Trends in Life Sciences
An International Peer-reviewed Journal
www.sciencejournal.in

ASSESSING MODEL-DATA FIT FOR MIXED FORMAT TESTS

Shirin Akbari Varmazyar[1*], Mohammad Reza Falsafinejad[2], Ali Delavar[3], Noorali Faroukhi[4]
[1]PhD Student in Assessment and Measurement, Faculty of Psychology and Educational Science,Allameh Tabataba'i University, Tehran, Iran.
[2]Assistant Professor, Faculty of Psychology and Educational Science, Allameh Tabataba'i University, Tehran, Iran.
[3]Professor, Faculty of Psychology and Educational Science, Allameh Tabataba'i University, Tehran, Iran.
[4]Assistant Professor, Faculty of Psychology and Educational Science, Allameh Tabataba'i University, Tehran, Iran.
*Corresponding Author

ABSTRACT
The assessing of model-data fit is of great importance in psychometric. Aim: This study is aimed to assess the goodness of fit of methods and parameter estimation models regarding Item response theory (IRT) in mixed format tests. At first, the assessment methods of fitting model are discussed, then by real data, six combination model, three dichotomous models, and two calibration procedures are compared based on fitting assessment indices in model and item level. The study population is all the items of academic achievement of third of high school in academic year 2012-2013. The English language test items (3) are selected as a sample of academic achievement tests of third of high school. To determine the properties of items by real data, 2084 participants in English language test (3) are selected by random sampling method. To assess model-data fit in model level, Akaikes Information Criterion (AIC), Corrected AIC (AICC) Bayesian Information Criterion (BIC)and Root Mean Squared Error of Approximation(RMSEA) statistics are used and S-X^2 statistic is used in item level (Orlando and Thissen, 2000). Findings: The findings show that among the applied models in this study, the combination of three parameter logistic model(3PLM) with each of polytomous models, show low misfitted items compared to combining one-parameter logistic model(1PLM) with each of *Polytomous* models. Among polytomous models, Generalized partial credit model (GPCM) shows low misfitted items compared to Graded response model (GRM). Regarding the fitting in the findings model level, better fitting of combinational model 2PL/GPCM with data can be shown. Regarding the comparison of combined models and dichotomous models, one-parameter model showed the highest misfitted items and three-parameter model and combination of three-parameter model with generalized partial credit model (3PL/GPCM) showed lowest number of misfitted items. Regarding the fitting in model level, among dichotomous models, two-parameter logistic model(2PLM) and among combined models, combination of two-parameter logistic model and generalized partial credit model have better fitting with the data in terms of all the indices. Finally, a comparison is made between simultaneous and separate calibration. The results of comparison among the dichotomous models showed better performance of separate calibration in two, three-parameter models compared to simultaneous calibration. There was no difference in polytomous models between two methods. Although there are various methods to determine the model-data fit, there is no full agreement regarding optimal method. Judgment regarding item-response models requires various methods but the best judgment is dedicated to the researcher.

KEYWORDS: Model- data fit, mixed format test.

INTRODUCTION

The tests are used as criteria to choose academic levels and jobs (admittance in University), verification of mastery of specific skill (language tests), and evaluation of various educational courses. Normally, to improve quality, validity, reliability, and reduction of costs, a combination of multiple choice and constructed response is used. We can measure many educational objectives and an important part of textbook content by multiple choices. This property improves reliability and content validity of test. Constructed response tests provide some models to evaluate some specific skills as problem solving; these items can be used to improve construct validity (Thissen and Wainer, 2001). Multiple-choice items and constructed responses have some advantages and disadvantages and their combination can compensate this problem. Both items are used to evaluate a construct or an attribute in capability evaluation. The tests including both are called mixed format tests in psychometric literature. One of the basic challenges of using mixed format tests is regarding estimation of item and ability parameters. This makes some challenges not only in meaning of validity; it is problematic also to estimate ability parameters and item. Thus, parameters estimation and model selection are of great importance when mixed format tests are used. Calibration as mentioned estimation is the process of achieving item

parameters and ability in Item response theory. One of the important issues in calibration of the data with mixed format is selection of calibration procedure. For example, mixed format tests can be calibrated simultaneously or separately. In simultaneous calibration, all the items in a test are calibrated only a single run of an IRT estimation program. This provides direct comparison of items by putting them on a common scale. Another procedure is using separate calibration procedure. In separate calibration, various formats are calibrated separately. This procedure is suitable namely, when various formats of items are administered separately (Chon et al., 2010).

