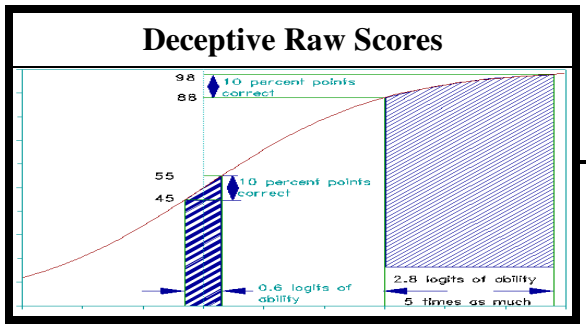


RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 20 No. 1 | Summer 2006 | ISSN 1051-0796

Data Variance Explained by Rasch Measures

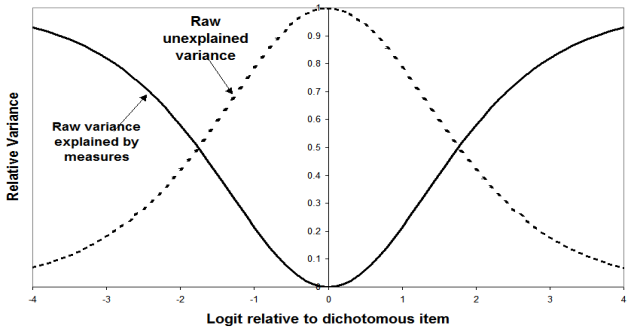


Fig. 1.Variance decomposition of a dichotomy.

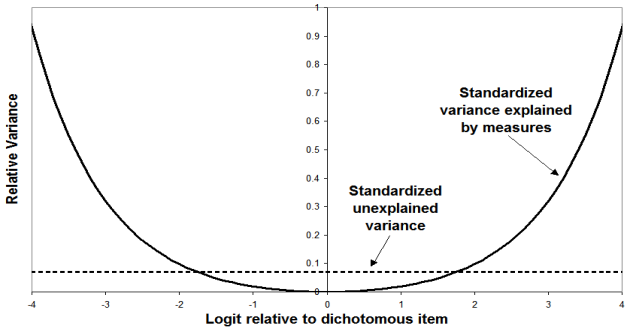
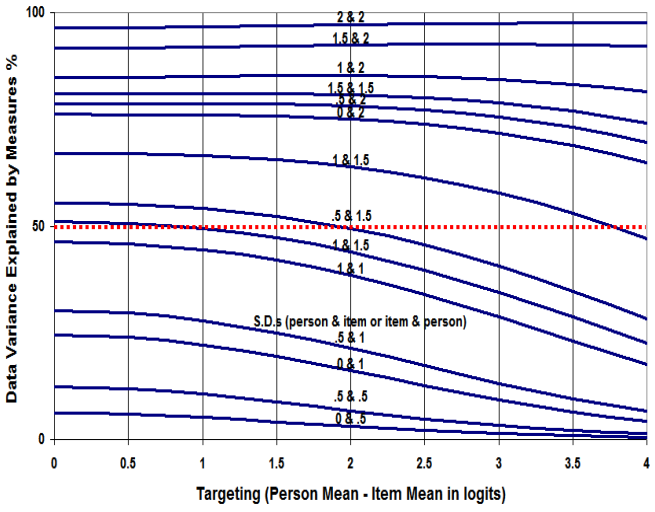


Fig. 2. Decomposition with standardized variance.

The Rasch model predicts that each observation will contain an exactly predictable part generated by the Rasch measures, and a well-behaved random component whose variance is also predicted by the model.

Figure 1 shows that, for dichotomous observations, as the logit measure difference between the person and the item increases (x-axis), the variance explained by the measures also increases (solid line) and the unexplained variance decreases (dotted line). When an item is as difficult as a person is able (0 on the x-axis), the outcome is completely uncertain. It is like tossing a coin. None of the variance in the observation is explained by the Rasch measures.



In Figure 2, the unexplained variance has been standardized to be the same size for every dichotomous observation. Thus each observation is modeled to contribute one unit of statistical information. An effect is to down-weight the central high-unexplained-variance observations. Standardized variances are used in the computation of standardized residuals which form the basis of several indicators of fit to the Rasch model.

In Figure 3, the decomposition of the variance in the data is shown for different combinations of item and person variance and item-person targeting. The unexplained variance has been standardized across observations a in Fig. 2. It is seen that the sum of the person S.D. and item S.D. must exceed 2 logits in order that over 50% of the standardized variance in the data be explained by the Rasch measures.

In the equivalent plot for raw variances, the y-axis values are about half of those plotted in Figure 3. Thus the sum of the person and item S.D.s must exceed 3 logits for over 50% of the raw observational variance to be explained.

