Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 5, March 2013 ISSN 1531-7714

Fitting the Mixed Rasch Model to a Reading Comprehension Test: Identifying Reader Types

Purya Baghaei, *Islamic Azad University*
Claus H. Carstensen, *University of Bamberg*

Standard unidimensional Rasch models assume that persons with the same ability parameters are comparable. That is, the same interpretation applies to persons with identical ability estimates as regards the underlying mental processes triggered by the test. However, research in cognitive psychology shows that persons at the same trait level may employ different strategies to arrive at the solutions. This is a major threat to the construct validity of a test since the construct representation of the test changes for different classes of respondents. In this study a reading comprehension test composed of 20 multiple-choice items is analysed with mixed Rasch model. Findings show that a two-class model fits the data best. After investigating class specific item profiles the implications of the study for test validation along with the contribution of the research to our understanding of reading processes are discussed.

Unidimensional Rasch models (Rasch, 1960/1980) assume that examinees with the same location on the ability continuum have similar interpretations as regards their abilities, skills and mental processes. However, research in cognitive psychology and language testing has showed that individuals at the same trait level, i.e., the same measures on the construct, may use totally different strategies and mechanisms for arriving at the solutions (Sigott, 2004, Sternberg, 1985). This implies that the construct and its substantive meaning changes for different examinees depending on the types of strategies they use for solving the items, which is a major threat to construct validity.

If we cannot have uniform interpretations regarding the mechanisms and strategies that examinees of similar trait level get involved in then comparison of examinees on one ability continuum is not possible. In other words, the instrument measures different underlying constructs for different subpopulations or classes of examinees and it is not justifiable to compare examinees on a common ability continuum. Classes are defined in terms of the set of processes, strategies and mechanisms that examinees use to solve the items. (Embretson, 2007; Glück & Spiel, 2007; Rost, Carstensen, & von Davier, 1997).

A test is unidimensional, i.e., measures the same underlying construct for everybody if the item difficulty order is stable for different subclasses of examinees. Constant item difficulty order indicates that performance on the test requires the same skills, knowledge and strategies for all examinees (Andrich, 1988; Linacre, 1996; Perline, Wright, & Wainer, 1979; Rasch, 1960). Different order of item difficulties shows that different mechanism and skills have been employed to solve the items. Therefore, the nature of the construct depends on the class to which an examinee belongs. In this case the correlation of the test with external criteria may also change, i.e., class membership acts a moderator variable which is further evidence of the change of the construct (Embretson, 2007). It is important to note that under the Rasch model, not only the order of items should remain constant across subpopulations but their estimated difficulty parameters and the distances among them should also remain invariant within

modeled error (Andersen, 1973; Andrich, 1988; Fischer, 1974; Rasch, 1960: Wright & Stone, 1979).

Another perspective on item difficulty invariance is investigating Differential Item Functioning (DIF). That is, checking the invariance of item parameters across known subpopulations. Changes in item parameters across subpopulations indicate changes in the underlying cognitive processes employed by the test-takers who belong to subpopulations. Andrich (1988) states that when DIF exists one cannot compare the means of the two groups; the differences between the groups are not differences in degrees but differences in kind.

Mixed Rasch model

Mixed Rasch model (MRM) or mixture distribution Rasch model (Rost, 1990) identifies latent classes of persons for whom the Rasch model holds separately. MRM is a combination of Rasch model and latent class analysis. The idea is that the Rasch model holds for classes of persons within a population with differing difficulty order for items in the latent classes. The model is a unidimensional model, however, the intended dimension changes across the classes. Under the standard unidimensional Rasch model item difficulty estimates should remain constant for different groups of people. MRM can account for data when difficulty patterns of items consistently differ in classes of population. MRM allows item parameters to vary across classes of population, i.e., when the unidimensional RM does not fit for the entire population (Rost, 1990; Rost & von Davier, 1995; Yamamoto, 1987).

Mixed Rasch model is a variant of the unidimensional Rasch model in which some item and population homogeneity assumptions are relaxed. This variant is still a Rasch model because each subset of population which is identified with the mixed RM can be scaled separately with a unidimensional RM (Rost, Carstensen, & von Davier 1997). This is desirable in situations where the heterogeneity of population is unavoidable. Instead of rejecting the entire dataset as Rasch unscalable we can fit a mixed RM and study different cognitive processes for latent classes of population (Rost, 1990).

In MRM the probability of a correct response to an item is a function of both the ability of the person which is a continuous variable and the grouping of the person which is a categorical variable, i.e., what type the person is or what set of strategies s/he uses. The role of mixed RM is to identify subclasses of population in which the assumptions of RM hold. The dichotomous form of mixed Rasch model is formally expressed as

$$P_{ni}(\mathrm{x}=1\,|\,\theta_n)=\sum_{g=1}^{G}\pi_g\frac{\exp(\theta_n-\beta_{ig})}{1+\exp(\theta_n-\beta_{ig})}$$

where P_{ni} is the response probability of person n to item i, π_g is the class size parameter or "mixing proportion", θ_n is the ability of person n and β_{ig} is the difficulty estimate of item i in latent class g (Rost, 1990). If there are two latent classes 1 and 2 with mixing proportions of say, .60 and .40, respectively then the item response function would be:

$$P_i(\mathrm{x}=1\,|\,\theta_n)=.60\frac{\exp(\theta_n-\beta_{i1})}{1+\exp(\theta_n-\beta_{i1})}+.40\frac{\exp(\theta_n-\beta_{i2})}{1+\exp(\theta_n-\beta_{i2})}$$