In the item-response theory applications, there are various models to describe examinees performance in the test. As both multiple-choice and constructed response are included in the single test for mixed format tests, a combination of dichotomous and polytomous models in item-response theory can be used. In multiple-choice items of dichotomous models as one, two, and three parameter models and for constructed response items of polytomous models as graded response model, generalized partial credit model and nominal response model can be applied. As there are various methods to estimate parameters in mixed format tests, by using model-data fit assessment methods, the best method is selected that the data of tests can adequately reflect the real conditions of test. The lack of selecting a correct model causes that the parameters estimation does not reflect the real conditions of examinees and items as methodology of model selection is of great importance. This methodology is not a novel concept and has been investigated within the context of multiple regressions, multilevel modeling, and structural equation modeling (Whittaker et al., 2012). The model selection receives much attention in item-response theory (e.g. Kang and Cohen, 2007; Kang et al., 2009). Selection of an incorrect model in the item-response theory leads not only to the different theoretical interpretation about data, but also inappropriate conclusions about other applications of item-response theory (Demars, 2010).it is recommended that the data fit the IRT model adequately prior to beginning the model comparison and selection process (De Ayala, 2009). Model-data fit is considered as a useful tool to select model. Normally, Model-data fit is considered in different levels in item-response theory. These levels include model level ,item level and person level. Although all aspects Model-data fit are important; in item-response theory, the assessment of fitting in item level is much important than fitting in test level as fitting in item indicates the fitting in test level and suitable fitting of all itemsof a test with model is sufficient condition for test fitting (Swaminathan et al., 2007).

To determine model –data fit, we should compare between a set of justified models of item-response theory to select a suitable model. There are two general approaches to select and compare the suitable model, the approach based on observed and expected scores distribution and approach of likelihood test more than other approaches in model selection (Swaminathan et al., 2007). The present study applies likelihood test-based approach. One of the methods of this approach is Likelihood ratio test (LRT). When item-response theory nested models are used, the selection model can be based on Likelihood ratio test (LRT) (Thissen et al., 1988). If the item-response theory models are not nested, LRT statistics is not suitable. The information-based criteria are used for non-nested models. These criteria are also included in likelihood test approach. The major advantage of these approaches is their application for different item-response theory models (nested and non-nested). One of the most famous methods is Akaikes Information Criterion (AIC). When this method is used to compare models, lower value of this index indicates better fitting of model with data. One criticism of the AIC is that it tends to over-fit models, meaning that tends to select more highly parameterized models (Shibata, 1976; cited in Whittaker et al., 2012). Corrected AIC (AICC) is designed to eliminate this problem. Although AIC and AICC are efficient model selection criteria, they are not consistent model selection criteria (Whittaker et al., 2012). Consistent selection criteria will select the correct/true model with probabilities close to or at 1.0 when the correct/true model is actually among the set of comparison model. One of the consistent information based criteria is Bayesian information criterion (BIC). Other suitable indices for comparing non-nested models are Root mean squared error of approximation. This index is denoted by RMSEA, is less affected by sample size, and is considered as a good and common size of fitness. The introduced indices are used to assess the fitness in model level. There are two general approaches to evaluating item fit, Heuristic or graphical approach and statistical approach. In heuristic approach, no statistical test is performed. In this approach, judgment about items fitting based on comparison between estimated item-response curves (IRC) with empirical IRC. In statistical approach, these curves are compared by statistical formula. For example, we can refer to Bock chi-square (1972) and Q1 statistics of Yen (1981). These indices are sensitive to sample size like many chi-square statistics and we cannot use them as tools for solid decision-making (Embretson and Reise, 2000). Based on the problems of traditional methods of item fitting assessment, alternative methods are presented and are used in various situations. Orlando and Thissen (2000) proposed fit statistics, S-X2 and S-G2based on joint likelihood distributions for each possible total score. They pointed out that grouping of

respondents in the traditional goodness of fit statistics is based on estimates that are model dependent, rather than on some observable statistics such as the number correct score. Alternatively, the S-X2 and S-G2statistics are not dependent on the model-based trait estimates, but on directly observable frequencies and thus are solely a function of data. In the studies, the findings show the suitability of these statistics as an alternative for traditional goodness of fit methods (e.g., Orlando and Thissen, 2000; Stone 2000; Glas and Suarez Falcon, 2003). Most of the studies in model selection are regarding merely dichotomous and polytomous models. A few studies have been conducted in mixed format tests (Chon et al., 2013; Chon et al., 2010).Although there are various methods to determine model fitting with data, there is no full agreement regarding optimal method. Despite wide use of mixed format tests in education and evaluation process of third of high school students as it is used as a criterion to select student in high education centers, no study has been conducted regarding systematic psychometric analysis of these tests. Based on the importance of study and priority of assessment of model fitting with data to model comparison and selection, as a basic issue for the researcher is which item-response models has acceptable fitting with the data of mixed format tests? On the other hand, this question is raised whether using various calibration methods (simultaneous and separate) makes any difference in model fitting with data in mixed format tests.

Study population

The study population of this study is all the items of academic achievement tests of third of high school in academic year 2012-2013. To determine the properties of items, the performance of students of third of high school of Khoramabad city participated in tests of Jun in academic year 2012-2013 is evaluated. Based on the statistics of education organization assessment center of Lorestan, 4105 people participated in the test, of which 2219 are girls and 1886 boys.

The sample and its selection method

In this study, the items of English language test (3) are selected as a sample of academic achievement tests of third of high school. To determine the properties of items by real data, 2084 participants in English language test (3) are selected by random sampling method. Of which, there are 1044 girls and 1040 boys. 1626samples learn in state schools and 458 samples in non-state schools.

Measures

The measure in this study is English language test of third of high school. This test is performed in Jun of 2013 in schools. This test is used due to its items format, the role of this test in determining average and access to high volume data. This test is composed of 76 items, of which there are 21 multiple-choice items, 16 matching items, 3 true- false items, 28 short- answer items and eight constructed response item. This test is performed in Jun 2013 in Iran. Based on initial analyses regarding reliability of this test, Cranach's alpha is 0.972 and standard error of measurement as 3.44.

