Misfits: People and their problems. What might it all mean?

David D Curtis
Flinders University, Centre for Lifelong Learning and Development and School of Education
david.curtis@flinders.edu.au

In the analysis of data, which arise from the administration of multiple choice tests or survey instruments and which are assumed to conform to a measurement model such as Rasch, it is normal practice to check item fit statistics in order to ensure that the items used in the instrument cohere to form a unidimensional trait measure. However, checking whether individuals also fit the measurement model appears to be less common. It is shown that poor person-fit compromises item parameter estimates and so it is argued that person-fit should be checked routinely in the calibration of instruments and in scoring individuals. Unfortunately, the meanings that can be ascribed to person-fit statistics for attitude instruments is not clear. A proposal for seeking the required clarity is developed.

Item Response Theory, Rasch, person-fit statistics, attitude

Three sets of data derived from the application of different attitude survey instruments have been analysed using Item Response Theory (IRT) based software packages Quest (Adams & Khoo, 1997) and Conquest (Wu, Adams, & Wilson, 1998). In all cases, items that fit a unidimensional scale have been found. Kline (1993) however, has noted that the Rasch (one parameter) IRT model may be insensitive to departures from the assumption of unidimensionality. However, closer examination of the data using Confirmatory Factor Analysis (CFA) has found that the structure, supported in IRT analysis, is regarded as only moderately fitting and that further refinement was required. A fundamental requirement of measurement is that the set of items represent a single, unidimensional construct (Michell, 1997; Weiss & Yoes, 1991; Wright & Masters, 1982).

In a review of these data sets, a significant number of poorly fitting cases were found. These cases may have compromised the calibration of the instrument and may have introduced factors other than the attitude construct, whose measurement was the goal of the instruments. The IRT software appears to have been less sensitive to sources of variation in the data than the CFA software. In IRT based measurement, the claim is made that item parameters are independent of the particular sample of respondents and that respondent scores are independent of the particular sample of items used to measure the underlying trait. Thus both items and persons are samples of all possible items and of the universe of persons represented by the sample. But are misfitting persons somehow not legitimate members of a coherent population? What criteria might be used to disqualify them?

In seeking answers to these questions, the person-fit statistics were reviewed. It seems possible that misfitting persons may introduce a source of variance that is detected through CFA but that is not detected using IRT software. If this is the case, it is possible that item calibrations are suspect and even that acceptable items have been rejected in the IRT refinement of the instrument. In addition, attitude survey instruments are the central concern of this research, and criteria for judging the appropriateness of response patterns are rather different from criteria that might be used in knowledge testing. In objective tests of knowledge, cheating and

guessing are threats to valid response patterns. However, in attitude surveys these are not viable threats, although others, such as a desire to appear to have favourable views, may be.

In order to investigate the effect of poorly fitting cases on item parameter estimates, data from the 1996 application of the Course Experience Questionnaire (CEQ) were reanalysed.

A REVIEW OF PERSON-FIT STATISTICS

Fit statistics for both items and persons are derived from deviations of observed responses from expected ones. The matrix of individuals' responses to items is tabulated and the rows and columns summed. Under the Rasch model, the person and item totals are used to estimate person ability (β) and item difficulty (δ). In turn, these values are used to estimate the expected response of each person to each item. In general, an observed response should be associated with a high probability of that response, and the deviation of the observed response from the expected one is an indicator of misfit.. The simplest misfit statistic is the unweighted mean square residual, found by taking the mean value of the squared differences between the observed and expected responses, and is the Outfit Mean Square reported in Quest.

Karabatsos (2000) is critical of fit statistics such as this, arguing that the expected value is a continuous quantity while the observed response is always a discrete one – most often a '1' or a '0'. This means that, even for very well fitting items and persons, there will always be a residual, albeit a small one. This may not be important, as fit is a matter of relative judgement, and for poorly fitting items or persons, residuals will be much higher than for those with reasonable fit. Karabatsos has proposed an alternative approach to item and person misfit, that itself is based on a logisitc model, using the measurement scale of the instrument.