When there is only one latent class the mixed Rasch model is equivalent to the standard Rasch model. Item and person parameters in MRM are estimated separately for each latent class g, therefore, the estimated parameters are conditional on latent class g. The probability of belonging to each of the latent classes can be estimated for each examinee based on her response patterns. These probabilities add up to one for each examinee. The latent class which has the highest probability for an individual is the class individuals are assigned to for further analyses. Class membership is a categorical variable and its relationship with other criteria such as gender, age, proficiency, etc. can be investigated to study the nature of the classes and the essence of the qualitative differences among them.

Rost, Häussler, and Hoffmann (1989) analysed a physics test with 10 items both with standard one-class Rasch model and mixed Rasch model. Results showed that one-class Rasch model did not fit the data. So they shifted to mixed Rasch model. The mixed Rasch model identified two classes of people for whom the Rasch model held. For one class Items 1-5 were easy and for the other class Items 6-10 were easy. Detailed examination of item contents showed

that the first five items were practical knowledge items and the second five items asked about more theoretical issues in physics. This was interpreted by Rost et al. (1989) as having two distinct classes of people in the population, namely, practically oriented examinees and theoretically oriented ones.

MRM has been used in personality testing to identify latent classes differing in the use of response scale. For example, Rost, Carstensen and von Davier (1997), analysed personality scales with MRM and showed that there were two latent classes: one with a tendency to endorse extreme ratings and the other moderate ratings. In another study using mixed IRT model Maij-de Meij, Kelderman, and van der Flier (2008) demonstrated that the identified latent classes differed in terms of tendency to socially desirable responding. They also showed that the middle category of "?" in a 3-point response scale was used differently by respondents in different classes. Along the same lines, Smith, Ying, and Brown (2012), using MRM, demonstrated that the middle category of "Neutral" in a 5-point scale did not function as expected.

Hong (2007) analysed a depression scale given to a sample of nonclinical Korean university students and identified three classes or types of depressed behavior. Zickar, Gibby, and Robie (2004) applied MRM to identify fakers in personality tests. They managed to identify groups with different degrees of faking from honest respondents to extreme fakers. Other researchers have employed MRM to identify solution strategies, (Mislevy & Verhelst, 1990), study the effects of test speededness (Bolt, Cohen, Wollack, 2002), and set proficiency standards (Jiao, Lissitz, Macready, Wang, & Liang, 2012).

Investigating the invariance of item parameters across subclasses of examinees is a well-documented way of checking model data fit in unidimensional Rasch models and a test of unidimensionality assumption (Andersen, 1973; Kubinger, 2005; Wright & Stone, 1979). However, the requirement of invariant item parameters gets violated quite often. It is very common to check the invariance of item parameter estimates across different subclasses of examinees divided by gender, ethnicity, score, etc. The strength of this method, however, depends on finding an appropriate partition of examinees. An optimal partitioning criterion is not necessarily scores or gender. MRM helps identify the partitions of population across which the item parameter estimates differ most and can direct test developers to more powerful partitioning of population for checking item invariance to investigate model-data fit in unidimensional Rasch models (Rost, 1990). Different item parameters can be due to poor item construction, employing of different strategies for solving items by individuals belonging to different classes, or different cognitive styles of individuals across subpopulations (Rost, 1990).

Mixed Rasch models can detect examinee heterogeneity and the associated item profiles, the latent score distribution and the size of latent classes. It can also help to test the fit of unidimensional Rasch models (Rost, 1990). Rost and von Davier (1995) argue that checking the fit of unidimensional Rasch models is one of the peripheral applications of MRM. Its main application is detecting qualitative differences among examinees and finding out how individuals perform the test tasks. The underlying abilities, motives and multitude of skills which are employed by respondents to complete tasks in educational and psychological tests might be more complex than those hypothesized by the instrument developer. Mixed Rasch modeling helps identify and detect these skills when simple unidimensional models are not sufficient to model the interactions between persons and items.

Mixture distribution models are a promising way of taking qualitative individual differences into account without losing the strong but necessary assumptions of the basic models-those models that hold for the unmixed data (i.e., the Rasch model in the present case). The Rasch model calls for this extension because its theoretical strength is better used for identifying groups of examinees who are really scalable, than for refuting the unidimensionality assumption for the entire population, and then moving on to a weaker model. Future applications of the model will show whether this promise is warranted (Rost, 1990, p. 281).

Having cited Rost (1990) and Rost and von Davier (1995) on the applications of MRM for testing

the fit of unidimensional models we also need to mention that Draxler (2002, cited in Kubinger, 2005) demonstrated that Andersen's likelihood ratio test leads too often to the rejection of the Rasch model when partitioning is done on the basis of MRM. This happens even when data are simulated to fit the Rasch model. Thus Kubinger (2005) refers to testing the fit of the Rasch model with Andersen's likelihood ratio test with MRM-based partitioning criterion as "artificial model check" and does not recommend it.

