iejee INTERNATIONAL ELECTRONIC JOURNAL OF ELEMENTARY EDUCATION

June 2019, Volume 11, Issue 5, 539-545

A Comparison of IRT Model Combinations for Assessing Fit in a Mixed Format Elementary School Science Test

Hacı Bayram Yılmaz*

Received: 12 March 2019
Revised: 22 May 2019
Accepted: 28 June 2019
ISSN: 1307-9298
Copyright © IEJEE
www.iejee.com

DOI: 10.26822/iejee.2019553350

Abstract

Open ended and multiple choice questions are commonly placed on the same tests; however, there is a discussion on the effects of using different item types on the test and item statistics. This study aims to compare model and item fit statistics in a mixed format test where multiple choice and constructed response items are used together. In this 25-item fourth grade science test administered to 2351 students in 35 schools in Turkey, items are calibrated separately and concurrently utilizing different IRT models. An important aspect of this study is that the effect of the calibration method on model and item fit is investigated on real data. Firstly, while the 1-, 2-, and 3-Parameter Logistic models are utilized to calibrate the binary coded items, the Graded Response Model and the Generalized Partial Credit Model are used to calibrate the open-ended ones. Then, combinations of dichotomous and polytomous models are employed concurrently. The results based on model comparisons revealed that the combination of the 3PL and the Graded Response Model produced the best fit statistics.

Keywords: Item Response Theory, Model Comparison, Mixed Format Tests, Item Fit

Introduction

Tests play crucial roles in individuals' lives. Exams are used for many reasons, such as selection and placement of individuals, determining which knowledge areas need to be improved, and planning and revising educational programs. Test design, analysis of test scores, and interpretation of test results have been important aspects of measuring examinees' trait levels (Kinsey, 2003). Public concern boosts discussions on tests regarding their reliability and validity, which are affected by many elements, such as test length, item format, and scoring.

Multiple choice (MC) items are the most common item types in tests. Despite the fact that the MC format is criticized since examinees can guess the answer correctly, many tests include only MC items due to not only budget and time constraints but also due to the difficulties in defending test scores to the public in plain terms. Although MC items are economically practical and they secure objective and reliable marking, it is difficult to measure higher order thinking with them. In addition, as Lissitz, Hou and Slater (2014) stress, if MC items are exclusively used in testing, the focus of instruction and learning will undermine the analysis, synthesis and evaluation skills of the learners, which in turn risk the loss of the active construction of knowledge. To eliminate these major limitations, it is possible to incorporate constructed response (CR) items in tests. On the other hand, CR items are difficult to score objectively and reliably despite they are considered to be measuring examinees' understanding of the content at a deeper level (Kim, Walker & McHale, 2008). Mixed format tests including both MC and CR items are highly effective measurement tools for teaching and learning to overcome the limitations stemming from their separate use. When they are combined, more reliable content total scores are obtained and a more precise latent trait is defined (Sykes & Yen, 2000). However, as Hollingworth, Beard and Proctor (2007) state, some educators and policy makers believe that constructed response items and multiple choice items do not measure the same construct when placed on the same tests.

The purpose of the present study was to investigate the applicability of separately and concurrently calibrating the dichotomous and polytomous items on a 4th grade science examination data using different Item Response Theory (IRT) models. Therefore, it would be possible to examine how model and item fit statistics vary when MC and CR items are analyzed separately and together. In addition, it will give insight regarding which IRT model is a better candidate for possible further use on achievement test data.

The Classical Test Theory (CTT) has been utilized in many testing systems; yet, it has many shortcomings such as the dependence of the values of item statistics (i.e., difficulty and discrimination) on a particular examinee sample, their average level of ability, and the range of scores. Another important shortcoming is that a valid comparison of examinees coming from different groups is possible only when the same or parallel tests are administered. In CTT, test reliability is described in terms of parallel forms although it is not practical in real world.