It appears to be common practice in the analysis of data sets under IRT, collected for purposes of validating instruments, to focus on item parameters and to pay little attention to case fit statistics (eg Waugh, 1999). It might be added that other methods of analysis do not attend to the issue of person-fit at all and that the sampling process is assumed to generate a genuinely representative set of individuals. In some of the original treatments of the Rasch Measurement Model (eg Wright & Masters, 1982), almost equal treatment is accorded to both item and person-fit statistics. Indeed, there appears to be a sound case for such equity. Inspection of the matrix of any data set being analysed using the Rasch model reveals that analogous operations are performed on the rows and columns to derive marginal totals and expected cell values. Whether the primary view is of items or of people, the mathematics of the analysis for items and persons are mirror images, and if the matrix is to be fitted to a measurement model, then both the items and the persons should be treated similarly.

A range of person-fit statistics has been used to assess conformity of the data set to the measurement model (Li & Olejnik, 1997). They tested five indices of misfit, including those that are employed in the Quest and Conquest programs – the Weighted Mean Square residuals (WMS or Infit Mean Square) and the Unweighted Mean Square residuals (UMS or Outfit Mean Square) developed by Wright and Stone (1979). Although Li and Olejnik reported that there was little difference among the performances of the indices that were tested, the performance of the two that are used in Quest was slightly inferior in both the detection of misfit and in reporting false positives, that is reporting as misfits cases that were reasonable fits. They further reported that these indices do not strictly follow a normal distribution, and recommend that their standardised versions, Infit t and Outfit t, should be given somewhat more latitude than the common ± 2.

In the current study, there are many cases (51,356) and relatively few items – 17 after refinement of the instrument. This means that item estimates are based on many cases, and very small misfits have very high Infit t values. Case estimates, based on relatively few items, can show substantial misfit but still have modest Infit t values. This raises the question: Should fit be judged on the Infit MS or the Infit t value? In addition, in large samples like the one

being reported here, there is a degree of clustering within course types and within institutions, and any analysis that treats all cases as being drawn from a single population will reveal aggregation bias and consequently false significance.

The study reported by Li and Olejnik (1997) simulated objective test data. In such tests, there are a range of threats to measurement model conformity. Most commonly, carelessness, cheating, lucky guessing, special knowledge, and miscoding of the data are proposed as possible reasons for atypical response patterns (Bond & Fox, 2001, p.178). These conditions normally lead to high (under-fitting) person-fit statistics. They also indicate that over-fitting can arise when the responses follow a deterministic response pattern whose fit statistics are "too low to be believed" (Bond & Fox, 2001, p.178). The reasons advanced to explain poor person-fit are relevant to objective testing, but not necessarily to attitude survey instruments. In these cases cheating or guessing seem irrelevant, but carelessness or even deliberate spoiling of the instrument and miscoding may be problems.

In the current study, attitude data derived from a survey have been re-analysed in order to examine possible sources of misfit and to ascertain the effect of person misfit on item parameter estimation.

METHOD

Data from the 1996 round of the Course Experience Questionnaire (CEQ) was analysed previously under IRT using Quest (Adams & Khoo, 1997). The CEQ is a 25 item instrument designed to gather the views of graduates on the quality of their recently completed courses. Respondents selected one of five options – Strongly Disagree, Disagree, Neutral, Agree, or Strongly Agree to each item. The CEQ was based on a theoretically informed view of the major contributing factors to quality of teaching and to perceived course quality and comprised five sub-scales. The refinement of the instrument revealed that only 17 of the items fitted both their intended sub-scale and an overall coherent scale of perceived course quality. Subsequent CFA revealed that a nested structure of one underlying course quality construct and the five proposed sub-scale factors was a reasonable model (Curtis, 1999, p.18).

In the current study, the 17 fitting items were taken as a starting point and Rasch scaled scores, standard errors, and fit statistics were generated for the 51361 cases for whom complete data were available. Of them, 65 (0.13 per cent) had either zero or perfect scores, 7093 (13.74 per cent) had low Infit MS (<-2), and 5384 (10.43 per cent) had high Infit MS values (>2).

In order to build a case for the exclusion of persons who demonstrated misfitting response patterns, the responses of the most poorly fitting cases, both those with very low and very high Infit MS values, were examined. The cases with zero or perfect scores are not included in item or case estimation as the algorithm cannot estimate scaled case scores nor produce fit statistics for them. These are people who have responded consistently to items in that they have either very positive sentiments (agreed strongly with positive items and disagreed strongly with negative items) or they have very negative views (having disagreed strongly with all positive items and agreed strongly with all negative items). There are methods for extrapolating from scaled cases and imputing scaled scores for these individuals, although no confidence interval can be associated with these imputed values. These cases are not of interest in the current study as they are excluded from Rasch analysis and do not influence item or person parameter estimation.