Data analysis method

In this study, one, two, and three parameter models are used for dichotomous items, GPCM, and GRM models are used for polytomous items. By combining dichotomous and polytomous models, six various combinational models including1PL/GRM, 2PL/GRM, 3PL/GRM, 1PL/GPCM, 2PL/GPCM, 3PL/GPCM is used. In addition, two simultaneous and separate calibration methods are used to analyze subjects' responses. Totally, 12 situations of estimation are created (six combinational models*two calibration procedures) and based on fitness indices in model and item level are compared. To assess fitness in model level, AICC،AIC، BIC and RMSEA statistics are used and in item level, S-X^2 statistic is used (Orlando and Thissen, 2000). NOHARM4 (Fraser, 1988) andIRTPRO2.1 (Cai et al., 2011) software are used for collected data analysis. According to item-response theory, at first the $^{\text{unidimensionality}}$ and local independence assumptions are evaluated for test items analysis. Various methods are proposed to evaluate $^{\text{unidimensionality}}$ of test. To determine $^{\text{unidimensionality}}$, non-linear factor analysis method is used raised by McDonald (1967). This approach is implemented by NOHARM4 computer program. This program does not estimate guess parameters and acts as fixed with the entered values (Wright and Stone, 1979). As multiple choice and true - false items had no negative score in this study, it is predicted that examinees applied guessing to answer the items, thus guessing of multiple choices is 0.25, and false and true items 0.5.

IRTPRO2.1 software is used to estimate item parameters and ability. One of the capabilities of this software is implementation of combined models based on item-response theory, presenting some goodness of fit indices (-2log

likelihood, AIC, BIC and RMSEA). Based on uncertainty of item parameters and ability, Marginal maximum likelihood estimation method is used for simultaneous estimation of parameters.

RESULTS

The study findings are presented in two sections of a) investigation of the model assumptions and b) investigation of model fitting indices.

a. Investigation of model assumptions

The true selection of model can be promoted by investigation of fundamental assumptions of unidimensional models of item-response theory. A basic assumption in item-response theory as widely is unidimensionality of test. Before estimation of item parameters, this assumption is investigated. The results of non-linear factor analysis by NOHARM4 software are shown in Table 1. NOHARM software calculates the Root mean square of residuals (RMSR) and it is an index to fit the model. Indeed, RMSR is equal to Root mean square of the difference of observed covariance and predicted covariance. Thus, small values of RMSR indicate the fitness of model with data. A criterion to interpret RMSR is to compare it with four times inversed square of sample size (residuals standard errors) (McDonald, 1997). Another index to evaluate fitness model is Tanaka. McDonald (1997) proposes that value 0.90 shows acceptable fitness, 0.95 indicates good fitness of model with data. If this index is 1.0, it shows full fitness of model. As shown in Table 1, RMSR in unidimensional is very small (0.0076) close to zero. This value is smaller than four times inverse of square of sample size (as 0.088 in this study). In addition, Tanaka index is 0.99 and it indicates good fitness of model with data. Based on these indices and residuals matrix, there are no adequate evidences to reject unidimensionality of English language test (3).For exact investigation of issue and assurance of above conclusion; the data are analyzed based on 2-D solution again. The two-dimensional solution residuals are very small like uni-dimensional solution residuals. RMSR value of two-dimensional solution is 0.0063 and it shows little reduction (0.001) compared to unidimensional solution RMSR value. In addition, Tanaka index as 0.993 is increased little compared to this value in unidimensional solution (0.003). Thus, we can say the test is unidimensional and we can use item-response theory unidimensional models for its scaling. According to Hambleton, Swaminathan and Rogers (1991) when unidimensionality assumption is met the local independence assumption is also satisfied. Therefore, the clues for the unidimensionality were used also for the local item independence.

Table 1- Investigation of dimensionality of English language test by NOHARM explorative solution

	Unidimensional		Two-dimensional	
Index	RMSR	Tanaka index	RMSR	Tanaka index
Value	0.0075	0.9901	0.0062	0.9933

B. The assessment of model fitness indices

This section is based on study questions.

First question: Which of item-response models has acceptable fitness of mixed format tests? This question is raised in the form of two following sub-questions:

Question 1-1: Which of combined models have better fitness with data?

To respond the above question, six different combinational models calibrate the study data simultaneously. These models include 1PL/GRM, 2PL/GRM, 3PL/GRM, 1PL/GPCM, 2PL/GPCM, and 3PL/GPCM. For each combinational model, fitness statistical indices as $S\text{-}X^2$ are calculated for fitness test in item level and AIC, AICC, BIC and RMSEA statistics to fit in test level. Significance level of 0.05 is used for misfitting test of items. The significant fitness statistics shows that item parameters are different in the observed scores groups and the model is not good for data. Table 2 shows items without any fitting with model at significance level 0.05. It can be said, in this test, items 1-42, 49-70, 73-76 are dichotomous and 43-48, 71-72 items are polytomous.