Another application of MRM is in investigating construct validity by testing the assumption of unidimensionality (von Davier & Yamamoto, 2007). In construct validation studies it is extremely important to demonstrate that only one ability or skill accounts for the observed response variances. Rosenbaum (1989, cited in Kreiner & Christensen, 2007) argues that unidimensionality, monotonicity, local independence, and the absence of DIF are the requirements of criterion related construct validity. Absence of DIF "requires the relation between the latent trait and the items to be the same in any subpopulation" (Kreiner & Christensen, 2007, p. 332). While DIF analyses use a priori known subpopulations, mixed Rasch model has a priori unknown grouping. The analyst does not need to have a known classification such as gender or language background to test the invariance of item parameter estimates. The model is capable of identifying classes of respondents across whom DIF exists.

In multidimensional Rasch and IRT models the probability of a correct response to an item depends on more than one person ability dimension. In MRM the probability of a correct response to an item depends on one person ability dimension and a categorical variable, namely, the latent class to which the person belongs. Dependence on a categorical variable can be a source of multidimensionality "since the different outcomes of the mixing variable moderate the conditional response variable in addition to one or more continuous person variables" (von Davier & Yamamoto, 2007, p. 114). Fit of a two-class model to data is an instance of multidimensionality and evidence that construct validity is compromised.

In this study we aim to apply MRM to an educational test to identify latent classes of examinees, if any. The primary objective is to demonstrate the applications of MRM in test validation via identification of latent classes who are qualitatively different. Existence of latent classes poses problems in test score interpretation and generalization and therefore is a threat to test validity. Such findings can also help substantive psychologists and educationalists in developing and revising construct theories.

Method

Participants and instrument

Participants were 1024 Iranian 3rd grade junior high school students aged 14 (605 girls, 419 boys) in Mashhad. The test was the reading comprehension section of their final achievement test in English as a Foreign Language in spring 2010. District-wide achievement tests are administered at the 3rd grade of junior high school in all subjects for decisions to be made on the type of high school candidates can attend. The reading comprehension section of the test comprised 20 dichotomously scored multiple-choice items.

Data analysis

The 20 reading items were subjected to mixed Rasch model analysis using WINMIRA (von Davier, 2001a). WINMIRA directly reads data from SPSS; analyses can be run by pointing and clicking and there is no need for programming or syntax writing. Screen shots are provided in the Appendix. WINMIRA can be purchased from http://www.von-davier.com/ or international distributers such as Kagi or Assessment Systems Corporation. A restricted demo version can also be downloaded for free.

As the number of classes is not a parameter to be estimated several alternative models with different number of classes are fitted and then the best fitting parsimonious model is selected. Since the models are not nested the deviance statistic (-2 log-likelihood) cannot be used for model selection. Competing models are selected by means of information criteria such as Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Consistent Akaike Information

Criterion (CAIC) (Bozdogan, 1987). These criteria are computed as follows:

$$AIC = -2 \log L + 2p$$

$$BIC = -2 \log L + p (\log N)$$

$$CAIC = -2 \log L + (\log N + 1) p$$

where L is the likelihood, N is sample size and p is the number of estimated parameters in the model. In WINMIRA the information indices are computed using the conditional likelihood (von Davier, 2001b). The number of parameters is included in the model as a penalty term for over parameterization (Kang & Cohen, 2007). *BIC* and CAIC were suggested because *AIC* is not asymptotically consistent as sample size is not used in its calculation. BIC and CAIC penalize more for the number of parameters and therefore chooses the models with fewer parameters compared to AIC. Models which have smaller information criteria are selected. According to Lin and Dayton (1997) the results of the statistics do not necessarily agree.

Results

Number of latent classes

To determine the appropriate number of latent classes competing models with one, two, three, and four latent classes were fitted to the data. Table 1 reports the BIC and CAIC for the four models. We employed BIC and CAIC because they are recommended more frequently in the literature (Read & Cressie, 1988; Rost, 1996). Table 1 shows that the two-class model has the smallest BIC and CAIC indices. Therefore, the model with two latent classes with sizes .50 and .49 was selected. The fact that a two-class model fits better than a standard one-class mode and the difficulty order of the items change across classes is evidence that the standard one-class Rasch model does not fit.

Class-specific item parameters

As mentioned before item parameters are conditional on latent classes in MRM. Comparing item parameters across classes is a particularly informative procedure about the qualitative differences among the latent classes. In such comparisons the focus is on the items which are relatively more difficult in one class but easier in other classes. Investigation of class-specific item parameters leads to understanding of the differences in the cognitive strategies and mechanisms involved in test performance. The difficulty order of the items on the Wright map (Wilson, 2005) in the two classes is shown in Figure 1.

