Psychological Test and Assessment Modeling, Volume 52, 2010 (3), 313-322

A comparison of three polychotomous Rasch models for super-item analysis

Purya Baghaei[1]

Abstract

Local dependency is a prevalent phenomenon in educational tests where several dichotomous items are based on a single prompt. This is a violation of one of the major assumptions of Rasch and other IRT models and poses restriction on the analysis of such tests with these models. To solve the problem, it has been suggested that the items which belong to a single prompt be bundled together and analysed as independent polychotomous super-items. However, in the last few decades there has been an array of polychotomous models with different properties and assumptions which makes the choice of the right model rather difficult. The purpose of the present study is two-fold: 1) to compare the performance of three psychotomous Rasch models for super-item analysis and 2) to check the consequences of using 'inappropriate' models when the assumption of equal distances between steps within and across items is violated. To this end, a reading comprehension test comprising six independent passages each containing six dichotomous items was analysed with three Rasch models, namely, Andrich's (1978) rating scale model (RSM), Andrich's (1982) equidistant model and Masters' (1982) partial credit model (PCM). Results show that there is not much difference in the three models as far as model data fit, standard error of parameter estimates and discrimination are concerned. Nevertheless, noticeable differences were observed in the estimates of the difficulty parameters across the three models.

Key words: Rasch model, partial credit model, rating scale model, equidistant model, item-bundle approach

[1] *Correspondence concerning this article should be addressed to:* Purya Baghaei, PhD, English Department, Islamic Azad University, Ostad Yusofi St., 91886-Mashad, Iran; email: pbaghaei@mshdiau.ac.ir

Introduction

One of the assumptions of Rasch or IRT models in general is the local independence assumption. The items that are put to Rasch model analysis are required to be independent of each other. That is, a correct or wrong answer to one achievement test item should not lead to a correct or wrong answer to another item. This means that there should not be any correlation between two items after the effect of the underlying trait of testees is partialized. The items should only be correlated through the latent trait that the test is measuring (Lord & Novick, 1968). If there are significant correlations among the items after the contribution of the latent trait is partialized, then the items are locally dependent or there is a subsidiary dimension in the measurement which is not accounted for by the main Rasch model based dimension (Lee, 2004; Linacre, 1998). In other words, performance on the items then depends to some extent on a trait other than the Rasch model dimension; this means a violation of the assumptions of local independence and unidimensionality.

If the assumption of local item independence is violated, any statistical analysis based on it would be misleading. Specifically, estimates of the latent variables and item parameters will generally be biased because of model misspecification, which in turn leads to incorrect decisions on subsequent statistical analysis, such as testing group differences and correlations between latent variables. In addition, it is not clear what constructs the item responses reflect, and consequently, it is not clear how to combine those responses into a single test score, whether IRT is being used or not (Wang, Cheng & Wilson, 2005, p.6).

However, in practical testing contexts the local independence assumption gets violated very easily. This happens when several questions are based on a single prompt such as in reading comprehension tests, C-Tests (Grotjhan, 2002, 2006, 2010) or in cloze tests (Alderson, 1978; Oller, 1979). It is argued in the literature that if the local independence assumption does not hold, the local item dependence (LID) itself acts as a dimension. If the effect of LID is substantial it is difficult to say what dimension the main Rasch model dimension is. Even if the effect is small, the derived parameters will be contaminated, i.e., the parameters partially reflect the LID dimension to the extent that LID exists. In fact, LID is a form of the violation of unidimensionality principle. LID also results in artificially small standard errors of estimates. This could be a very severe problem in computerized adaptive testing where standard errors of measurement are the criteria for terminating the test. It can result in premature termination of the test (Zenisky, Hambleton & Sireci, 2003).

When a set of items are locally dependent there are two major approaches to address the problem. One is the item-bundle approach (Rosenbaum, 1988, Wainer & Kiely, 1987) in which the scores of dichotomous items which are based on the same stimulus are aggregated as ordered polychotomous super-items. That is, the set of items which are related to a common stimulus are considered as one polychotomous item to control the influence of local item dependence among items within each super-item. Polychotomous Rasch models or IRT models such as Andrich's rating scale model (1978), Andrich's equidis-

tant model (1982), Samejima's (1969) graded response model or Masters' (1982) partial credit model are then applied to analyse the super-items. The drawback of collapsing dichotomies into polytomies, however, means the loss of information.