John M. Linacre

Table of Contents

Toward Practical Significance of Rasch Scores in International Studies in Education: More than Statistical Significance and Effect Size

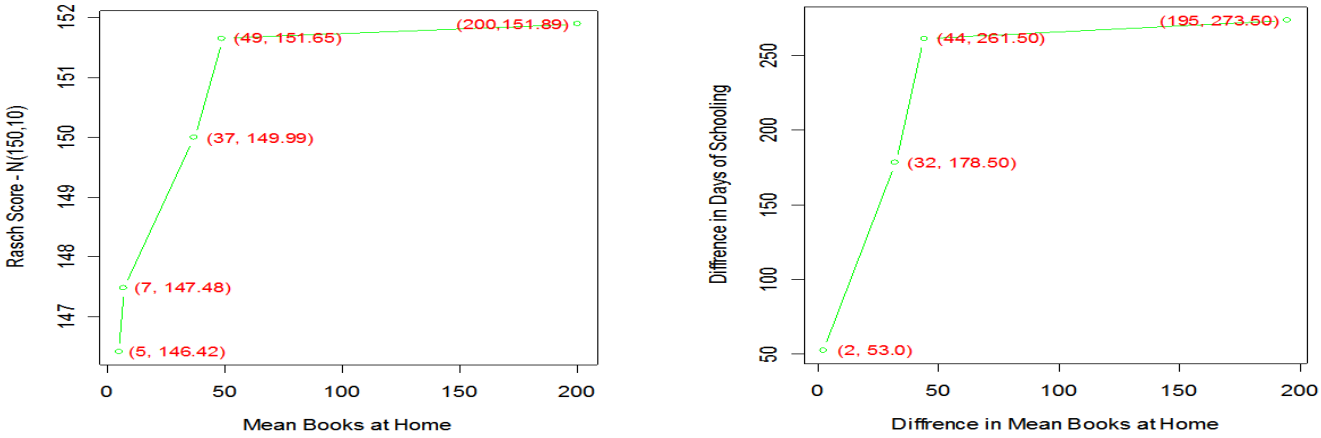


Figure 1. *Relation between Rasch and TEIMS 2003 scores, schooling days and number of books at home.*

Both in social sciences and in education, the scales that are used can easily and accurately lead to statistical significance, calculation of effect size indices, or computation of the confidence interval . However, unfortunately, practical significance of these scores can scarcely be obtained. For example, what does it mean to a teacher to be two points happier with his job than another one on a Likert scale? Almost worse: What does it mean to a school to be compared to the average national results in mathematics? Just to say that there is a statistical significant difference between the school score and the national average scores is not very informative for the school.

Statisticians proposed a compromise to these kinds of situations. For this purpose, they suggest using effect size indices and confidence intervals in replacement of significance tests. They are interesting devices, though not as much when it comes to the practical significance of scores and to the difference between these scores.

In the context of comparisons made with the results of international and national surveys in education, like the TEIMS, PIRLS and PISA, practical significance of these scores would lead researchers to more useful interpretations. It would lead to more meaningful comparisons of countries and states. Important decisions would be easier to make.

In the context of these international surveys, researchers tried to find a way to lead to practical significance of the obtained mathematics, science, or reading literacy scores. For example, Cartwright, Lalancette, Mussio, and Xing (2003) related these international scores to results at a smaller provincial scale in Canada. They found that the provincial results in reading of British Columbia can be related with a good precision to the international PISA reading comprehension literacy score. While these findings don't permit a direct practical interpretation, they tell us, at least, that it is possible to link results obtained at one survey, the PISA, to those obtained at another one, the British-Columbia.

More interesting, and more practical, is Willms' proposition (2004a, 2004b). Willms tries to translate the international Rasch scores results from PISA into a metric of *school years* and *school days*. This is an attractive idea, because it leads directly to a metric that allows administrative decision and teaching effect interpretation. Thus, it would be legitimate to say that to reach an

augmentation of one point on one of PISA literacy scales, a fixed number of schooling days would be necessary. Willms, in preliminary works, finds out that a difference of one point on the PISA scale is equivalent to about 5 schooling days. He didn't report on which literacy scale the equivalence was computed. He did this estimation by considering the fact that, because of the annual inscription date in each state and country, the 15-years-old students participating in the survey can be on two different school levels. This fact permits him to compute the effect of one schooling year on the PISA scores. The average schooling year in the 12 countries, where he was able to obtain the information about the level and the date of inscription, was equal to 172 days. Wherefore he observed that a schooling year correspond to a difference score of 34.30, he extrapolates that a one point difference is equivalent to about 5 schooling days (172/34.30).

By themselves, Willms' findings are of interest, being now able to give a clear interpretation of differences in literacy scores. However, if we also consider Cartwright *et al.*, we can think that the equivalence formula obtained by Willms can be related to the one of other international and national educational surveys. Consider that the PISA scores are on a scale with a mean of 500 and a standard deviation of 100 and that, for example, the TEIMS Rasch scale has a mean of 150 and a standard deviation of 10. Thus, only using the standard deviation ratio, a one point difference at the TEIMS scale would be equivalent to a 10 points difference at the PISA scale: σ_{PISA} / σ_{TIMSS} = 100/10. An illustration of the relation between TIMSS 2003 scores, schooling days, and number of books at home is presented in Figure 1. It can be seen that a mean difference of 32 books at home (37 - 5) corresponds approximately to one school year (178.50 days).

A note of caution is to be considered. More research work is still to be done on this topic and it will have to consider the specific literacy scale, how old the students are, as well as their school level. More important, however, we think that the equivalence between the schooling days and the difference in the literacy scores would have to be computed independently for each country, not by an average on the 12 countries used by Willms which are so different in schooling practice.

Gilles Raîche, Université du Québec à Montréal
Claire IsaBelle, Université d'Ottawa
Martin Riopel, Université du Québec à Montréal

Cartwright, F., Lalancette, D., Mussio, J. and Xing, D. (2003). *Linking provincial student assessments with national and international assessments.* Report no 81-595-MIE2003005. Ottawa, Canada: Statistics Canada.