Table 1. Information criteria values for the mixed Rasch model with different number of classes

Model	BIC	CAIC
One-class	24697	24718
Two-class	24351	24394
Three-class	24356	24421
Four-class	24437	24524

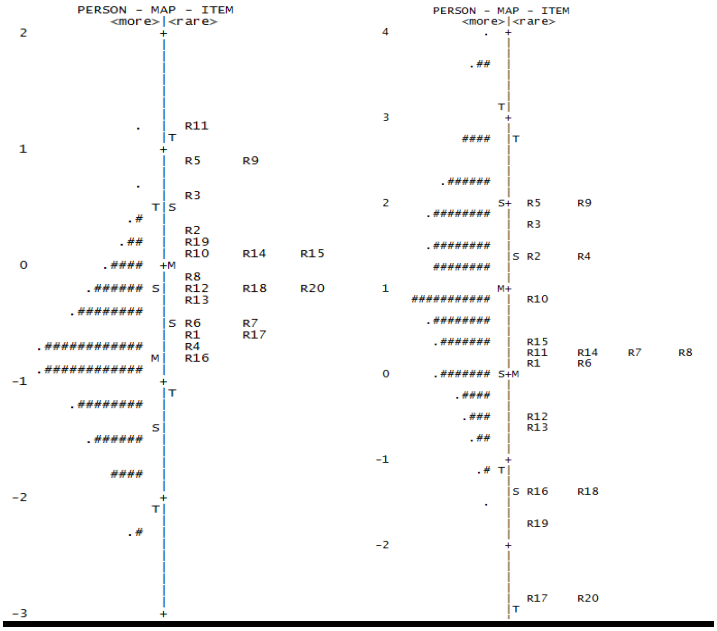


Figure 1. Item difficulty hierarchy and person distribution in Class 1 (left) and Class 2 (right)

R1 to R20 stand for the 20 reading comprehension items of the test; '#' and '.' represent persons. Items on the top are more difficult and those falling towards the bottom are easier. The difficulty estimate of each item can be read from numbers printed vertically on the left of the graphs.

Figure 2 shows the class-specific item parameters for the two latent classes in this study. The horizontal

axis shows the 20 items and the vertical axis shows the logit difficulty scale. Points lower on the scale indicate that the item was relatively easier for the class and points higher on the scale indicate that the item was harder for the class.

Figure 2 displays that the two classes have different difficulty parameters. The patterns for Class 1 and 2 show that there are certain items on which the two classes seem to diverge. In general both classes have found the second part of the test easier. There is one interesting difference between the classes. Class 1 has found the first 10 items easier than Class 2 and Class 2 has found the second 10 item easier than Class 1. The lines swap positions exactly after the 10th item, except Items 14 and 15 which do not neatly fall into this pattern. The item parameter differences for these two items across classes are very close, though.

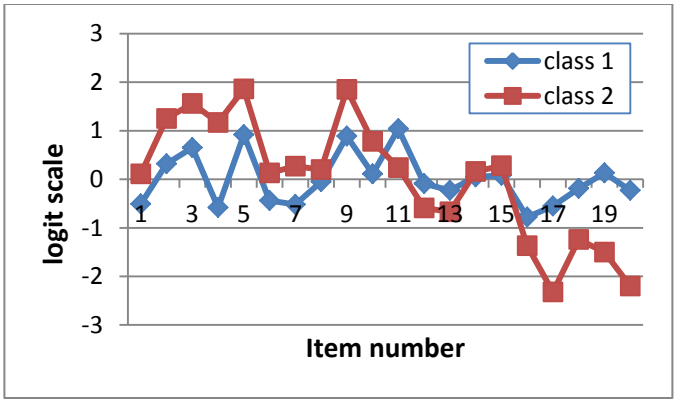


Figure 2. Class specific item parameter profiles

Unfortunately, the researchers did not have access to the items. But specifications of the test were available. According to the specification the first 10 items were based on 10 short passages (20-30 words), one question for each passage and the second 10 items were based on two long passages (400-500 words), five questions for each passage. Figure 2 shows that items based on long texts are easier for all examinees. This means that processing longer texts is easier than processing short texts perhaps because there are more contextual clues in longer texts and as a result there is more contextual support for the readers. Figure 2 indicates that Class 1 is more proficient in reading short texts and Class 2 is more proficient in reading long texts.

The relationship between latent classes and proficiency

To further investigate the latent classes, each individual was assigned to the latent class s/he belonged with the highest probability and the means of the two latent classes on the reading test were compared using an independent samples t-test. Results showed that mean of Class 2 (M=.90, SD=.95) on the reading test was significantly higher than the mean of Class 1 (M=-.90, SD=.86), t (1022) =-31.72, p=0.00 (two-tailed), with effect size of .49 (eta squared) indicating that 49% of the variance in the reading measures was accounted for by latent classes.