Table 2- The results of $S\text{-}X^2$ statistic to assess fitting items with model

Models	**Misfit items**	**N**	**%**
1PL/ GPCM	3,6,7,10,17,18,19,20,22,23,25,26,27,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,45,49,50,52,53,54,57,	51	67

	58,59,60,61,63,64,65,66,67,68,69,70,72,73,74,75,76		
1PL/ GRM	3,6,7,10,17,18,19,20,22,23,25,26,27,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,44,45,49,50,52,53,54,57, 58,59,60,61,63,64,65,66,67,68,69,70,72,73,74,75,76	52	68
2PL/GRM	1,6,17,19,20,22,25,28,29,30,31,35,37,38,39,40,45,50,52,60,61,6 3,69,70,74,76	26	34
2PL/GPCM	1,17,19,20,22,25,28,29,30,31,35,37,38,39,40,50,52, 60,61,63,69,70,72,74,76	25	33
3PL/GPCM	17,20,28,29,35,39,40,46,49,50,53,61,63,68,70,72	16	21
3PL/GRM	17,20,28,29,35,39,40,44,45,46,49,61,63,67,68,70,72	17	22

As shown in Table 2 and Figure 1, compared to other applied models, combinational models 1PL/GRM and 1PL/GPCM are highest misfitted items and combinational models 3PL/GPCM and 3PL/GRM have lowest misfitted items.

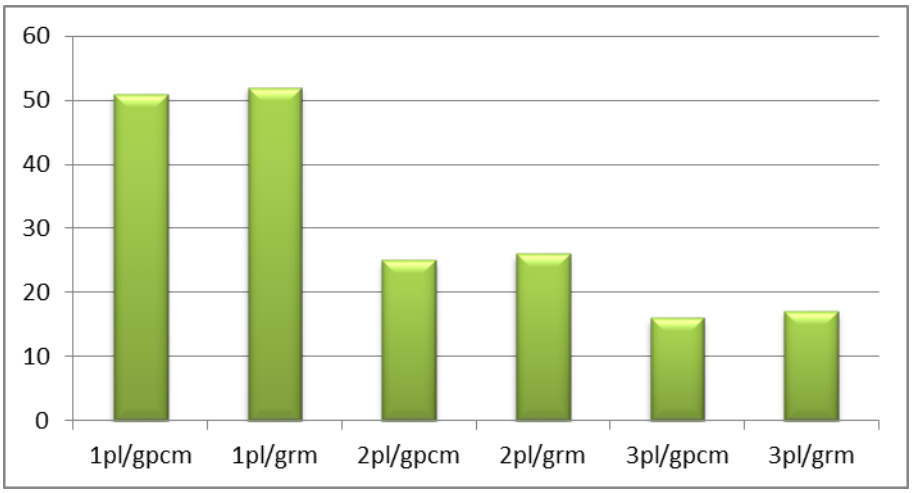

Figure 1- Frequency distribution of misfitted items based on investigated model

To test the fitness in model level, AICC, AIC, BIC and RMSEA statistics are used and the results are presented in Table 3.

Table 3- The results of fitness statistics in model level

Model / **Fitness statistics**	AIC	AICC	BIC	RMSEA
1PL/ GPCM	173343.45	173355.60	173958.43	0.05
1PL/ GRM	173345.60	173357.75	173960.59	0.05
2PL/GRM	167639	167671.67	168632	0.03
2PL/GPCM	167623.26	167655.93	168616.26	0.03
3PL/GPCM	167758.23	167823.24	169134.89	0.03
3PL/GRM	167701.99	167767	169078.65	0.03

Regarding AICC‹AIC and BIC indices, value about to zero indicates good fitness and among them, the smallest value indicates model with better fitness. RMSEA value as deviation test in degree of freedom is less than 0.05 for the models with good fitness ,higher values to 0.08 show reasonable errors for approximation, the values equal or bigger

than 0.1-show weak fitness of data with model.Hu and Bentler (1999, cited in Homan, 2009) as good fitness, RMSEA smaller or equal to 0.06 is proposed. The results in Table 3 show better fitness of combination model 2PL/GPCM based on all investigated indices. Based on the results of **S-X^2**statistic and fitness indices in model level, using the combination of two-three parameters models with Generalized partial credit model and graded response model is suitable than combining one-parameter model with these models. If nested models of item-response theory are used, model fitness can be done based on Likelihood ratio test (LRT). Here, -2log likelihood (-2LL) of data fitness is compared with two required models. LRT is the difference between -2LLs of two nested models with Chi-square distribution. Degree of freedom in LRT is equal to the difference of estimated parameters in two models. Significant LRT indicates better fitness of model estimating more parameters and insignificance of this statistics shows the lack of difference between the models and selecting a nested model is economical (the model with low parameter) . As all the investigated models have not nest, this statistic is not used to compare all models, and combinational models of GRM family and combinational models of GPCM as separately are investigated by this statistic.