Willms, J. D. (2004a). *Reading achievement in Canada and the United States : Findings from the OECD programme for international student assessment*. Final Report no SP-601-05-04E. Ottawa, Canada: Human Resources and Skills Development Canada.
auxweb.unb.ca/applications/crisp/pdf/0404i.pdf

Willms, J. D. (2004b). *Variation in literacy skills among Canadian provinces: Findings from the OECD PISA.* Report no 81-595-MIE2004012. Ottawa, Canada: Statistics Canada.

Rasch-related Coming Events

Sept 2006 - Dec 2007 3-day Rasch courses, Leeds, UK
home.btconnect.com/Psylab_at_Leeds/

Oct. 14-15, 2006, Sat.-Sun. Introduction to Rasch Measurement: Theory and Applications, Chicago IL (Smith & Smith) www.jampress.org

Oct. 20-Nov. 17, 2006, Fri.-Fri. Practical Rasch Measurement (Winsteps) on-line course (Linacre)
www.statistics.com/courses/rasch

Nov. 1, 2006, Wed. Using Rasch to Measure Services and Outcomes (Conrad & Bezruczko)
www.eval.org/eval2006/aea06.PDW.htm

Dec. 11-13, 2006, Mon.-Wed. Objective Measurement in the Social Sciences - ACSPRI Conference, Australia
www.acspri.org

April 9-13, 2007, Mon.-Fri. AERA Annual Meeting
Chicago, Illinois, www.aera.net

ACSPRI Social Science Methodology Conference

The University of Sydney, Sydney, Australia

December 10-13, 2006

Call for Papers and Participation

The Australian Consortium for Social and Political Research, Inc. (ACSPRI) Conference focuses on current issues in social science methodology. Plenary presentations will be given by:
Prof. Merrill Shanks, Univ. of California, Berkeley.
Prof. Michael Greenacre, Univ. Pompeu Fabra, Barcelona, Spain.

The Conference will accept papers based on an abstract of 500 words. It is expected that, if accepted for the conference, presenters will submit a draft of a full paper before Dec 1, 2006, for inclusion in the conference proceedings. Papers can be peer-reviewed and this will be indicated in the Conference Proceedings. The Conference Website contains details of the submission procedure.

Oct 2, 2006 : Final deadline for proposals

Nov 1, 2006 : 'early bird' enrolments close

December 10-13, 2006: (Sunday afternoon through Wednesday morning): Conference

For information and conference updates visit
www.conference2006.acspri.org.au

Unidimensionality Matters! (A Tale of Two Smiths?)

Introduction

The Rasch measurement model is a unidimensional measurement model and this attribute has been the subject of much discussion in the Transactions (Stahl J 1991; Wright BD 1994; Linacre JM 1994; Fisher WP 2005). In an early article Wright and Linacre tell us that 'whether a particular set of data can be used to initiate or to continue a unidimensional measuring system is an empirical question (Wright BD, Linacre JM 1989). The only way it can be addressed, they argue, is to 1) analyze the relevant data according to a unidimensional measurement model, 2) find out how well and in what parts these data do conform to our intentions to measure and, 3) study carefully those parts of the data which do not conform, and hence cannot be used for measuring, to see if we can learn from them how to improve our observations and so better achieve our intentions'. (MESSA Memo 44, reprinted from Wright BD, Linacre JM 1989). Smith uses simulation to investigate which technique is better at discovering dimensionality (Smith RM, 1996). A review of these findings in RMT (9:4, 1996) argues that the conclusions are simple. 'When the data are dominated equally by uncorrelated factors, use factor analysis. When they are dominated by highly correlated factors, use Rasch. If one factor dominates, use Rasch'.

In Rasch analysis the understanding and detection of unidimensionality in the context of medical and psychological studies has developed and changed much in the past 15 years. Early published articles subscribed to the notion that fit to the model supported the unidimensionality of the scale and little else was done to confirm that assumption (Tennant A, et. al 1996). In the 1990's Wright had put forward a Unidimensionality Index (Wright BD, 1994), and gradually greater emphasis was placed on analysis of the residuals and particularly a Principal Component Analysis (PCA) of the residuals to detect second factors after the 'Rasch Factor' was removed. Originally interpretation of this was difficult as the proportion of variance attributable to the first residual factor was reported, but the total variation in the data was unknown. Subsequently Winsteps (Linacre JM, 2006) has incorporated the total variation into its reporting, so the magnitude of the first residual factor against the Rasch factor can be determined. In 2002 Smith reported an independent t-test approach to testing for unidimensionality (Smith EV, 2002, JAM) which is being incorporated into the latest RUMM2020 software (Andrich, D., Lyne A, Sheridan B., Luo G, 2003). Elsewhere, others have used classical factor analytical approaches to testing for unidimensionality prior to fitting data to the Rasch model (Bjorner JB, Kosinski M, Ware JE Jr, 2003).

A review of the literature suggests that there are three main approaches to assessing dimensionality:

a) prior testing using classical approaches, such as factor analysis;

b) those which hold to the assumption of fit equals unidimensionality – a fit only approach;

c) those which involve post-hoc testing, having undertaken the Rasch analysis and supposing fit to the Rasch model (e.g., PCA of the residuals).