Table 2. Item statistics in the two classes

Item	Class 1			Class 2		
	Esti-mate	Err	Q-index (ZQ)	Esti-mate	Err	Q-index (ZQ)
1	-.51	.09	20 (-.41)	.11	.10	.18 (-1.08)
2	.32	.10	.24 (.70)	1.25	.10	.19 (-.69)
3	.65	.11	.22 (.03)	1.56	.10	.20 (-.71)
4	-.58	.09	.29 (1.76)	1.17	.09	.31 (2.08)*
5	.92	.12	.25 (.97)	1.86	.10	.23 (.09)
6	-.44	.09	.18 (-1.03)	.13	.10	.23 (.19)
7	-.52	.09	.21 (-.42)	.27	.10	.28 (1.31)
8	-.05	.09	.25 (.46)	.20	.10	.19 (-.90)
9	.89	.12	.23 (.57)	1.85	.10	.21 (-.38)
10	.11	.10	.24 (.16)	.78	.09	.21 (-.44)
11	1.04	.12	.25 (.40)	.24	.10	.22 (-.36)
12	-.09	.09	.19 (-.70)	-.59	.11	.22 (-.01)
13	-.23	.09	.20 (-.62)	-.67	.11	.13 (-1.92)
14	.05	.10	.24 (.17)	.16	.10	.18 (-.09)
15	.08	.10	.24 (.11)	.28	.10	.26 (.79)
16	-.78	.09	.21 (-.56)	-1.37	.14	.35 (1.39)
17	-.55	.09	.19 (-1.16)	-2.32	.20	.27 (-.05)
18	-.19	.09	.24 (-.07)	-1.24	.13	.29 (.48)
19	.13	.10	.24 (-.25)	-1.50	.15	.36 (1.03)
20	-.23	.09	.27 (.79	-2.20	.19	.36 (.43)

Item fit for each class

The fit of the 20 reading items was assessed within each class using the Q index (Rost & von Davier, 1994) implemented in WINMIRA. The Q index provides information about the relationship of items to the latent trait. The Q index "is based on the log−likelihood of the observed item-pattern….The fit of an item i is evaluated with regard to the conditional

probability of its observed item response vector" (von Davier, 2001b, p. 76). The *Q* index varies between 0 and 1, where 0 indicates perfect fit and 1 indicates perfect misfit or negative discrimination. A *Q* index of .50 indicates no relation of the item to the trait or random response behaviour. The standardized form of *Q* index, *ZQ*, with zero mean and variance of unity which can be assumed to be asymptotically normal is also available. The familiar ±1.96 boundary of a 95% confidence interval can be applied to standardized *Q* index. *ZQ* indices show that all the items fit well in both classes except Item 4 which misfits in Class 2.

Discussion and Conclusion

A two-class Mixed Rasch model with sizes .505 and .494 proved to fit better to the data than a standard one-class model for a reading comprehension test composed of items based on short and long passages. Class 1 was more proficient in short text items and Class 2 was more proficient in long text items. The latent classes differed with respect to reading competence, with Class 2 having a significantly higher reading mean. Item fit assessed by *Q* index showed that the items fit well within the classes except one item which had poor fit in Class 2. The item profiles for the classes showed some significant differences in item parameters across classes for 17 out of the 20 items. This descriptive analysis is further evidence showing that the one-class model does not fit the data.

When a standard one-class model does not hold the major concern is the comparability of person measures across the latent classes. Different item parameters across latent classes imply that the construct assessed is different across the two classes. Therefore, all the concerns and ramifications when DIF occurs across reference and focal groups in standard one-class models apply here. Person measures within different classes need to be transformed onto the same scale so that we can compare test-takers across classes. Rost, Carstensen and von Davier (1997) state that if item parameters are substantially different across latent classes the test measures different traits for the two classes and person parameters cannot be compared across classes. However, if the item parameters are close the same trait is measured in both classes.

To solve the psychometric problem of scoring and score interpretation across classes Embretson (2007) suggests that we can include both ability estimates and class membership in interpretation and use of test scores. Although this method is practical for test developers and psychometricians, it is complicated to explain to examinees, and other non-specialist test-users. It is rather awkward to tell examinees that their scores on the same test has different meanings and are indicants of different abilities. Embretson's second suggestion is to make all members of different classes use the same strategies by teaching and intervening through test preparation courses, i.e., removing sources of class distinction.

Another solution is to identify items which function the same way across classes and impose equality constraints on these items across latent classes (Maij-de Meij, Kelderman, & van der Flier, 2008; Kelderman & Macready, 1990; von Davier & Yamamoto, 2004).

Mixed RM is a very valuable model to study the strategies test-takers employ to solve test-tasks, which has been the focus of psychometric research for several decades (Mislevy & Huang, 2007). This is in line with the concept of construct validity. Validity according to Messick (1989) is not just prediction of some behavior but explanation of the strategies and processes that take place in the mind of respondents.

Figure 2 demonstrated that items based on long and short passages are equally difficult for Class 1 readers while items based on long passages are substantially easier for Class 2 readers. It also showed that Class 1 readers are better than Class 2 readers at short-context items and Class 2 readers are better at long text items. Therefore, one can conclude that there must be two subtypes of reading: Class 1 readers are 'short text processors' and Class 2 readers are 'long text processors'. Results also indicated that the two latent classes were different with respect to reading proficiency with Class 2 having a significantly higher reading mean. Class 1 readers apply their 'short text processing skills' to long text items, which require 'long text processing skills' to get solved. It seems that the application of short text processing skills to long-context items is not very helpful. Class 2 readers

apply their long text processing skills to short text items. It appears that the application of long text processing skills to short-context items works to a certain degree otherwise Class 2 readers would not have had a significantly higher reading mean.

