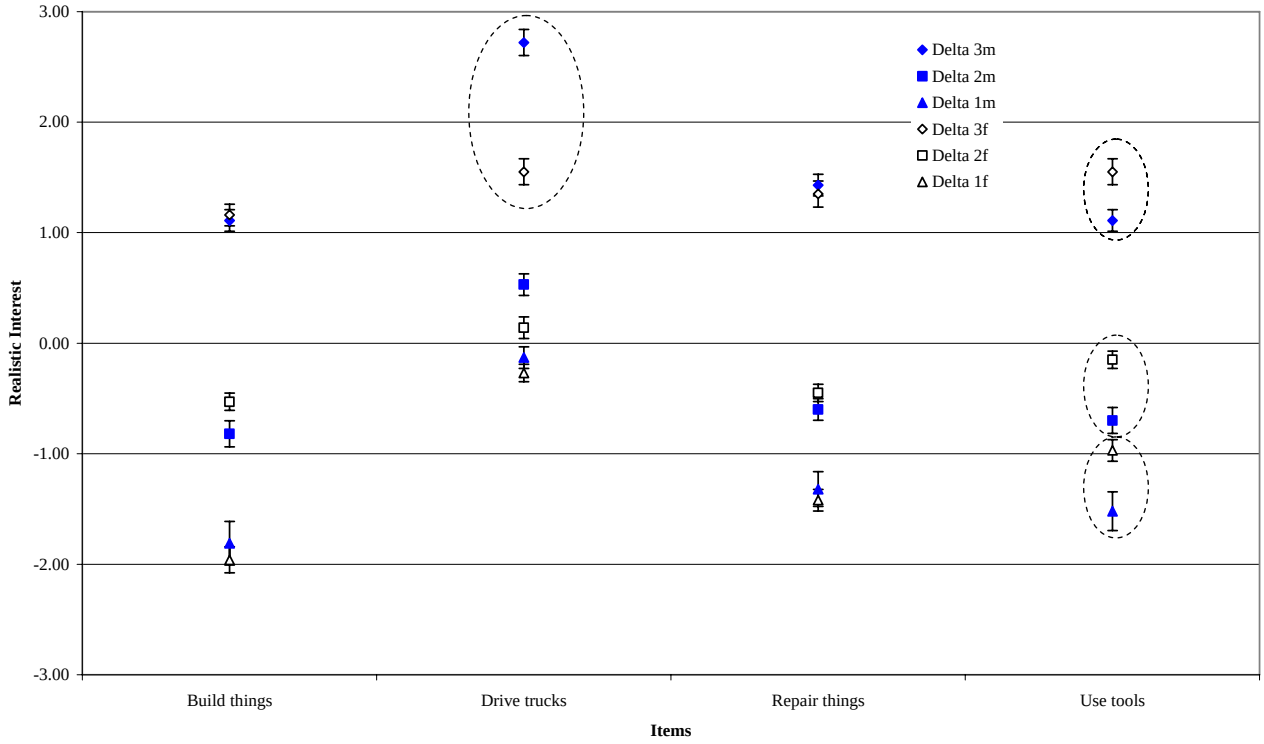

Figure 2: Thresholds for males and females on the Realistic Vocational Interests scale

SUMMARY

In the paper, we have drawn attention to three levels of analysis, namely the macro- or scale-level, the meso- or item- and person-level, and the micro- or threshold-level. At each level, Rasch analysis software produces several indicators that should be considered by analysts. At the macro-level, indices for the reliability of both item and person estimates should be examined. In addition, the degree to which the instrument is targeted for the sample should be inspected. Poorly targeted instruments will not lead to precise estimates of individuals on the trait.

At the meso-level, item fit statistics are routinely examined. The case has been made that person fit statistics and, in particular, the standard errors of person estimates must be examined to ascertain that the measures are of sufficient precision for the purposes to which the data are put.

---

[3] According to Holland's (1985) typology, realistic interests are those most closely associated with traditional trade occupations, and males dominate employment in many of these occupations.

[4] Alternative plots of male and female thresholds on x- and y-axes also show the effects of differential item function.