Thus it is possible to conceive of a broad selection of tests which may be undertaken for any given data set. For the everyday user of Rasch software working in the health and social sciences, how can they be sure that they are truly dealing with a unidimensional construct? How far do these various tests detect multidimensionality in the data?

Methods

The aim of this present study is to contrast commonly used techniques from each of the three main approaches identified above by applying them to a set of simulated datasets with known dimensionality characteristics. Each data set is based upon 20 polytomous items with 5 response options (0-4) and 400 cases. Details of the

Table 1. Details of Datasets

Dataset	Structure	Contents
1	Unidimensional	20 items.
2	Two orthogonal dimensions (r<.05)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest) . Interlaced items with item 1 assigned to dimension1, item2 assigned to dimension 2.… to ensure equal difficulty for each dimension
3	Two orthogonal dimensions (r<.05)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest) . Dimensions stacked with easy items 1-10 in Dimension 1, and hardest items 11-20 in Dimension 2
4	Two orthogonal dimensions (r<.05)	16 items in Dimension 1 and 4 items in Dimension 2 . (items 5,10,15,20). Items generated in difficulty order (1=easiest, 20=hardest)
5	Two correlated dimensions (r =.70)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest) . Interlaced items with item 1 assigned to dimension1, item2 assigned to dimension 2… to ensure equal difficulty for each dimension
6	Two correlated dimensions (r =.70)	10 items in each dimension. Items generated in difficulty order (1=easiest, 20=hardest) . Dimensions stacked with easy items 1-10 in Dimension 1, and hardest items 11-20 in Dimension 2

datasets are outlined in Table 1. A series of analyses were conducted on each of the 6 data files to assess dimensionality (Table 2). SPSS Version 14.0 was used to conduct factor analysis, and both Winsteps and RUMM2020 were used to conduct Rasch analysis. The data were simulated using SIMMsDepend (Marais I,2006).

We have chosen procedures from SPSS because it is widely available and easy to use. Principal components analysis (PCA) was used to extract the factors followed by oblique rotation of factors using Oblimin rotation (delta = 0). Kaiser's criterion, which retains eigenvalues above 1, was used in Procedure 1.1 to guide the identification of relevant factors. In Procedure 1.2 Horn's parallel analysis (Horn JL, 1965), which has been identified as one of the most accurate approaches to estimating the number of components (Zwick & Velicer, 1986), was used. The size of eigenvalues obtained from PCA are compared with those obtained from a randomly generated data set of the same size. Only factors with eigenvalues exceeding the values obtained from the corresponding random data set are retained for further investigation. Parallel analysis was conducted using the software developed by Watkins (2000). Analyses were also conducted using a non-linear Factor Analysis (HOMALS) available in SPSS. Using curve estimation and a quadratic function, the values exported from the HOMALS procedure can be tested to determine the number of dimensions in the data.

For the Rasch procedures we set both Winsteps and RUMM2020 to have identical convergence criteria. As none of the data sets satisfied the assumptions of the rating scale model, we use the unrestricted (partial credit) polytomous model. A number of different fit statistics are reported. OUTFIT ZSTD in Winsteps and Residuals in RUMM are equivalent, with any variation reflecting the difference in the underlying estimation procedures. We use the value 2.5 and above for both (≈ 99% significance) to determine misfit to model expectation. Usually the two statistics provide similar magnitudes of fit to the model.

INFIT and OUTFIT MNSQ (Winsteps) are also reported with acceptable ranges of 0.9-1.1 and 0.7-1.3 respectively, following Smith's recommendations for sample size adjustment (Smith RM et al, 1998). RUMM Chi-Square probabilities are also reported, Bonferroni adjusted to 0.0025 and unadjusted. We also report the RUMM Chi Square Interaction Fit Statistic which is a summary fit statistic and widely used to indicate overall fit to the model. We also report Wright's Unidimensionality Index which is the person separation using model standard errors, divided by the person separation using real (misfit inflated) standard errors (Wright BD, 1994). A value above 0.9 is indicative of unidimensionality; 0.5 and below of multidimensionality and everything between is the usual grey area of uncertainty!

We report the usual Principal Component Analysis (PCA) of the residuals, including the percentage of variance attributable to the Rasch factor and the first residual factor (usually identical in Winsteps and RUMM), and the percentage of variance attributable to the first residual factor out of total variance (Winsteps).

Finally, we report on a comparison of person estimates based upon subsets of items. In practice where there is a conceptual basis for multidimensionality estimates are made from the a-priori dimensions. In the present case with this simulated data, we use the item loadings on the first factor of the PCA of the residuals. Person estimates derived from the highest positive set of items (correlated at 0.3 and above with the component) are contrasted against those derived from the highest negative set. A series of independent t-tests are undertaken to compare the estimates for each person and the percentage of tests outside the range ±1.96 is computed, which follows Everett Smith's general approach (Smith EV, 2002). A