What is evident from these findings is that short and long text processing skills in reading in a foreign language do not develop linearly as a result of increased reading ability. That is, long text and short text processing strategies are two distinct skills which develop independently. One cannot argue that learners who possess long text skills have already mastered short text skills. If this was the case then short text items would have been easier for Class 2 readers who are more proficient readers. Mislevy and Huang (2007) state that the reason why examinees belong to different latent classes could be different educational systems and curricula or application of different strategies for responding to tasks. Substantive examination of item contents can shed light on the qualitative differences among the examinees. Understanding reduced context texts such as signs, notes, and newspaper advertisements is not necessarily easier for more proficient readers. In fact, understanding reduced context texts could be extremely difficult if reading courses have not provided enough training and practice on them.

This implies that short and long text processing skills do not stand on a reading dimension but form taxa. In other words, as far as short and long text processing skills in reading in English as a foreign language are concerned reading ability is taxonomical and not dimensional.

Another major finding of the study which ensues from the application of MRM is that texts with different lengths have different cognitive demands which in turn have an impact on the internal validity of the test in terms of its fit to the Rasch measurement model. Short text items are included in reading comprehension tests mainly because of time constraints. The other reason for having short texts in reading comprehension tests is that understanding context-reduced texts such as classified ads, signs and notices is a common practice in real life reading. Therefore, to measure candidates' abilities to process short texts such items are included in the test. The problem which arises here is that text length can potentially affect the cognitive processes which are triggered in candidates' minds in terms of knowledge structures, mechanisms and strategies. When the cognitive demands of short and long texts are very dissimilar then the validity of the test is questioned. We also provided empirical evidence for the existence of two categorically distinct subtypes of reading or levels of understanding, i.e., short texts and long text skills.

Mixed RM has potential applications in developmental psychology. A developmental psychologist can investigate if there are different types of learners with different patterns of learning and if those learner types can be associated with external factors such as age, sex, motivation, first language, etc. "…differences in item difficulty patterns may be more than just "noise" that needs to be removed in test development. They may reflect interesting processes of change that can contribute to our understanding of development" (Glück & Spiel, 2007, p. 292). The application of mixed RM in this area can be both confirmatory and exploratory. That is, a researcher might have some idea about the factors that affect differential patterns of acquisition and then collect data accordingly to test her hypotheses. Or it can be totally exploratory, i.e., the researcher first determines the number of classes in the dataset and then tries to associate them with possible external factors to develop hypotheses regarding the possible determinants of different acquisition patterns.

MRM can also be applied in the investigations of the effects of strategy training on strategy use by examinees (Glück & Spiel, 2007). Consider a situation where we want to study whether teaching appropriate reading strategies affect strategy application of the learners. We need a pretest-posttest design with strategy training as a treatment. Suppose that a two-class model fits the pretest data which is an indication of heterogeneity in the strategy application of the examinees. If a one-class model fits the posttest data, it means that the training has been effective in aligning the learners reading strategy use. If say, a four-class solution fits the pretest data and a two-class solution fits the post-test data, we can have a similar conclusion.

A limitation should be pointed out in the current study. We did not have access to the actual reading items. Only the data along with the test specifications were provided for the researchers for secondary analyses. Detailed examination of the contents of items which had differing difficulty estimates across latent classes was not possible. Such examinations can provide deeper insight into the development and processings involved in reading comprehension. Prospective mixed Rasch model analyses of reading tests with well-designed items may ultimately answer questions about the nature of reading comprehension construct and subtypes of reading ability and the relation between these subtypes and other manifest criteria.

As concluding remarks, a limitation with the application of mixed RM should be mentioned. The mixed RM requires large sample sizes especially for the polytomous extension of the mixed RM (von Davier, & Rost 1995). Since, the number of parameters to be estimated increases in polytomous items and there are more than one class to be estimated the sample size required to do a unidimensional analysis should be multiplied by the number of classes in order to carry out a reasonably accurate mixed Rasch analysis. It is, of course possible to estimate the parameters with smaller samples but the standard error of the estimates will be high (von Davier & Yamamoto, 2007). The problem exacerbates when the number of items and categories to be estimated increase as well. von Davier (2002) has implemented bootstrap fit analyses in WINMIRA software that allows testing the stability of results in small samples (cited in Glück & Spiel, 2007).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andrich, D. (1988) *Rasch models for measurement*. Newbury Park, CA: Sage.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika, 52*, 345-370.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (pp. 235-253). New York: Springer Verlag.

Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

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

Glück, J., & Spiel, C. (2007). Studying development via Item Response Model: A wide range of potential uses. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 281-292). New York: Springer Verlag.

Hong, S. (2007). Mixed Rasch modeling of the self-rating depression scale. *Educational and Psychological Measurement, 67*, 280-299.

Jiao, H., Lissitz, R.W., Macready, G., Wang, S., & Liang, S. (2012). Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling, 53*, 499-522.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331-358.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307-327.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 329-346). New York: Springer Verlag.

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

Lin, T. H., & Dayton, M. C. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249-264.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.

Mislevy, R., & Huang, C.-W. (2007). Measurement models as narrative structures. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 15-35). New York: Springer Verlag.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195 -215.

Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237–255.

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).

Read, T.R.C., & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271 – 282.

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

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer, & I. W. Molanaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 257-268). New York: Springer Verlag.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Munster, Germany: Waxmann.

Rost, J. (1996). Logistic mixture models. In W. van der Linden, & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). Berlin: Springer Verlag.

Rost, J., Häußler, P., & Hoffmann, L. (1989). Long term effects of physics education in the Federal Republic of Germany. *International Journal of Science Education, 11*, 213-226.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt/am: Peter Lang.

Smith, E.V., Ying Y, & Brown, S.W. (2012). Using the mixed Rasch model to analyze data from the beliefs and attitudes about memory survey. *Journal of Applied Measurement*, *13*, 23-40.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

von Davier, & Yamamoto, K. (2007). Mixture-distribution and hybrid Rasch models. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99-115). New York: Springer Verlag.

von Davier, M., & Carstensen, C. H. (Eds.) (2007). *Multivariate and mixture distribution Rasch models: extensions and applications*. New York: Springer Verlag.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406.

von Davier, M. (2001a). WINMIRA [Computer Software]. 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.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molennar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371-379). New York: Springer Verlag.

Weir, C. J. (2005). *Language testing and validation*: An evidence-based approach. New York: Palgrave Macmillan.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

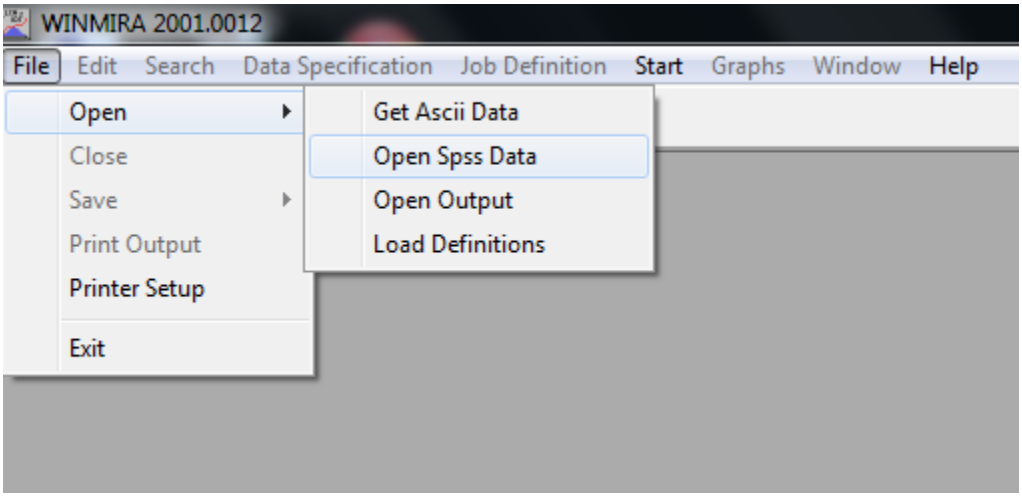
Yamamoto, K. Y. (1987). A model that combines IRT and latent class models. Unpublished doctoral dissertation, University of Illinois Urbana Champaign.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*, 168-190.

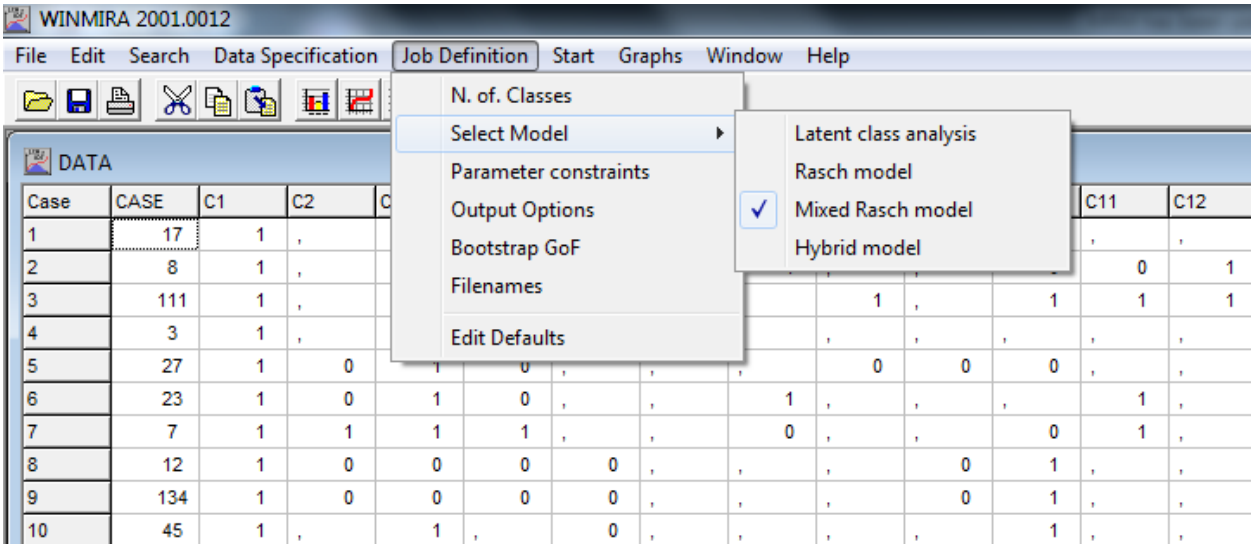
.