In combined models of GRM including 1PL/GRM,2PL/GRM,3PL/GRM, calculated chi-square is used to compare 1PL/GRM with 2PL/GRM as 5840.6 and based on degree of freedom 68 at confidence interval 99% (P<0.01) is significant. This result indicates good fitness of 2PL/GRM model with data. Calculated chi-square to compare 2PL/GRM with 3PL/GRM is equal to 73.01. This statistic with degree of freedom 68 is not significant at level P<0.05. Based on parsimony principle, it is better to select the model with lower parameters. Thus, 2PL/GRM model is selected as the model with better fitness. In combinational models of GPCM including 1PL/ GPCM,2PL/ GPCM,3PL/GPCM, calculated chi-square to compare 1PL/ GPCM and 2PL/ GPCM is equal to 5854.19 and based on degree of freedom 68 is significant at confidence interval 99% (P<0.01). This result indicates better fitness of 2PL/GRM with data. Calculated chi-square to compare 2PL/ GPCM with 3PL/ GPCM is 1.03 and this statistic with degree of freedom 68 at confidence interval 95% (P<0.05) is not significant. Based on parsimony principle, it is better to select the model with lower parameters and 2PL/GPCM in this family is selected as the model with best fitness.

Question 1-2: Is the fitness of data different with the model by combinational models (combination of dichotomous and polytomous models) compared to dichotomous models in mixed format tests?
To answer the above question, the study data are calibrated once by six combinational models in question 1-1 and another time by three dichotomous models including one, two, and three parameter models as simultaneously. For each of the studied models of statistical index of fitness as S-X^{2}are calculated to test fitness in item level and AICC،AIC، BIC and RMSEA to fit at test level. Significance level 0.05 is applied to test misfitting of items. Table 4 shows the items without fitness at significance level 0.05 with dichotomous models.

Table 4- The results of S-X^2 statistic in dichotomous models

Models	Misfitted items	N	%
1PL	1,3,6,7,,10,12,13,17,18,19,20,22,23,25,26,27,28,29, 30,31,32,33,34,35,36,37,38,39,40,43,44,45,46,47,4 8,49,50,51,52,53,54,57, 58,59,60,61,64,65,66,67,68,70,71,74,75,76	56	74
2PL	1,7,13,16,19,22,25,26,28,29,30,31,35,37, 38,47,48,51,60,61,70,74,75	23	30
3PL	1,13,25,29,31,35,37,38,47,48,49,51,61,67,68,70	16	21

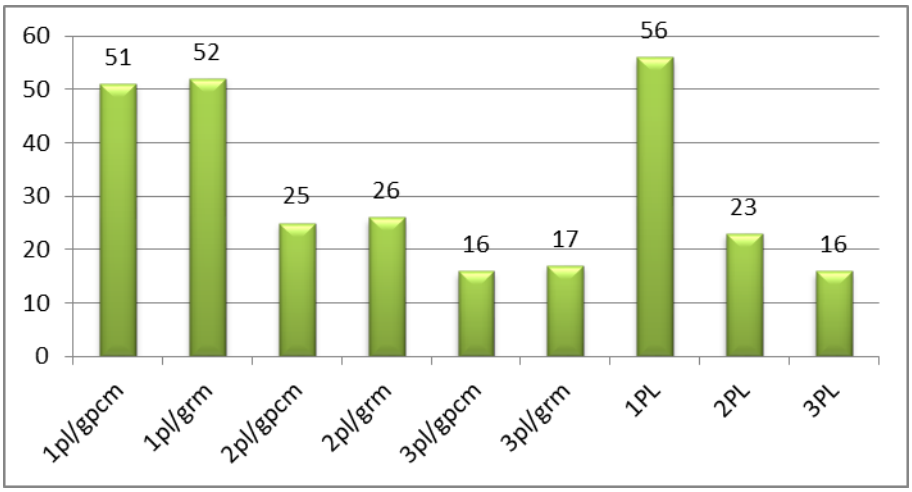

Figure 2- Frequency distribution of misfitted items based on studied models

As shown in Tables 2, 4, among dichotomous models, one-parameter model has the highest misfitted items and three-parameter model has the lowest misfitted items. Comparing the combined models and dichotomous models, one-parameter model has the highest misfitted item and three-parameter model and combination of three-parameter model with generalized partial credit model has the lowest misfitted items.

AICC، AIC، BIC and RMSEA statistics are used to fit at model level. The results are shown in Table 5.

Table 5- The results of fitness statistics in model level

Model / **Fitness statistics**	AIC	AICC	BIC	RMSEA
1PL	156174.34	156180.33	156608.78	0.06
2PL	149897.19	149921.28	150754.78	0.04
3PL	150250.82	150307.11	151537.20	0.04
1PL/ GPCM	173343.45	173355.60	173958.43	0.05
1PL/ GRM	173345.60	173357.75	173960.59	0.05
2PL/GRM	167639	167671.67	168632	0.03
2PL/GPCM	167623.26	167655.93	168616.26	0.03
3PL/GPCM	167758.23	167823.24	169134.89	0.03
3PL/GRM	167701.99	167767	169078.65	0.03

Based on Table 5, among dichotomous models, two-parameter and among combined models, the combination of two-parameter model with generalized partial credit model has better fitness with data based on all the investigated indices.