Table 2. Details of Procedures

Prior testing	1.1 Default SPSS Principal Components Analysis using Kaiser's criterion, retaining eigenvalues above 1.
	1.2 Default SPSS Principal Components Analysis with Horn's parallel analysis to determine significant eigenvalues.
	1.3 HOMALS non linear factor analysis
Fit to the Rasch model	2.1 Percentage of items which misfit the (polytomous) model OUTFIT ZSTD (Winsteps).
	2.2 Percentage of items which misfit the (polytomous) model Residuals (RUMM).
	2.3 Percentage of items showing INFIT MNSQ misfit (Winsteps).
	2.4 Percentage of items showing OUTFIT MNSQ misfit (Winsteps).
	2.5 Percentage of items showing Chi-Square misfit (RUMM).
	2.6 Percentage of items showing Chi-Square misfit (RUMM), Bonferroni corrected
	2.7 Summary Fit statistics.
	2.8 Wright's Unidimensionality Index.
	2.9 Person Separation Index (RUMM)
	2.10 Person Separation (real) (Winsteps)
Post Hoc tests	3.1 Percentage of variance attributable to the Rasch factor
	3.2 Percentage of variance attributable to the first residual factor
	3.3 Ratio of variance attributable to first residual factor compared with Rasch factor (Winsteps)
	3.4 Percentage of individual t-tests outside the range ± 1.96 (RUMM2030) with Binomial Test for Proportion confidence intervals where appropriate.

Binomial Proportions Confidence Interval can be calculated for this percentage. The Binomial CI should overlap 5% for a non-significant test. The results of these analyses are reported in the Table 3.

Results

The default factor analysis (1.1) failed to identify the single dimension, instead, identifying two 'difficulty' dimensions. The HOMALS procedure failed to detect the situation (specified in Set 4) where only four items belonged to a second dimension, and consistently failed where the correlation between factors was ≈ 0.7. The Rasch model fit statistics performed poorly where dimensions were interlaced and where the correlation between factors was ≈ 0.7. Wright's Unidimensionality Index appeared insensitive to multidimensionality. Little can be gleaned from the percentage of variance attributable to the Rasch factor, as this seems consistently high, irrespective of the underlying dimensionality. In Set 1 the percentage of variance attributable to the first residual factor was substantially lower than in other sets, but the percentage of variance out of the total variance was low, except for the orthogonal data sets. The independent t-test approach consistently identified the unidimensional and multidimensional data sets.

These results have a number of implications for everyday practice of Rasch analysis. In the construction of a new polytomous scale where the intention is to create a unidimensional construct, Rasch fit statistics may mislead if there are two dimensions where the items are interlaced in difficulty. Supporting Richard Smith's (1996) recommendation, exploratory factor analysis should be undertaken at the outset to make sure that dimensionality is not going to be a problem, or to identify which items may be problematic so as to inform the iterative Rasch analysis procedure. As we cannot know in advance whether or not two interlaced dimensions may exist, this analysis should be undertaken as a matter of routine. The simplest way to undertake this is with the default factor analysis procedure using the parallel analysis to determine the number of significant eigenvalues.

Although the PCA of the residuals may give clues to multidimensionality in the data, their interpretation is not straightforward. The percent of variance of the first residual factor (out of total variance in the residuals) does show a clear increasing trend from the unidimensional data, through the correlated factors to the orthogonal factors. However, at what point does this figure shift from a unidimensional indicator to a multidimensional indicator?

The individual t-test approach proposed by Everett Smith seems the most robust in that it clearly identifies dimensionality. This test has importance not just for the interpretation of unidimensionality, but also the meaning of multidimensionality in the data. Note that the proportion of t-tests outside the range is high across Sets 2-6, even when the factors are correlated at ≈ 0.7. In

Table 3. Summary of Results of Analyses

Test	Dataset:	1	2	3	4	5	6
	Prior Tests – Number of Factors						
1.1	EFA with eigenvalue>1. (% Variance 2nd factor)	2 (6%)	2 (30%)	2 (31%)	2 (14%)	2 (63%)	2 (63%)
1.2	EFA with parallel analysis	1	2	2	2	2	2
1.3	HOMALS – number of factors	1	3	2	1	1	1
	Rasch Fit						
2.1	% OUTFIT ZSTD out of range	0	0	0	100	5	0
2.2	% Residuals outside range	0	0	0	85	0	0
2.3	% INFIT MNSQ out of range	5	0	5	100	20	15
2.4	% OUTFIT MNSQ out of range	0	0	0	60	0	0
2.5	% Chi-Square significant	0	5	70	100	0	0
2.6	% Chi-Square significant (Bonferroni adjusted)	0	0	35	70	0	0
2.7	Item-Trait Interaction Fit statistic	0.74	0.09	0.00	0.00	0.97	0.12
2.8	Wright's Unidimensional Index	1.08	1.11	1.11	1.12	1.07	1.08
2.9	Person Separation Index ≈ α	0.91	0.88	0.89	0.93	0.95	0.95
2.10	(Real) Person Separation	3.12	2.44	2.56	3.59	4.04	4.09
	Post Hoc tests						
3.1	% variance attributable to the Rasch factor.	82.0	70.0	70.6	76.9	85.2	84.8
3.2	% variance attributable to first residual factor	7.4	48.8	47.5	25.4	26.3	23.8
3.3	% variance attributable to first residual factor out of total variance	1.4	14.3	14.1	6.4	3.8	3.7
3.4	Percentage of individual t-tests outside range ± 1.96 (95% CI) where needed	7.0 (5-9%)	55.0	51.5	45.3	38.8	35.0

practice this means that person estimates differ by between 1 to 2 logits, depending upon which set of items are being used for that estimate. This variability in person estimate is unsustainable when scales are to be used for individual clinical use, for example where cut points are often used to determine clinical pathology. The variability of person estimates where multidimensionality exists also raises fundamental questions about Computer Adaptive Testing approaches which rely upon estimates based upon just a few variables. Clearly, only the strictest form of unidimensionality must be used to avoid significantly different person estimates driven by multidimensionality.