The other approach is to model local dependency with the approach of Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Wang, 2007; Wang, Bradlow & Wainer, 2002. This is a 4-PL IRT model which adds a super-item parameter to the familiar 3-PL model. In this way the super-item effect can be separated from the test-takers' ability. The model can be applied to a mixture of independent items and super-items and provides a super-item parameter estimate for each super-item which is on the same scale as the ability parameter. The higher the super-item parameter the greater is the connectedness of the items which are in a super-item. This model is formally expressed as:

$$P(y_{ij}=1) = c_j + (1-c_j)\frac{\exp[a_j(\theta_i - \delta_j - \gamma_{ijk})]}{1+\exp[a_j(\theta_i - \delta_j - \gamma_{ijk})]}$$

Where γ_{ij} is the score for item *j* received by examinee *i*, a_j is the slope parameter, δ_j is the difficulty parameter, c_j is the guessing parameter, θ_i is the examinee's ability and γ_{ijk} is the super-item effect parameter showing this effect for examinee *i* in the super-item *k* to which item *j* belongs.

The purpose of the present paper is (1) to account for local dependency in a reading comprehension test using the item-bundle approach and compare the performance of three polychotomous Rasch models for this purpose and (2) to investigate the consequences of violating the assumptions of each of these models.

Polychotomous Rasch models

After the first formulation of the Rasch model (Rasch 1960/1980) which was designed for items that could be answered either correctly or wrongly, i.e., the so called dichotomous Rasch model, several other extensions for the model have been invented. The first of these models is probably Samejima's (1969) graded response model (Thissen & Steinberg, 1986). Samejima developed her model completely independent of Georg Rasch; but there are at least two approaches by Rasch himself in 1966 (cf. Kubinger, 1989) which predicts the probability of responses in each category *k* or above. The model is an extension of Birnbaum's item response theory to categorical data and unlike Rasch models accommodates a slope parameter.

$$P_{kni} = \frac{\exp[\alpha_i(\theta_n - \lambda_{ik})]}{1+\exp[\alpha_i(\theta_n - \lambda_{ik})]}$$

The model predicts that the probability person *n* reaches category *k* on item *i* depends on θ_n the ability of the person, λ_{ik} the value of *k*'th item boundary and the item's discrimination a_j.

Two other widely used polychotomous models which are implemented in most Rasch model software are Andrich's (1978) rating scale model (RSM) and Masters' (1982)

partial credit model (PCM). The formal expressions of Andrich's RSM and Masters' PCM are given below.

RSM (Andrich, 1978)

$$P_{xni} = \frac{\exp\left[-\sum_{j=0}^{x}\tau_j + x(\theta_n - \delta_i)\right]}{\sum_{k=0}^{m}\exp\left[-\sum_{j=0}^{k}\tau_j + k(\theta_n - \delta_i)\right]}$$

In this function θ_n and δ_i are the locations of person *n* and item *i* respectively, τ_j is the location of the *j*'th step in each item and *k* is categorie.

PCM (Masters, 1982)

$$P_{xni} = \frac{\exp\sum_{j=0}^{x}(\theta_n - \delta_{ij})}{\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\theta_n - \delta_{ij})}$$

x*=0, 1, ... *m

The model expresses that the probability of person *n* scoring *x* on the *m*-step item *i* is a function of person's location θ_n and the difficulties of the *m* steps. In this model the thresholds are combined with the item parameter estimates, i.e., $\delta_{ij} = \delta_i + \tau_j$.