To investigate the model fitness in dichotomous models as 1PL, 2PL, 3PL, likelihood ratio test is used. Calculated chi-square to compare 1PL with 2PL is 6427.15 and based on degree of freedom 76 is significant at confidence interval 99%($P<0.01$) and this result shows better fitting of model 2PL with data. Calculated chi-square to compare 2PL model with 3PL is 201.63. This statistic with degree of freedom 76 is significant at confidence interval 95% ($P<0.05$) and this result shows better fitting of 3PL model with data. Second question: Is using various calibration procedures (separate and simultaneous) makes any difference in data fitting with model in mixed format tests?

To respond the above question of data, by six combinational models in question 1-1, calibration is done separately at first by simultaneous and then separate procedure. In separate calibration, dichotomous items are calibrated by dichotomous models and polytomous items by polytomous models. In the next stage, fitness includes S-$X^2$to test fitness in item level and AICC، AIC، BIC، RMSEA statistics for fitness at test level. Significance level 0.05 is applied to test misfitting of items. Table 6 shows the items misfitting at confidence interval 0.05 with the studied models.

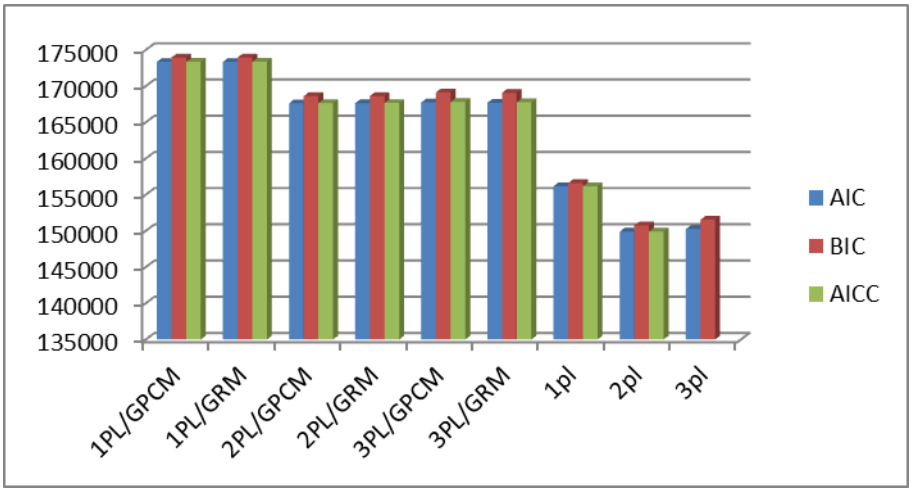

Figure 3- The results of fitness statistics in model level

Table 6- The results of S-X^2in combinational models based on calibration procedure

Calibration method	**Simultaneous calibration**			**Separate calibration**		
Models	Misfitted items based on S-X^2statistic	N	%	Misfitted items based on S-X^2 statistic	N	%
1PL/GPCM	3,6,7,10,17,18,19,20,22,23,25,26,27,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,45,49,50,52,53,54,57, 58,59,60,61,63,64,65,66,67,68,69,70,72,73,74,75,76	51	67	1,3,7,10,17,18,19,20,22,23,25,26,27, 28, 29,30,31,32,33,34,35,36,37,38,39,40, 42,43,44,45,46,47,48,49,50,51,52,53, 54,57,58,59,60, 61, 64,65,66,67,68,70, 73,74,75,76	54	71
1PL/ GRM	3,6,7,10,17,18,19,20,22,23,25,26,27,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,44,45,49,50,52,53,54,57, 58,59,60,61,63,64,65,66,67,68,69,70,72,73,74,75,76	52	68	1,3,7,10,17,18,19,20,22,23,25,26,27, 28, 29,30,31,32,33,34,35,36,37,38,39,40, 42,43,44,45,46,47,48,49,50,51,52,53, 54,57,58,59,60, 61, 64,65,66,67,68,70, 73,74,75,76	54	71
2PL/GRM	1,6,17,19,20,22,25,28,29,30,31,35,37, 38,39,40,45,50,52,60,61,63,69,70,74,76	26	34	1,20,22,28,29,30,32,35,37,38,40,43,44,45,46,47,48,49,58,60,61,70,74,75	24	31
2PL/GPCM	1,17,19,20,22,25,28,29,30,31,35,37, 38,39,40,50,52, 60,61,63,69,70,72,74,76	25	33	1,20,22,28,29,30,32,35,37,38,40,43,44,45,46,47,48,49,58,60,61,70,74,75	24	31
3PL/GPCM	17,20,28,29,35,39,40,46,49,50,53,61,63,68,70,72	16	21	29,35,43,44,45,46,47,48,49,53,61,67, 74,75	14	18
3PL/GRM	17,20,28,29,35,39,40,44,45,46,49,61,63,67,68,70,72	17	22	29,35,43,44,45,46,47,48,49,53,61,67, 74,75	14	18