Appendix: Screen shots from WINMIRA

Selecting data from SPSS

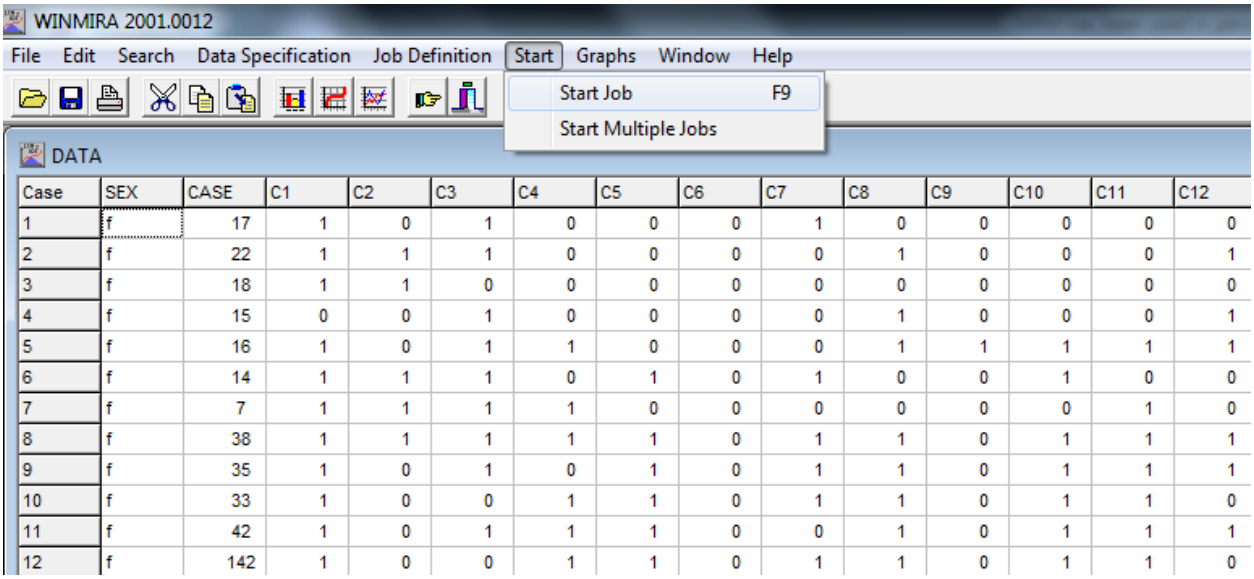


Choosing the model



Case	CASE	C1	C2	C								C11	C12
1	17	1	,									,	,
2	8	1	,									0	1
3	111	1	,						1	,	1	1	1
4	3	1	,						,	,	,	,	,
5	27	1	0	1	0	,	,	,	0	0	0	,	,
6	23	1	0	1	0	,	,	1	,	,	,	1	,
7	7	1	1	1	1	,	,	0	,	,	0	1	,
8	12	1	0	0	0	0	,	,	,	0	1	,	,
9	134	1	0	0	0	0	,	,	,	0	1	,	,
10	45	1	,	1	,	0	,	,	,	,	1	,	,

Running the analysis



Case	SEX	CASE	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
1	f	17	1	0	1	0	0	0	1	0	0	0	0	0
2	f	22	1	1	1	0	0	0	0	1	0	0	0	1
3	f	18	1	1	0	0	0	0	0	0	0	0	0	0
4	f	15	0	0	1	0	0	0	0	1	0	0	0	1
5	f	16	1	0	1	1	0	0	0	1	1	1	1	1
6	f	14	1	1	1	0	1	0	1	0	0	1	0	0
7	f	7	1	1	1	1	0	0	0	0	0	0	1	0
8	f	38	1	1	1	1	1	0	1	1	0	1	1	1
9	f	35	1	0	1	0	1	0	1	1	0	1	1	1
10	f	33	1	0	0	1	1	0	1	1	0	1	1	0
11	f	42	1	0	1	1	1	0	0	1	0	1	1	1
12	f	142	1	0	0	1	1	0	1	1	0	1	1	0

Acknowledgment:

Alexander von Humboldt Foundation in Germany is greatly acknowledged for financing this project by granting a fellowship to the first author.

Citation:

Baghaei, Purya & Carstensen, Claus H. (2013). Fitting the Mixed Rasch Model to a Reading Comprehension Test: Identifying Reader Types. *Practical Assessment, Research & Evaluation*, 18(5). Available online: http://pareonline.net/getvn.asp?v=18&n=5

Authors:

Purya Baghaei (corresponding author)
Department of English
Islamic Azad University, Mashhad Branch
91871-Mashhad, Iran.
Pbaghaei [at] iaum.ac.ir

Claus H. Carstensen
Department of Psychology
Methods of Educational Research
University of Bamberg
96045-Bamberg, Germany.
claus.carstensen [at] uni-bamberg.de