The analysis we have undertaken is only at the simplest level, reflecting what is most likely to be used in everyday research practice in the health and social sciences. We have, for example, not used Monte Carlo simulation or other methods to look at ranges of variance explained. Neither have we looked at different sample sizes or different test lengths. We have not addressed dichotomous items, which bring their own set of problems to factor analysis. Nevertheless, we believe that this simple analysis has shown that great care needs to be taken in confirming the assumption of unidimensionality of data when fitted to the Rasch model. Perhaps others may pursue some of the issues we have omitted.

Conclusion

When developing new polytomous scales, an exploratory factor analysis used a priori, with parallel analysis to indicate significant eigenvalues, should give early indications of any dimensionality issues prior to exporting data to Winsteps or RUMM. This should identify the situation of equal number of items on two factors which will not be detected by the Rasch analysis fit statistics and where the PCA of the residuals may be indeterminate. After fit of data to the Rasch model, careful examination of the PCA of the residuals should provide clues to any remaining multidimensionality. Comparison of person estimates derived from these subsets of items, using the independent t-test approach, should confirm or reject the unidimensionality of the scale.

Alan Tennant PhD[1], and Julie F. Pallant PhD[2].

[1] Academic Unit of Musculoskeletal & Rehabilitation Medicine, Faculty of Medicine and Health, The University of Leeds, UK.

[2] Faculty of Life and Social Sciences, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia

References

Andrich D, Lyne A, Sheridan B, Luo G. (2006). RUMM 2020. Perth: RUMM Laboratory

Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the application

of item response theory to the headache. Quality of Life Research 2003; 12: 913–933.

de Bonis M, et al. The Severity of Depression. Rasch Measurement Transactions, 1992; 6:3 p. 242-3

Horn JL A rationale and test for the number of factors in factor analysis. Psychometrika 1965; 30:179-185.

Linacre JM. DIMTEST diminuendo. Rasch Measurement Transactions, 1994, 8:3 p.384

Linacre JM. Winsteps Rasch measurement computer program. Chicago: Winsteps.com, 2006.

Fisher W.P. Jr. Meaningfulness, Measurement and Item Response Theory (IRT). Rasch Measurement Transactions, 2005, 19:2 p. 1018-20

Marais I. SIMMsDepend. Murdoch University, Western Australia, 2006.

Raîche G. Critical eigenvalue sizes in standardized residual Principal Components Analysis. Rasch Measurement Transactions, 2005, 19:1 p. 1012

Schumacker RE, Linacre JM Factor analysis and Rasch. Rasch Measurement Transactions 1996, 9:4 p. 470.

Smith EV. Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement 2002; 3:205-231.

Smith RM. A Comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modeling 1996; 3:25-40.

Smith RM et al. Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement 1998; 2:66-78

Stahl J. Lost in the Dimensions, Rasch Measurement Transactions, 1991; 4(4):120

Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Brit J Rheum 1996; 35: 574-578.

Watkins MW: Monte Carlo PCA for Parallel Analysis [software]. State College, PA: Ed & Psych Associates; 2000. www.personal.psu.edu/mww10/Watkins3.html

Wright BD. Unidimensionality coefficient. Rasch Measurement Transactions, 1994; 8:3 p.385

Wright B.D. Rank-ordered raw scores imply the Rasch model. Rasch Measurement Transactions, 1998, 12:2 p. 637-8.

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Archives of Physical Medicine and Rehabilitation 1989; 70: 857-860.

Zwick, WR, Velicer WF. Comparison of the rules for determining the number of components to retain. Psychological Bulletin, 1986; 99: 432-442.

Dichotomous Equivalents to Rating Scales

There are numerous ways to conceptualize rating scales. One useful conceptualization is to imagine that the rating scale is equivalent to a set of dichotomous items. Huynh Huynh investigated this: Huynh H. (1994*) On equivalence between a partial credit item and a set of independent Rasch binary items.* Psychometrika, 59, 111-119, and Huynh H. (1996) *Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items.* Psychometrika, 61, 31-39.

A crucial finding is that the Rasch-Andrich thresholds must advance (i.e., not exhibit "threshold disordering") for a polytomy to have the mathematical properties of a set of dichotomies. But merely advancing is not enough.

Consider a polytomy with *m+1* ordinally advancing categories. There are *m* transition points, so this could be conceptualized as *m* dichotomies. As the Rasch-Andrich thresholds for the polytomy become further apart then the set of dichotomous items would have a wider difficulty range. The boundary condition is that the *m* dichotomies be of equal difficulty. Then a score of *k* on the polytomous item would be equivalent to scoring *k* on *m* equally-difficulty dichotomies.

A set of equally difficulty dichotomies constitute a set of Bernoulli (binomial) trials. The polytomous Rasch model for this is (with the familiar notation):

$$\log\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = B_n - D_i - \log\left(x/(m-x+1)\right)$$

This provides the lower limits by which Rasch-Andrich thresholds must advance in order that a polytomy have the same mathematical properties as a set of dichotomies. A useful rule-of-thumb is "thresholds must advance by one-logit". The exact values are tabulated below.