Inspection of the response patterns of 20 cases showing the highest misfits revealed that respondents had selected either 'Strongly Disagree' or 'Strongly Agree' to all items, even though eight of the items were reversed. Thus their responses were inconsistent with a coherent expression of the trait targeted in the instrument. The pattern of these cases is that they have checked all responses down either the left or the hand side of the response column, irrespective of the sense (positive or negative) of the items. Both the item fit statistics and the

observed pattern of responses indicate that there is a reasonable case for removing the responses of these people.

The response patterns of the cases with the lowest person-fit statistics (with Infit MS values around 0.10) were less simple. These cases are characterised by the choice of middle range response options, either 'Agree' or 'Neutral', but more importantly by an invariant response pattern. Typically these respondents selected 'Agree' to all items but one, and selected an adjacent category, say 'Neutral' to the remaining item. Item locations vary and a person with a true level on the underlying trait of perceived course quality that falls close to the range represented by this response category ('Agree') might be expected to select an adjacent category for some of the items. However, selecting the 'Agree' response option is not an unlikely event for a person who has a moderate overall trait score. Therefore the case for removing these persons from the analysis is far from convincing. Thus, for persons with low Infit MS values, the person-fit statistics might suggest removal of the cases, at least for calibration purposes. However, the pattern of responses cannot be interpreted as easily as in the situation for under-fitting cases described above. On one hand, it could be argued that the invariant selection of a single response option suggests a 'patterned and thoughtless response' but the responses are not improbable given the underlying trait value. This analysis suggests that other criteria require examination before a decision is made to retain or to exclude over-fitting cases.

In order to ascertain the influence of misfitting cases on item parameter estimates, the original data set with all 51,631 cases was used to generate item parameters. Subsequently, under-fitting cases (Infit $t > 2$) were removed and item parameters re-estimated. Finally, all misfitting cases (Infit $t < -2$ and Infit $t > 2$) were removed and item parameters again re-estimated. In order to establish further the influence of person-fit on overall model fit, confirmatory factor analyses were undertaken. A refinement process using the Quest software was also conducted beginning with all 25 items and successively removing misfitting items with both the under-fitting cases removed and then with all misfitting cases excluded. This was done to examine whether misfitting cases would bias the fit statistics of items and therefore the items removed during the refinement process.

RESULTS

The Influence of Person-fit on Item Locations

While for some items, for example Item 1, the effect of removing misfitting cases has little impact, for the majority the effect is significant. This is shown in ***Table 1***. Here, Item 1 shows a small change, but Item 5 shows a shift of 0.13 logits when under-fitting cases are removed from the calibration. This is taken to be a significant effect given that the standard errors of the location estimates are 0.01 for all items. In summary, the removal of misfitting cases leads to a greater spread of item locations. This effect is more pronounced for under-fitting cases and suggests that misfitting cases add to error variance but not to information about items.

The Influence of Person-fit on Item Thresholds

The influence of cases with poor person-fit statistics were shown also to influence item threshold parameters. The four threshold parameters for each of the 17 items with all cases included, under-fits removed, and all misfits removed are shown in ***Table 2***. Standard errors for the threshold parameter estimates, which are not shown in the table, vary from 0.01 to 0.06, with most at 0.02 logits. The thresholds under the three conditions vary by considerably more than could be attributed to random error. In general, the inclusion of misfitting cases, especially under-fitting ones, compresses the width of the item steps on average by 0.32 logits with under-fitting cases are removed and by 0.25 logits with all misfitting cases are removed.

Of particular note are the locations of the first and fourth thresholds. The inclusion of all cases results in a compression of the thresholds compared with the situation in which the under-fitting cases are excluded. This is

illustrated most effectively by tabulating differences between the first and fourth thresholds under the three conditions and is shown in

Table 3. The removal of under-fitting cases extends the range by almost a logit while also removing over-fitting cases extends the range by approximately one half of a logit compared with the situation in which all cases are retained. This has implications for later applications of a scale. Calibrating the scale on one sample and anchoring item parameters, so that case scores may then be estimated for other samples, may lead to better person separation in subsequent samples.