Micro-level analyses are especially informative. Having the desired measurement range well populated with precisely quantified thresholds adds to the information available for estimating person locations. The most parsimonious scales will have the full measurement range populated with thresholds but without redundant ones. It is in this respect that the "new rule" of measurement proposed by Embretson (1999) that "shorter tests can be more reliable than longer ones" may apply. Shorter tests may be better than longer ones if the thresholds are estimated with precision and if they cover adequately the desired measurement range. Thresholds within items should be well separated and certainly should not show reversals. Where these thresholds are not distinct, problems within items should be suspected. Other problems, such as differential item function, can be investigated through an examination of threshold locations for different sub-groups of the sample.

In addition to ensuring that structural and distributional properties of the data set, examination of measurement data at three levels – macro, meso and micro – inform analysts about the quality of their measurement instruments and about the quality of the measures generated through the application of their instruments. Information from each level of analysis can assist in the interpretation of information at other levels.

REFERENCES

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

Adams, R. J., & Khoo, S. T. (1999). Quest: the interactive test analysis system (Version for PISA) [Rasch analysis software]. Melbourne: Australian Council for Educational Research.

Anderson, L. W. (1997). Attitudes, measurement of. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (pp. 885-895). Oxford: Pergamon.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, 9*(1), 95-104.

Andrich, D. (1997). Rating scale analysis. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (pp. 874-880). Oxford: Pergamon.

Bentler, P. M., & Houck, E. L. (1996). Structural equation models of multiple sclerosis disease status. *The International Test Commission Newsletter, 6*(1), 11-13.

Boman, P. (2002). *Optimism, pessimism, anger, and adjustment in adolescents.* Unpublished PhD Thesis, University of South Australia, Adelaide.

Boman, P., Curtis, D. D., Furlong, M. J., & Smith, D. C. (2006). Cross-validation and Rasch analyses of the Australian version of the Multidimensional School Anger Inventory - revised. *Journal of Psychoeducational Assessment, 24*(3), 225-242.

Boman, P., Smith, D. C., & Curtis, D. D. (2003). Effects of pessimism and explanatory style on development of anger in children. *School Psychology International, 24*(1), 80-94.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Curtis, D. D. (2003). *The influence of person misfit on measurement in attitude surveys.* Unpublished EdD dissertation, Flinders University, Adelaide.

Curtis, D. D., & Boman, P. (2004). *The identification of misfitting response patterns to, and their influences on the calibration of, attitude survey instruments.* Paper presented at the 12th International Objective Measurement Workshop, Cairns, QLD.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement. What every psychologist and educator should know* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum and Associates.

Gottfredson, L. S. (2002). Gottfredson's theory of circumscription, compromise, and self-creation. In D. L. Brown (Ed.), *Career choice and development* (4th ed., pp. 85-148). San Francisco: Jossey-Bass.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Li, M. N. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215-231.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*(2), 422-423.

Linacre, J. M. (1997). Investigating Judge Local Independence. *Rasch Measurement Transactions, 11*(1), 546-547.

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

Linacre, J. M., Wright, B. D., Gustafsson, J.-E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(2), 370.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54*(2), 99-104.

Muthen, L. K., & Muthen, B. O. (2006). MPlus (Version 4.0) [Statistical analysis with latent variables]. Los Angeles, CA: Muthen & Muthen.

Rowe, K. (2002). The measurement of latent and composite variables from multiple items or indicators: Applications in performance indicator systems. Retrieved 7 March, 2006, from http://www.acer.edu.au/research/programs/documents/MeasurmentofCompositeVariables.pdf

Smith, D. C., Furlong, M. J., Bates, M., & Laughlin, J. D. (1998). Development of the Multidimensional School Anger Inventory for males. *Psychology in the Schools, 35*(1), 1-15.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

SPSS Inc. (2003). SPSS for Windows (Version 12.0.1) [Statistical analysis program]. Chicago: SPSS Inc.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: John Wiley.

Thomson, S., Cresswell, J., & de Bortoli, L. (2004). *Facing the future: A focus on mathematical literacy among Australian 15-year-old students in PISA 2003*. Melbourne: ACER and OECD.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: theory and applications* (pp. 69-95). Boston: Kluwer Academic Publishers.

Wright, B. D. (1993). Thinking with raw scores. *Rasch Measurement Transactions, 7*(2), 299-300.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest generalised item response modelling software (Version 1.0) [Statistical analysis software]. Melbourne: Australian Council for Educational Research.