IRT has been employed to compute scale scores for achievement tests by most of the testing agencies throughout the world. When there is a reasonable fit between the selected model and data, IRT models produce invariant item statistics and ability estimates. As Hambleton and Swaminathan (1991) explained, the IRT estimate of an examinee's ability does not depend on a particular sample of test items. Also, the precision of ability estimates is known, and free to vary from one examinee to another (Baker, 2001). However, as Bergan (2010) reports, IRT model selection is often based solely on philosophical considerations rather than empirical tests. In general education policies dictate the choice of IRT model which results in a danger of misinterpretation of the data being analyzed as measures of relative fit are ignored (Brown, Templin & Cohen, 2015). Therefore, it is imperative to compare relative fit of competing models to avoid misleading interpretations about the data and making wrong decisions about test takers' performance.

[a,*]Correspondance Details: Hacı Bayram Yılmaz, Alanya Alaadin Keykubat University, Faculty of Education, Department of Educational Sciences, Antalya, Turkey. E-mail: bayram.yilmaz@alanya.edu.tr

IRT Models

Many different approaches have been developed to calibrate items in the IRT framework. The current study focuses on item calibrations based on the 1-, 2-, and 3- Parameter Logistic Models (1PL, 2PL, 3PL), the Generalized Partial Credit Model (GPCM), and the Graded Response Model (GRM). The roots of the 1PL model were introduced by a Danish mathematician, Georg Rasch. He demonstrated that item difficulties and examinee ability are sufficient statistics for measurement and introduced the Rasch Model (Rasch, 1960). In the 1PL model which was developed based on the Rasch's work, the probability of getting a correct response is plotted as a function of ability.

$$P_i(\theta_j) = \frac{e^{(1.7(\theta_j - \beta_i))}}{1 + e^{(1.7((\theta_j - \beta_i))}}$$

where θ_j is the ability and β_i is the difficulty parameter. The letter *e* is the base of natural logarithms (*e*≈ 2.118) and the 1.7 in the exponent lets the logistic function approximate the normal function (Warm, 1978). Although Rasch Model and 1PL are philosophically different (Andrich, 2004; Linacre, 2005), the differences between them are not in the scope of the current study. The 1PL model assumes an equal discrimination among all items, and a guessing parameter is not included in the model as it assumes that ability parameter is the sufficient statistic to compare individuals taking a particular test (Baghei and Carstensen, 2013). The two-parameter model was developed by Lord (1952) based on cumulative normal distribution. Birnbaum (1968) replaced the two-parameter logistic function with the two-parameter normal ogive function to model item characteristics (Hambleton, Swaminathan & Rogers, 1991). He modeled the probability of getting a correct response as a function of difficulty and discrimination parameters.

$$P_i(\theta_j) = \frac{e^{\alpha_i(\theta_j - \beta_i)}}{1 + e^{\alpha_i(\theta_j - \beta_i)}}$$

where α_i is the discrimination parameter.

Birnbaum (1968) modified the 2PL model by adding a parameter that represents the contribution of guessing to the probability of correct response (Baker, 2001). That is, the probability of correct response depends on guessing besides difficulty and discrimination in the 3PL model.

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{-\alpha_i(\theta_j - \beta_i)}}{1 + e^{-\alpha_i(\theta_j - \beta_i)}}$$

where c_i is the guessing parameter.