Table 1: Item locations for three person-fit conditions

	Item Locations		
Item No.	All cases	Under-fits removed	All misfits removed
1	0.06	0.07	0.06
2	-0.46	-0.58	-0.55
5	-0.53	-0.66	-0.63
6	-0.03	-0.05	-0.05
7	0.64	0.77	0.73
10	-0.12	-0.19	-0.18
11	-0.37	-0.44	-0.41
12	-0.09	-0.06	-0.05
13	0.06	0.05	0.05
14	0.17	0.25	0.24
15	0.37	0.46	0.44
17	0.41	0.50	0.48
18	0.32	0.39	0.37
19	-0.42	-0.48	-0.47
20	0.17	0.21	0.19
22	-0.42	-0.52	-0.50
24	0.24	0.29	0.27

Standard Errors are not shown, but for all items, the standard error of the estimate is 0.01 logits.

Confirmatory Factor Analyses

In the original analyses of the 1996 CEQ data, confirmatory factor analyses were undertaken in order to identify the most suitable model for representing the structure of the instrument. These analyses suggested that a nested model was the most appropriate (Curtis, 1999). In the present study, the nested model was taken as established, and the model was re-run under LISREL 8.12a (Joreskog & Sorbom, 1993) in order to examine the impact of the removal of misfitting cases on model fit. ***Table 4*** shows that the removal of misfitting cases, and most particularly the removal of under-fitting cases, results in a slight improvement in model fit. However, the parsimony fit index is rather low in all three conditions, suggesting that the model might be further simplified.

RASCH REFINEMENT OF ITEMS

In order to examine whether the removal of items in the original refinement was a result of the inclusion of misfitting cases, the refinement process was conducted after excluding under-fitting cases and again following the removal of all misfitting cases. With both reduced data sets, all 25 items were included and misfitting items were removed in an iterative process. For both reduced data sets, the same eight items that had been cut in the original refinement were

again removed. This result suggests that, despite having a sample of which 25 per cent of cases showed poor fit, item fit statistics are not strongly influenced by the presence of these cases and the refinement process is a reasonably robust one.

Table 2: Item thresholds for three person-fit conditions

	Tau 1			Tau 2			Tau 3			Tau 4		
Item No.	All cases	No under-fits	No misfits	All cases	No under-fits	No misfits	All cases	No under-fits	No misfits	All cases	No under-fits	No misfits
1	-1.85	-2.35	-2.19	-0.78	-0.86	-0.66	0.35	0.45	0.34	2.28	2.76	2.51
2	-1.46	-2.00	-1.87	-0.65	-0.68	-0.50	0.11	0.26	0.22	2.00	2.43	2.15
5	-1.40	-1.95	-1.81	-0.77	-0.82	-0.64	0.15	0.31	0.27	2.02	2.46	2.19
6	-1.58	-2.07	-1.90	-0.62	-0.67	-0.49	0.11	0.21	0.11	2.09	2.54	2.28
7	-1.60	-1.97	-1.73	-0.64	-0.74	-0.61	0.48	0.54	0.38	1.76	2.17	1.97
10	-1.53	-2.10	-1.94	-0.91	-0.95	-0.77	0.28	0.41	0.34	2.16	2.64	2.37
11	-1.05	-1.47	-1.33	-0.46	-0.50	-0.32	-0.01	0.09	0.05	1.52	1.88	1.60
12	-1.09	-1.43	-1.26	-0.63	-0.69	-0.51	0.31	0.37	0.27	1.40	1.76	1.50
13	-1.73	-2.26	-2.09	-0.72	-0.78	-0.59	0.37	0.48	0.38	2.08	2.56	2.30
14	-1.92	-2.43	-2.23	-0.82	-0.94	-0.75	0.24	0.30	0.16	2.51	3.07	2.82
15	-1.56	-2.01	-1.79	-0.75	-0.84	-0.69	0.48	0.56	0.41	1.84	2.30	2.07
17	-1.74	-2.19	-1.97	-0.70	-0.80	-0.63	0.40	0.46	0.30	2.05	2.53	2.30
18	-1.74	-2.24	-2.02	-1.10	-1.19	-1.03	0.65	0.74	0.60	2.19	2.68	2.45
19	-1.61	-2.22	-2.08	-1.29	-1.33	-1.13	0.83	0.97	0.90	2.07	2.59	2.31
20	-1.57	-2.04	-1.84	-0.97	-1.05	-0.89	0.38	0.47	0.34	2.16	2.62	2.38
22	-1.21	-1.80	-1.66	-0.78	-0.80	-0.58	0.09	0.23	0.16	1.90	2.36	2.08
24	-1.81	-2.34	-2.14	-0.84	-0.93	-0.74	0.53	0.63	0.49	2.12	2.64	2.39

Note that standard errors are not shown. They vary from 0.01 to 0.06 with 0.02 being the modal value.