Andrich advanced RSM for ordered item categories data. This model assumes equal category thresholds across the items. Step difficulties in ordered item categories are deemed to be governed by a predefined set of response categories which are repeated for all the questions. Since the same responses alternatives such as 'strongly disagree', 'disagree', 'undecided', 'agree' and 'strongly agree' are given for all the items it is assumed that step difficulties do not vary across the items. In other words, the distance between 'strongly agree' and 'agree' in all the items throughout the test is the same (Masters & Wright, 1984). That is, the increment in the level of the construct as a result of endorsing 'strongly agree' rather than 'agree' is equal for all the items. However, the model does not require that the distances between 'strongly disagree', 'disagree', 'undecided', 'agree' and 'strongly agree' be equal within a single question. The level of increment in the construct can be different when a respondent endorses 'strongly agree' rather than 'agree' compared to when he endorses 'undecided' rather than 'disagree'.

Then Andrich (1982) proposed a model called 'equidistant model'. This model assumes that the distances between the thresholds within the items are equal but not necessarily across the items. The model was especially suggested to account for local dependency in educational tests where several items are based on one prompt by forming super-items. The probability that *X* gets the value *x*, (*x*=0, 1, 2, ... *m*) given person parameter θ and item parameter δ is expressed as:

$$P\{X = x\,|,\theta,\delta,m\} = \frac{1}{\sum_{k=0}^{m}\exp(k_k + k(\theta-\delta)]}\exp[x(m-x)\theta + x(\theta-\delta)]$$

x*=0, 1, 2, ... *m

Where k_m = are category coefficients and are expressed in terms of *m* thresholds $\tau_1,\tau_2,...,\tau_m$.

Masters' PCM, on the other hand, is less restrictive than Andrich's RSM and equidistant model in that it does not require equal distances between the steps neither within items nor across items. Therefore, each item has a unique rating scale structure. That is, the distances between the steps can vary for all the items and within each single item and even the number of steps can vary. This property of PCM makes it the model of choice for analyzing educational tests where the assumption of equal step difficulties across items is very unrealistic.

Obviously, since we have fixed response categories in RSM and all the items share the same rating scale structure the number of response categories should also be fixed in all the items. This means that it is not possible to have some items with five response categories like 'strongly disagree', 'disagree', 'undecided', 'agree' and 'strongly agree' and some items with say, three response categories like 'agree', 'undecided' and 'disagree' in one instrument. However, as indicated PCM accommodates items with different number of response categories.

The assumption of equal distances between steps across the items, which is required by RSM, is certainly not met in the context of a reading test which has several different passages with different questions. Furthermore, the assumption of equal distances between the steps within each item, i.e., between each dichotomous question within a passage, which is the requirement of the equidistant model, cannot be met either in the context of a reading test.

These assumptions all suggest that Masters' partial credit model, which is the least restrictive model in terms of the distances between the steps within and across items, is the most appropriate model to analyse a reading comprehension test or any other educational test in which several items are based on a prompt. In the following section the consequences of using 'inappropriate' models for a reading comprehension test are investigated.

Method

A reading comprehension test comprising six passages was given to 160 Iranian undergraduate students of English. Six to eight items were based on each passage. In order to make the number of items on each passage equal the last one or two items of those passages with seven and eight items were not entered into the analysis. Therefore, the data-

set contained six passages each having six dichotomous items. The test was analysed three times with three polychotomous Rasch models: 1) Masters' partial credit model, 2) Andrich's rating scale model and 3) Andrich's equidistant model, considering each passage as a polychotomous item with seven categories. WINMIRA (von Davier, 2001a) programme was used for data analysis.

Results

Dichotomous analysis

Although all the cited models are members of the so called specific objective models (Rasch, 1960/1980; Fischer, 1974) it was decided not to follow the respective approach of model testing (cf. Kubinger, 2005) [you may see that this is true in Kubinger, 1989]; it was preferred here rather the universal approach of generalized linear models with statistics and psychometrics just to try for goodness of fit. This is due to the fact that within language educational test tradition the latter approach is common but the model testing approach according to specific objectivity is not. One may study this approach with respect to the used models in Kubinger (1989).

A preliminary dichotomous Rasch model analysis of the data, assuming local independence, indicates violation of the unidimensionality principal. Out of the 36 items nine have out of range mean square outfit values, the acceptable range being 0.7-1.3 (Bond & Fox, 2007). Principal component analysis of standardized residuals clearly indicates the existence of a secondary dimension. The variance explained by measures was equal to 41%. The unexplained variance in the first contrast was 5.3% which was equal to the strength of 3.2 items which is much larger than the minimum level of 1.5 (Smith, 2002). Rasch model separation reliability of the test was equal to 0.82. This reliability drops to 0.74 when the items are combined in super-items. This drop in reliability partly shows the effect of local dependence among the items which has resulted in spurious higher reliability for the 36-item test. One should bear in mind that the drop in the reliability is also partly due to the decrease in the number of items.