John Michael Linacre

Minimum Rasch-Andrich threshold advances for a polytomy to be equivalent to a set of dichotomies

Thresholds: ------ **Categories:**	**1 to 2**	**2 to 3**	**3 to 4**	**4 to 5**	**5 to 6**	**6 to 7**	**7 to 8**	**8 to 9**	**9 to 10**
3	1.39								
4	1.10	1.10							
5	.98	.81	.98						
6	.92	.69	.69	.92					
7	.88	.63	.58	.63	.88				
8	.85	.59	.51	.51	.59	.85			
9	.83	.56	.47	.45	.47	.56	.83		
10	.81	.54	.44	.41	.41	.44	.54	.81	
11	.80	.52	.42	.38	.36	.38	.42	.52	.80

Rating Scale Equivalence to Dichotomies

Question: We have 5 dichotomous items based on the same reading text. What would happen if we replace the dichotomies with a rating scale item scored 0-5?

Reply: This is an attractive option if the target dichotomies exhibit more local dependence than the other items on the test. Let us first suppose that the 5 items fit the Rasch model about as well as the other items on the test. Then here is a standard polytomous rating scale model:

$$\log\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = B_n - D_{ix}$$

and here is an example of one of the five items:

$$\log\left(\frac{P_{nk1}}{P_{nk0}}\right) = B_n - D_k$$

Then, for these to be equivalent,

$$D_{i1} = -\log\left(\sum_{k=1}^{5} e^{-D_k}\right)$$

$$D_{i2} = -\log\left(\sum_{k=1}^{5}\sum_{j=k+1}^{5} e^{-D_k - D_j}\right) - D_{i1}$$

$$D_{i3} = -\log\left(\sum_{k=1}^{5}\sum_{j=k+1}^{5}\sum_{h=j+1}^{5} e^{-D_k - D_j - D_h}\right) - D_{i1} - D_{i2}$$

and similarly for D_{i4}, D_{i5}. The overall difficulty of a polytomous item is given by

$$D_i = \sum_{k=1}^{m} D_{ik} \Big/ m$$

giving $F_{ik} = D_i - D_{ik}$ as the Rasch-Andrich thresholds.

In general, the overall difficulty of the polytomous item will not match the average difficulty of the equivalent dichotomous items. If the local origin of the measurement scale is set at the average difficulty of the items, it is seen that the local origin will also change location, so that the person measures will change value. But all person measures will change by about the same amount.

Example: Dichotomous item difficulties are {1,2,3,4,5} logits. Then the D_{ix} are {.55, 1.87, 3.0, 4.13, 5.45} logits.

So what will happen if your 5 items are more locally dependent than the other items on the test? If the local dependence increases their fit to the Rasch model, then they will be more discriminating than 5 independent Rasch dichotomous items. Consequently the $\{D_{ik}\}$ will be less dispersed than the values given above.

But if the local dependence is on a secondary dimension, so that it decreases their fit to the Rasch model, then the $\{D_{ik}\}$ will be more dispersed than the values given above.

John Michael Linacre

Meaningfulness, Sufficiency, Invariance and Conjoint Additivity

Consider the following statements from widely respected authorities in statistics and measurement:

"If there exists a minimal sufficient statistic [i.e., one that is both necessary and sufficient] for the individual parameter Theta which is independent of the item parameters, then the raw score is the minimal sufficient statistic and the model is the Rasch model" (Andersen, 1977, p. 72).

"The set of invariant rules based on a sufficient statistic is an essentially complete subclass of the class of invariant rules" (Arnold 1985, p. 275; citing Hall, Wijsman, & Ghosh, 1965).

"The hallmark of a meaningless proposition is that its truth-value depends on what scale or coordinate system is employed, whereas meaningful propositions have truth-value independent of the choice of representation, within certain limits. The formal analysis of this distinction leads, in all three areas [measurement theory, geometry, and relativity], to a rather involved technical apparatus focusing upon invariance under changes of scale or changes of coordinate system" (Mundy, 1986, p. 392).

Andersen (1977) shows that summing ratings to a score is meaningful and useful only if that score is a minimally sufficient statistic, and if that statistic exists, then the Rasch model holds. Arnold (1985) and Hall, Wijsman, and Ghosh (1965) show that statistical sufficiency is effectively equivalent with measurement invariance. Mundy (1986) shows that meaningful propositions all share the property of invariance. Luce and Tukey (1964) show that conjoint additivity is another way of arriving at the invariance characteristic of fundamental measurement.

These principles of meaningfulness, sufficiency, invariance, and conjoint additivity are ubiquitous in the production of scientific knowledge, which explains why we find so many strong statements in the history of science to the effect that measurement and quantification are absolutely essential to any science worthy of the name (Michell, 1990, pp. 6-8). We have, unfortunately, confused the mere use of number with meaningful measurement, when, in fact, it is the realization of the qualitatively mathematical ideal of invariance that makes science what it is. Even as unlikely a philosopher as Heidegger (1967, pp. 75-6), who was held by some to be, at best, a poet, understood that the broad qualitative sense of the mathematical is "the fundamental presupposition of all 'academic' work" and "of the knowledge of things."

Multiple harmonious definitions of meaningful measurement are effectively embodied in Rasch models (Fischer, 1995; Fisher, 2004; Wright, 1997). It then follows that the Rasch model's "singular significance for measurement is that it is a unique (necessary and sufficient) deduction from the (fundamental) measurement requirements of joint order and additivity" (Wright, 1984).