Table 3: Item threshold range for three person-fit conditions

	Threshold range Tau 4 - Tau 1		
Item No.	All cases	Under-fits removed	All misfits removed
1	4.13	5.11	4.70
2	3.46	4.43	4.02
5	3.42	4.41	4.00
6	3.67	4.61	4.18
7	3.36	4.14	3.70
10	3.69	4.74	4.31
11	2.57	3.35	2.93
12	2.49	3.19	2.76
13	3.81	4.82	4.39
14	4.43	5.50	5.05
15	3.40	4.31	3.86
17	3.79	4.72	4.27
18	3.93	4.92	4.47
19	3.68	4.81	4.39
20	3.73	4.66	4.22
22	3.11	4.16	3.74
24	3.93	4.98	4.53

DISCUSSION

This study has shown that the inclusion of misfitting cases, especially under-fitting ones, during instrument calibration has a significant effect on item location parameters and on item threshold estimates. In particular, variation among item locations is reduced and item steps are truncated by retaining misfitting cases. However, poor person-fit does not seem to influence item fit statistics and the refinement of instruments to detect and remove misfitting items can proceed while retaining misfitting cases. Retaining poorly fitting cases does seem to compromise the overall model fit slightly when the hypothesised structure of the instrument is being evaluated using confirmatory factor analysis.

Table 4: Summary fit statistics from confirmatory factor analyses

	Nested factor model with all cases	Nested factor model with no under-fits	Nested factor model with no misfits
N	51631	46182	39089
χ^2/df	143.50	121.53	103.59
RMSEA	0.053	0.051	0.051
RMR	0.027	0.025	0.026
GFI	0.97	0.97	0.97
PMI	0.65	0.65	0.58

Note: RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual; GFI = Goodness of Fit Index; PMI = Parsimony Fit Index

The influence of person misfit on item parameters has implications for the calibration of attitude instruments and for routine procedures in checking model fit under the Rasch measurement model. Criteria for identifying person misfit in attitude surveys must be more firmly established, as these instruments cannot be taken as completely analogous to objective tests.

First, some of the causes of misfit in objective tests, such as guessing and cheating, do not apply to attitude surveys, but other threats to their measurement validity, such as carelessness or disinterest, may. Attitude surveys are not normally high stakes tests for the individuals taking them, and this is certainly true for the CEQ. The 51,631 cases analysed in this study represent a sub-set of all respondents. These were selected originally because they responded to all 25 items. Some 12,000 cases with incomplete forms had been removed previously. Many of those had only answered the first few items. Of the 5,384 under-fitting cases, it is not clear that all had followed a simple and thoughtless pattern of choosing only the first or the last option for all items. Closer analysis of patterns of responses is required before a sound rationale can be established for removing all under-fitting cases from data sets, and cut-off criteria that engender greater confidence than the simple 'Infit t > 2' rule that was applied in this analysis must be found.

Second, judgments of misfit are based on conditional probabilities of selecting particular response options to items for individuals with a given level on the latent trait. However, differences in the probabilities of selecting a particular response option compared with adjacent ones may not be great. This is partly attributable to the categories that are offered to participants in attitude survey instruments. What does 'Strongly Agree' mean compared with 'Agree' to a respondent, and does it have the same meaning for all respondents? Answers to these questions are far from clear, and this may account for much variance in respondents' choices of options. It is desirable to attach a much more precise meaning to each response option and to convey this clearly to respondents. It is instructive to compare this situation with

that of a rating scale in judging performance. Similar data are generated, but with judged rating scales, there are more exacting criteria to determine which option on the scale is to be applied. Judged rating scales are likely to yield more precise estimates of thresholds and of persons being rated than are attitude surveys unless greater clarity can be given to response options.