Examining the correlation of standardized item residuals, which is an indication of local dependence (Linacre, 2007) shows that eight pairs of items which are in the same super-items have residual correlations of 0.22 and above.

Polychotomous analyses

Table 1 shows super-items' statistics in the three polychotomous analyses. The columns 'Measure' show the difficulty location of each passage or testlet, 'SE' is the standard error of super-item difficulty estimates, 'Q-index' (Rost & von Davier, 1994) is an item fit statistic that can range between 0 and 1. Zero indicates perfect discrimination and 1 indicates perfect negative discrimination. Values close to .50 show random responses. Therefore, we expect values which are closer to 0. 'Zq' is the transformation of the 'Q-

index' which is approximately normally distributed and therefore the usual boundary of -2 to +2 can be applied to it. High positive values indicate unmodeled noise or underfit and low negative values indicate redundancy, local dependency or model overfit (von Davier, 2001b). As Table 1 shows there is not much difference between the three models as far as model data fit is concerned. Fit of the data to the model supports unidimensionality (Kubinger, 2005), therefore, the test is unidimensional no matter which Rasch model is used. The standard errors of the super-item estimates are almost equal in the three models indicating that the super-items have been calibrated with equal precision in the three models. Nevertheless, the difficulty estimates of the super-items vary greatly across the models, which is a cause for concern as one does not know which model shows the real difficulty estimates.

Table l:
Fit statistics, difficulty measures and standard errors

Super-item	Partial credit model				Rating scale model				Equidistant model			
	Measure	SE	Q-Index	Zq	Measure	SE	Q-Index	Zq	Measure	SE	Q-Index	Zq
1	-.30	.07	.16	.00	-.18	.07	.16	.24	-.17	.07	.17	.00
2	-.66	.08	.17	-.16	-.89	.08	.17	.08	-1.04	.08	.18	-.03
3	.35	.07	.16	-.03	.44	.07	.16	1.43	.58	.07	.16	-.18
4	.22	.07	.15	.03	.01	.07	.15	.04	.04	.07	.15	.00
5	-.01	.06	.12	-.17	.01	.06	.12	-1.38	.12	.06	.12	-.19
6	.40	.06	.14	.42	.15	.07	.14	-.83	.46	.06	.15	.43

Information criteria were used to compare different models. Table 2 shows that according to AIC partial credit model seems to be a better model since it has the smallest AIC value. However, since in the calculation of AIC sample size is not used this statistics is biased and BIC is a preferable statistic. According to BIC index RSM is the model of choice. The reason why RSM has been indicated as the best model by BIC index is the smaller number of parameters that are estimated in RSM, since the BIC penalizes more for over parameterization than AIC.

Table 2:
Information Criteria for the three models

Model	AIC	BIC
Partial credit model	3177	3291
Rating scale model	3226	3263
Equidistant model	3225	3265

Table 3 shows the mean, standard deviation and the Rasch model separation reliability (Linacre, 1997) of the test in the three analyses. The means of the sample in the three analyses are very close. The three analyses have equally widely distributed the persons and therefore, there is not any difference in the discrimination power of the three models. The last column of the table shows that the three models yield equal reliabilities.

Table 3:
Sample Statistics in the three analyses

Analysis	n	Mean	Standard Deviation	Reliability
Partial credit model	160	-.09	.70	.74
Rating scale model	160	-.10	.70	.74
Equidistant model	160	.06	.70	.74

Note. n = number of persons

Conclusions

In this study the performance of three different polychotomous Rasch models for analyzing a reading comprehension test comprising six super-items was compared. Since the assumptions of equal distances between the steps within and across the items cannot be met for a reading comprehension test, Masters' partial credit model which does not assume such equalities is usually recommended for super-item analysis. The results of the present study showed that the three models perform equally well as far as model data fit, reliability and discrimination are concerned. This means that the rating scale model and equidistant model are very robust to violations of equal distances between the steps and for practical measurement purposes all three models are appropriate.