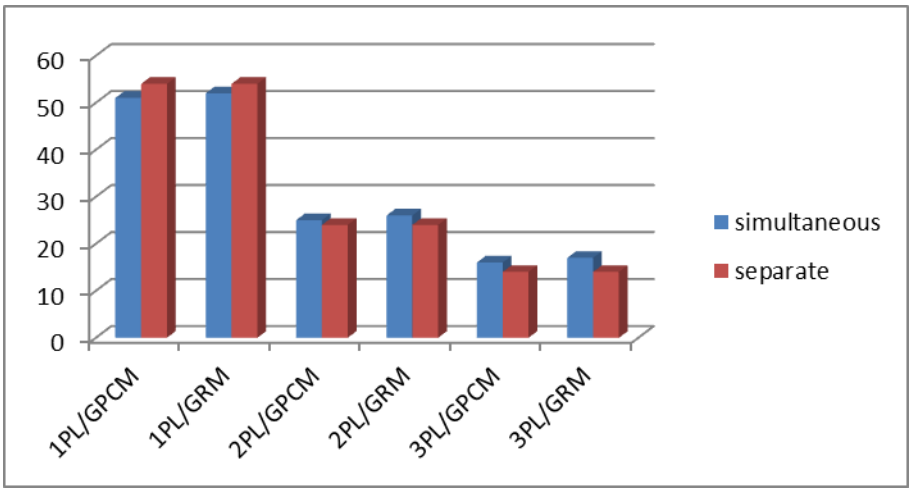

Figure 4- The results of S-X^2 statistic in combinational models based on calibration procedure

As shown in Table 6, when the items are calibrated separately, there is no difference between the combination of one-parameter models with generalized partial credit model and graded response model, combination of two-parameter model with generalized partial credit model and graded response model and combination of three –parameter model with generalized partial credit model and graded response model. This lack of difference is used among polytomous models. In separate calibration, three-parameter model has the lowest misfitted items and one-parameter model has the highest misfitted items. There is no difference between polytomous models. In comparison of simultaneous and separate calibration, separate calibration in two, three parameter models has better performance compared to simultaneous calibration and simultaneous calibration in one-parameter model compared to separate calibration .AICC, AIC, BIC and RMSEA statistics are used for fitting in model level and the results are shown in Table 7.

Table 7- The results of fitting statistics in model level by separate calibration

Model / **Fitting statistics**	AIC	AICC	BIC	RMSEA
1PL	141694.72	141699.52	142084.03	0.06
2PL	136022.43	136041.57	136789.75	0.04
3PL	136189.65	136231.01	137301.14	0.04
GPCM	34105.48	34107.08	34331.16	0.03
GRM	33946.11	33947.72	34171.79	0.03

As shown in Table 7, among dichotomous models, two-parameter model and among polytomous models, graded response model has better fitting with data in terms of all studied indices.

To investigate the model fitness with data in dichotomous models as 1PL, 2PL, 3PL, likelihood rate test is used. Calculated chi-square to compare 1PL model with 2PL model is 5806.29 and based on degree of freedom 68 is significant at confidence interval 99% ($P<0.01$). This result shows the better fitting of 2PL model with data. Calculated Chi-square to compare 2PL model with 3PL is 42.22 and this statistic with degree of freedom 76 is not significant at confidence interval 95% ($P<0.05$). Based on parsimony principle, it is better to select the model with lower parameters and 2PL model in this family is selected as model with better fitting.

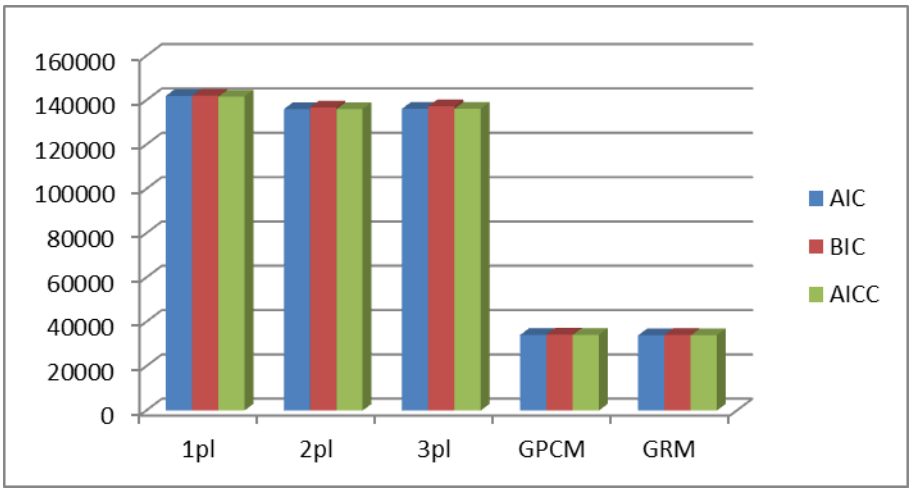

Figure 5- The results of fitness statistics in model level by separate calibration