Third, in the Rasch measurement model, the claim is made that item parameters are (person) sample independent and that case estimates are item independent. However, it must be understood that the items used represent a universe of possible items that reflect the trait being assessed. Items that are judged not to 'fit the scale' are rejected as not being part of the universe of trait-related items: they reflect, or are contaminated by, a different trait. Similarly, persons who do not fit are judged not to conform to the typical response pattern of the majority of respondents. There may be several reasons for this. Some ill-fitting individuals may have responded carelessly, but there are other plausible reasons for poor person-fit. Some items may have particular salience for a subset of respondents and their responses to these items are likely not to conform with the 'standard' pattern. This is a case of differential item function (DIF). The inclusion of these persons in calibration may distort the parameter estimates for those items. However, with DIF there is always uncertainty about whether the item discriminates against some persons, that is whether there is item bias, or whether these persons have a genuinely different level of attitude (or performance) on that item. There is some evidence of DIF in the CEQ data as respondent sex, age, and NESB status were all found to be related to perceived course quality. If the structure of the instrument is to be validated, alternatives to simple confirmatory factor analysis are required. Where clear evidence of careless patterned responses is found, those cases can be removed. Where individual difference attributes are thought to be related to responses, such patterns need to be factored into the response model.

Although the current study has found partial answers to some of the questions that were posed, some new issues have surfaced through this analysis and they require further investigation. These matters are now canvassed briefly.

RECOMMENDATIONS

The results that have been found in this study warrant replication using other data sets obtained from the administration of attitude survey instruments. The conclusions of the present study are that item parameters are influenced by the inclusion of misfitting cases, but that the items retained through refinement using the Rasch measurement model are not. Although in this case, the refinement of the instrument was verified through the use of CFA, it seems, from the analysis of other attitude data sets not reported here, that CFA is more sensitive to response variance than is the Rasch model. That is, sets of items found to cohere under Rasch analysis do not always show such good fit under CFA. It is possible that CFA is more sensitive than Rasch modelling to this 'error variance' and that the poor model fit that emerges from CFA may be somewhat spurious. To resolve this issue, an analysis of residuals is required and through this it must be shown that item residuals are not inter-correlated in order to support the claim for unidimensionality that is implicit in the Rasch modelling. However, it is also possible that correlations among residuals can occur, not because of a failure of the unidimensionality of the item set, but because of a lack of homogeneity within the person sample. Some evidence of this was reported in the study of the CEQ (Curtis, 1999). Therefore, a more complex form of analysis that examines both the hypothesised structure of the instrument interacting with aspects of the person sample is required. For this, a form of multi-trait multi-method analysis is suggested.

In order to develop a better understanding of the effect on person-fit statistics of different pattern of responses simulated data studies are suggested. However, care must be taken to ensure that the synthetic data generated for these studies reflect the particular characteristics of attitude data. Attitude data appear to show greater variation in response choices for a given trait level, and this results in higher standard errors in both item parameter estimates and scaled

person scores. Attitudes can also be highly varied between individuals and in order to obtain reliable estimates of persons with more extreme levels of the trait, items that tap these extremities are required. However, for the majority of individuals, such items are far from their locations and responses to those items tend to be quite skewed. These skewed response patterns may influence the estimation of items parameters and case fit statistics, and in turn must influence the criterion values of those fit statistics. A study using carefully generated simulated data may shed light on the difficult question of person-fit criteria.

REFERENCES

Adams, R.J., & Khoo, S.T. (1997). Quest: the interactive test analysis system [Statistical analysis software]. Melbourne: Australian Council for Educational Research.

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. (1999). The 1996 Course Experience Questionnaire: A Re-Analysis. Unpublished Ed. D. dissertation, The Flinders University of South Australia, Adelaide.

Joreskog, K.S., & Sorbom, D. (1993). LiSRel for Windows (Version 8.12a) [Statistical analysis software]. Chicago: Scientific Software International.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. Journal of Applied Measurement, 1(2), 152-176.

Kline, P. (1993). Rasch scaling and other scales. The handbook of psychological testing. London: Routledge.

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.

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

Waugh, R.F. (1999). Approaches to studying for students in higher education: A Rasch measurement model analysis. British Journal of Educational Psychology, 69, 63-79.

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

Wright, B.D., & Stone, M.H. (1979). Best test design. 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.