The partial credit model (PCM) was introduced in 1982 by Masters, who decomposed the response to an item into a series of ordered pairs of adjacent categories, then applied a dichotomous model to each pair assuming equal discriminations across the items (De Ayala, 2009). On the other hand, Muraki (1992) extended the equal discrimination assumption and applied the 2PL model to polytomously scored items and introduced the GPCM. This model assumes that the probability of choosing the k^{th} category over the *(k-l)*th category is expressed as the logistic dichotomous response model (Muraki, 1992), expressed as,

$$P_{jk|k-1,k}(\theta) = \frac{P_{jk}(\theta)}{P_{j,k-1}(\theta) + P_{jk}(\theta)} = \frac{\exp\left[Da_j\left(\theta - b_{jk}\right)\right]}{1 + \exp\left[Da_j\left(\theta - b_{jk}\right)\right]}$$

where, *k* represents the *n*= 2, 3, ...m, which are the response options. The GPCM is, then, written as

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} Z_{jv}(\theta)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} Z_{jv}(\theta)\right]}$$

and

$$Z_{jk}(\theta) = Da_j(\theta - b_{jk}) = Da_j(\theta - b_j + d_k)$$

where, *D* is a scaling constant (1.7) that sets the *θ* in the same metric as the normal ogive model, b_{jk} is an item category, and b_j is an item location parameter. While b_j represents the slope, d_k is the category parameter (Muraki, 1993).

GRM was developed by Fumiko Samejima (1969). Within the GRM, the b-parameter for each response category indicates the probability of an examinee whose *θ* is equal to the value of location parameter *(b)*, scoring *x* or higher is 50% on the CCRF (Tang, 1996). Samejima modeled the probability of a person responding in category *k* or higher versus responding categories lower than *k* as

$$p^*_{ix}\theta = \frac{\exp(Da_i(\theta - b_{xi}))}{1 + \exp(Da_i(\theta - b_{xi}))}$$

where, $P^*_{ix}(\theta)$ is the cumulative category response function (CCRF) representing the probability of scoring *x* or above on item *i* by an examinee with the proficiency level of *θ*. Probability of each score category is as follows:

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{ix+1}(\theta)$$

and the score category response function (SCRF) of the GRM can be written as

$$P_{ix}(\theta) = \frac{\exp\left[-Da_i(\theta - b_{ix+1})\right] - \exp\left[-Da_i(\theta - b_{ix})\right]}{\left[1 + \exp\left[-Da_i(\theta - b_{ix+1})\right] - \exp\left[-Da_i(\theta - b_{ix})\right]\right]}$$

The Partial Credit and Generalized Partial Credit Models are generalized from the dichotomous IRT models to describe an examinee's probability of selecting a possible score category among all score categories. Dichotomous IRT models describe how likely individuals at a certain ability level reach the score category *k* rather than *k-1*. So, *k* and *k-1* categories of polytomously scored items can be viewed as dichotomous categories. While the Partial Credit Model assumes that discrimination indices of all items are constant, the Generalized Partial Credit Model releases them free. These differences between the PCM and the GPCM are similar to those between the Rasch or the 1PL and 2PL models (Tang & Eignor, 1997). GRM, on the other hand, assumes that the boundary parameters of the categories are ordered. That is, each score category has a point where the probability of that category is highest.

Method

To realize the aims of the study, different IRT models were applied on the data collected form 2351 fourth grade students in 35 elementary schools in Turkey. The exam was part of a formative assessment initiative. Twenty five science items were asked to all participants, 14 of which were scored dichotomously and the remaining 11 were scored polytomously.

Before the items were written, 10 science teachers were selected as item writers based on school administrators' references and peer ratings about their teaching quality. A two-day training on item-writing was provided to teachers by two educational measurement specialists. Seventy five items were generated by those item writers and 25 items were selected based on content validity indicators set in accordance with the 4th grade science curriculum. The two educational measurement specialists participated in the item selection process along with the item writers. After the selection process, answer keys for each item was prepared by three item writers and the two specialists. During this process, possible and plausible answers for the graded items were prepared and a detailed rubric was developed. After the implementation of the exam, constructed response items were coded 0 if the answer was incorrect. It is coded as 1 if the answer was partially correct, 2 if it was correct, and number 9 was used to symbolize unattempted items.