However, there were substantial differences in super-item difficulty estimates across the three models. This raises the important question of which model depicts true item difficulty estimates. A replication of the study with simulated data where the location of the items is known is needed to find out which model performs best in recovering item estimates.

Furthermore, local dependency is something that is assumed to exist in the data analysed for this study. Further research needs be done by simulating datasets with varying degrees of local dependency among the items and then compare the performance of the models.

Nonetheless, one can be certain that the assumption of equal distances between the steps across and within the items, which is required by rating scale model and equidistant model respectively, has been violated in these data. The results showed that RSM and equidistant model performed as well as PCM in terms of fit, reliability and precision of estimates although their assumptions were violated. But the issue of unequal parameter estimates across the models is an open question which should be addressed with a simulation study.

Acknowledgment

This study was funded by Alexander von Humboldt Foundation in Germany. It was conducted when I was visiting Bamberg University as a Humboldt Fellow. I would like to thank Claus Carstensen professor of methodology at Bamberg University for providing excellent working conditions during my stay and many helpful discussions.

References

Alderson, J. C. (1978) *A study of the cloze procedure with native and non-native speakers of English*. Unpublished Ph.D dissertation, University of Edinburgh.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters, *Psychometrika, 47*(1), 105-113.

Bond, T. G., & Fox, C. M. (2007) (2nd ed.) *Applying the Rasch model: fundamental measurement in the human sciences*. Lawrence Erlbaum.

Bradlow, E. T., Wainer, H., & Wang, X., (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Fischer, G. H. (1974) *Einführung in die Theorie psychologischer Tests [Introduction to the theory of psychological tests]*. Bern: Huber.

Grotjahn, R. (ed.) (2002). *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen*. Vol. 4. Bochum: AKS-Verlag.

Grotjahn, R. (ed.) (2006). *Der C-Test: Theorie, Empirie, Anwendungen/The C-test: theory, empirical research, applications*. Frankfurt am Main: Peter Lang.

Grotjahn, R. (ed.) (2010). *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research.* Frankfurt am Main: Peter Lang.

Kubinger, K. D. (ed.) (2009). *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen [Modern psychometrics – A brief survey with recent contributions*]. Munich, Germany: Psychologie Verlags Union.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: some critical suggestions on traditional approaches. *International Journal of Testing, 5*(4), 377-394.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74-100.

Linacre, J. M. (1998). Structure in Rasch residuals: why principal component analysis? *Rasch measurement transactions*, *21*(2), 636. Available: http://www.rasch.org/rmt/rmt122m.htm.

Linacre, J. M. (1997). KR-20 or Rasch Reliability: Which Tells the "Truth"? *Rasch Measurement Transactions, 11*(3), 580-1. Available: http://www.rasch.org/rmt/rmt113l.htm.

Linacre, J. M. (2007). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago, IL: winsteps.com.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of Rasch models. *Psychometrika, 49*(4), 529-544.

Oller, J. W. Jr. (1979). *Language tests at school*. London: Longman.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The university of Chicago Press, 1980).

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349-359.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18*(2), 171-182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika,* Monograph Supplement No. 17.

Smith, E. V. Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205-231.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

von Davier, M. (2001a). *WINMIRA* [Computer Sofrware]. Groningen, the Netherlands: ASC-Assessment Systems Corporation. USA and Science Plus Group.

von Davier, M. (2001b). *WINMIRA user manual*. Groningen, the Netherlands: ASC-Assessment Systems Corporation. USA and Science Plus Group.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet Response Theory and its applications.* New York: Cambridge University Press.

Wang, W., Cheng, Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65*(1), 5-27.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26,* 109-128.

Zenisky, A. L., Hambelton, R. K., & Sireci, S. G. (2003). Effects of local item dependence on the validity of IRT item, test and ability statistics. *Association of American Medical Colleges (AAMC).* Available: http://www.aamc.org/students/mcat/research/monograph5.pdf.