Analytic methods implementing Rasch measurement test the hypothesis of qualitative yet mathematical meaningfulness more effectively, easily and efficiently than any other available methods. It is the norm today to presume scientific status and the achievement of measurement even when sufficiency and invariance have not been tested or established. The day may soon be coming when such hubris will be considered tantamount to fraud. When that day arrives, research employing Rasch models will be sought after as paradigmatic examples of mathematically meaningful methodology.

William P. Fisher

Andersen, E. B. (1977). Sufficient statistics and latent trait models. Psychometrika, 42(1), 69-81.

Arnold, S. F. (1985, September). Sufficiency and invariance. Statistics & Probability Letters, 3, 275-279.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 15-38). New York: Springer-Verlag.

Fisher, W. P., Jr. (2004, October). Meaning and method in the social sciences. Human Studies: A Journal for Philosophy and the Social Sciences, 27(4), 429-54.

Hall, W. J., Wijsman, R. A., & Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. Annals of Mathematical Statistics, 36, 575-614.

Heidegger, M. (1967). What is a thing? (W. B. Barton, Jr. & V. Deutsch, Trans.). South Bend, Indiana: Regnery/Gateway.

Michell, J. (1990). An introduction to the logic of psychological measurement. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mundy, B. (1986). On the general theory of meaningful representation. Synthese, 67, 391-437.

Wright, B. D. (1984). Despair and hope for educational measurement. Contemporary Education Review, 3(1), 281-288 www.rasch.org/memo41.htm

Wright, B. D. (1997, June). Fundamental measurement for outcome evaluation. Physical Medicine & Rehabilitation State of the Art Reviews, 11(2), 261-88.

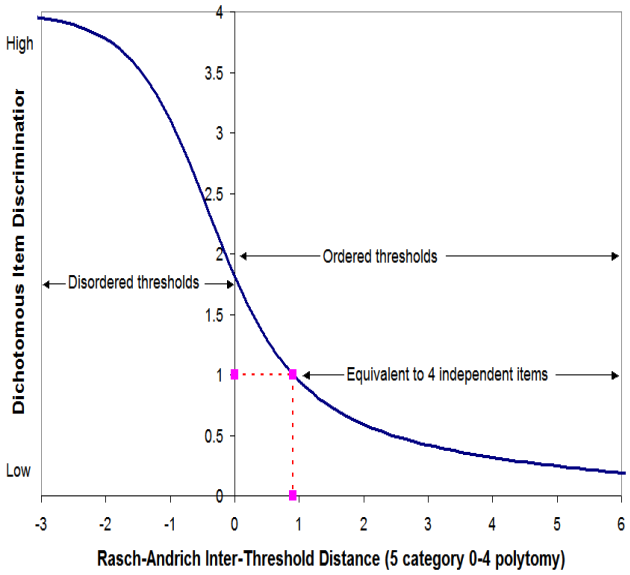
Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
www.rasch.org/rmt/
Editor: John Michael Linacre

SIG Chair: Thomas O'Neill, Secretary: Ed Wolfe
Program Chair: Trevor Bond
SIG website: www.raschsig.org

Item Discrimination and Rasch-Andrich Thresholds

Rasch dichotomous items are modeled to have the same (or known) discrimination. The inter-adjacent-category discrimination of polytomous Rasch items is also modeled to be the same (or known), but the overall discrimination of polytomous items depends on the distance between its Rasch-Andrich thresholds and so can vary across items and instruments..

Consider a five category rating scale, modeled by a set of 4 equally spaced Rasch-Andrich thresholds. Then the overall item discrimination can vary from very steep to almost flat depending on the distance between the threshold values.



The plot shows the relationship between uniform threshold spacing, *x,* and the item-discrimination slope, *a,* of an equivalent logistic ogive with the range y = 0-4 score points given by

$$\log\left(\frac{y}{4-y}\right) = ax$$

A discrimination of $a >= 1.0$ implies that the polytomy is equivalent to summing 4 independent dichotomous 0-1 items. When $a = 1.0$ the items are of equal difficulty. When the inter-threshold distance is negative, the Rasch-Andrich threshold are "disordered".

It can be seen that one advantage of using Rasch polytomies over independent dichotomous items is that a polytomy can provide higher item discrimination while maintaining the desirable measurement properties of a Rasch model. This is useful for items targeting pass-fail decisions and computer-adaptive testing. This is also one situation in which disordered thresholds can be advantageous.

John Michael Linacre

Making Metrics Less Arbitrary

"A metric, once made meaningful, can be used to provide perspectives about such things as the magnitude of change that occurs on an underlying dimension as a function of an intervention. Evidence that an intervention causes movement along a scale that has nonarbitrary meaning can reveal the real-world consequences of this change."

"It can be difficult and time consuming to conduct the research needed to make a metric less arbitrary. Fortunately, the issue of metric arbitrariness is irrelevant for many research goals, so not all researchers must tackle this issue. ... However, there are applied situations in which researchers need to address the issue if they are going to fulfill their research goals. Tying metrics to meaningful, real-word events provides a viable means of making metrics less arbitrary, but there will always be some guesswork involved. No new methodology is going to expose psychological constructs to the naked eye. Best estimates of where people stand on psychological dimensions are always that, estimates. Nevertheless, awareness of this limitation is of value to the psychologist. A researcher who appreciates the gap between a psychological metric and a psychological reality knows to look past a person's score and search for something meaningful."

Blanton, H., and J. Jaccard. 2006. Arbitrary metrics in psychology. American Psychologist 61(January):27-41.