DISCUSSION AND CONCLUSION

The present study aimed to assess goodness of fit of combined models based on item-response theory in mixed format tests. In the present study, three issues of assessment of model data fit in mixed format tests are considered. At first, the fitness in combined models is investigated based on item-response theory. The results show that among the applied models in this study, combining three-parameter model with each of polytomous models shows low misfitted items compared to combination of one-parameter model with each of polytomous models. Among polytomous models, generalized partial credit model shows lower misfitted items compared to graded response model. Regarding the fitting in model level, the findings show better fitness of combined model 2PL/GPCM with required indices. These results are in consistent with finding in Swaminathan et l., (2007); Chon et al., (2007). Then, a comparative analysis was done between combined and dichotomous models. Comparing combinational and dichotomous models, one-parameter model has the highest misfitted items and three-parameter model and combination of three-parameter with generalized partial credit model showed the lowest misfitted items. Regarding the fitting in model level, among dichotomous models, two-parameter model and among combinational models, combination of two-parameter model and generalized partial credit model showed better fitness with data. Finally, a comparison was made between simultaneous and separate calibration procedures, the results of comparison among dichotomous models showed better performance of separate calibration in two, three-parameter models compared to simultaneous calibration. There was no difference between two procedures in polytomous models. Regarding the fitting in item level in all assessed models, one-parameter model shows the highest misfitted items. This is due to difference of items discrimination parameter and unsuitability of this model for data, the invariance issue of item parameters and ability should be considered. As it was said, models fitness is an active area in item-response theory. Some points are as first, any considerable application of item-response theory should be measured by an assessment of a construct, and its fundamental conceptual model can be started. In the next step, using one or some statistics are proposed for person and item fitness. Here, it is not required to do hypothesis test for person or item fitness as formally and the researcher by these indices can identify the irrelevant items and subjects with measurement model. For example, big statistics of item fitness can guide the researcher to correct the items written with unclear terms or eliminate them of the test. The identification of subjects can guide the researcher for better estimation of item parameter. The studies of model data fit do not receive much attention among the researchers in this field. This issue can be due to complexity of fitness assessment, lack of perception of fitness statistics and the lack of comprehensive software in this regard (Zhao, 2008).As stated previously, there are no procedures that result in a researcher stating definitively that a particular model does or does not fit, or is or is not appropriate. Much like the assessment of practical fit in covariance structure modeling judging fit n IRT models calls for a variety of procedures to be implemented, and ultimately, a scientist must use his or her best judgment. (Embretson and Reise, 2000). This study can be repeated for the samples with different sizes and tests with various items with simulated and real data. Comparison of various procedures of calibration of items and various assessment methods of model fitness with data can be investigated in future studies.

REFERENCE

Baker F. B . and Kim S. H. (2004). Item response theory: Parameter estimation techniques (2nd ed.). New York: Marcel Dekker.

Bock R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*. 37: 29-51.

Cai L., Thissen D. and Du Toit S. H. C. (2011). IRTPRO for Windows [Computer software] Lincolnwood, IL: Scientific Software International.

Chon, K. H., Lee, W., Ansley, T. N. (2007). Assessing IRT model-data fit for mixed format tests. (CASMA Research Report) Iowa City, Iowa: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Pp. 26.

Chon K. H., Lee W. C. and Dunbar S. B. (2010). A comparison of item-fit statistics for mixed IRT models. *J. Ed. Measure.* 47: 318-338.

Chon K.H.a , Lee, W. C.b . and Ansley T.N.b (2013). An Empirical Investigation of Methods for Assessing Item Fit for Fit for Mixed Format Tests. *Appl. Measur. Edu.*, 26 (1): 1-15.

De Ayala R. J. (2009). The theory and practice of item response theory. New York, NY: Guilford.

DeMars C. (2010). Item response theory. Oxford, England: Oxford University Press.

Embretson S. E. and Reise S. P. (2000). Item Response Theory for Psychologists. NewJersey: Lawrence Erlbaum Associates, Inc., Publishers.

Fraser C. (1988). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [computer software]. Armidale, Australia: The University of New England.

Glas C. A. W. and Suarez Falcon J. C. (2003). A comparison of item-fit statistics for the three parameter logistic model. Applied PsychologicalMeasurement, 27: 87-106.

Hambleton R. K. Swaminathan H. and Rogers H. J. (1991). Fundamentals of item response theory. Newbury Park, CA:Sage.

Human H. (2009). Structural equations modeling with Lisrel software. Tehran. SAMT. [In Persian]

Kang T. and Cohen A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychol. Measurement.* 31: 331-358.

Kang T., Cohen, A. S., Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychol. Measurement.* 33: 499-518.

McDonald R.P. (1967). Non- linear factor analysis (Psychometric Monograph Psychometric Society. Pp. 15.

McDonald R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden and R. K. Hambleton(Ed.), Handbook of Modern Item Response Theory. New York: Springer Verlag. 258-269.

Orlando M. and Thissen D. (2000). New item fit indices for dichotomous item response theory models. *Appl. Psychology Measurement.* 24: 50-64.

Stone C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit_ test statistic in IRT models. *J. Edu. Measurement*, 37: 58-75.

Swaminathan H., Hambleton R. K. and Rogers H. J. (2007). Assessing the fit of item response theory models.In C. R. Rao, S. Sinharay (Eds.), *Handbook of Statistics.* 26: 683-718.

Thissen D., Wainer, H. (Eds) (2001). Test Scoring. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen D., Steinberg L. and Wainer H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Erlbaum. Pp. 147-169.

Whittaker T. A., Chang W. and Dodd B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Appl. Psychology Measurement.* 36: 159-180.

Wright B.D. and Stone M.H. (1979). Best test design. Chicago: MESA.

Yen W. M. (1981). Using simulation results to choose a latent trait model. *Appl. Psychological Measurement.* 5: 245-262.

Zhao Y. (2008). Approaches for addressing the fit of item response theory models to educational test data. Unpublished doctoral thesis, University of Massachusetts Amherst.