After data collection, all answer sheets were graded by at least two teachers who also participated in the item writing process. In case there were discrepancies between the ratings of an item, the raters convened, discussed, and decided on the final mark. After the data collection, the dichotomous items were calibrated utilizing the 1PL, 2PL, and 3PL models, and the constructed response items were calibrated through GPCM and GRM. Then, all items were calibrated concurrently using mixed models. After estimating a model, it was compared with a competing more complicated better fitting one. Model selection was based on RMSEA, -2LL, number of unfitting items and item fit statistics as more parsimonious model is preferable. Violating the principle of parsimony creates unnecessarily complicated models and reduces predictions about new data sets (Kang, Cohen & Sung, 2009).

Results

Before performing analysis with IRTPRO (Cai, du Toit & Thissen, 2011), the unidimensionality assumption was tested by performing a categorical confirmatory factor analysis (Cat-CFA) with Mplus (Muthen & Muthen, 2012). A χ^2 value of 1549.42 with a 275 degrees of freedom indicated a poor fit ($p< .01$); however, it is known that Chi-square is affected by the sample size and this result is not surprising. An investigation of TLI (NNFI) (.90) and CFI (.91) results indicated a reasonable fit (Hu & Bentler, 1999). In addition, an obtained RMSEA value of .04 represented a good fit (Steiger, 2007). Therefore, the data set was considered unidimensional.

The first step of the IRT analysis in the current study was calibrating the MC and CR items separately and determining the number of misfitting items. Orlando and Thissen's (2000, 2003) *S-X*^{2} statistics were computed to evaluate item misfits throughout the study. This statistic was originally developed for dichotomous IRT models and was found to perform better than the traditional item-fit statistics. *S-X*^{2} was generalized to the polytomous models by Kang and Chen (2008, 2010). Dichotomously scored 14 items were calibrated with the 1PL, 2PL, and 3PL models. Table 1 below includes the results of the analyses.

Table 1. *Comparison of 1PL, 2PL, and 3PL Models*

	1PL	2PL	3PL
RMSEA	.05	.04	.04
Marginal Reliability	.56	.57	.62
-2LL	33170	33111.03	33037.58
Number of Misfitting Items	2	1	0

As seen above, the 1PL, 2PL, and 3PL models fit the data well based on the RMSEA statistics, each of which has an RMSEA value of .05 or less, indicating a close approximate fit (Kline, 2005). However, the reliability statistics were considered low, which may be due to the small number of items. It is important to note that although marginal reliability in the IRT framework is similar to the reliability in the CTT framework in that it is a measure of the overall test, marginal reliability is based on the average conditional standard errors at various levels on the measurement scale (Geen, Bock, Linn, & Reckase, 1984). Marginal reliability can be expressed as

$$\hat{\rho}=1-\frac{\int\sigma_e^2\left(\hat{\theta}\right)f\left(\hat{\theta}\right)d\hat{\theta}}{\sigma_x^2}$$

in which $\sigma_e^2\left(\hat{\theta}\right)$ is the conditional error variance and $f\left(\hat{\theta}\right)$ is the population density (Florida Department of Education, 2015). The literature suggests that the deviance test based on -2log likelihood (-2LL) statistics can be used to assess the model improvement. The difference in the -2LL statistic is distributed as a χ^2 statistic with the degrees of freedom equal to the difference in the number of parameters between the two models. If the difference in the -2LL is greater than the critical value, the addition of the extra parameters contributes significantly to the fit of the model (Hambleton, Swaminathan & Rogers, 1991). The difference in -2LL between 1PL and 2PL ($\chi^2(13)= 58.97$, $p< .05$) was found statistically significant. Similarly, that difference between 2PL and 3PL ($\chi^2(14)= 73.45$, $p< .05$) was also significant. These findings indicate that, as the parameters are added, model fit gets better. Furthermore, while 2 out of 14 items showed misfit in the 1PL model, 1 item showed misfit in the 2PL. All of the items fit the 3 PL model well. Table 2 includes detailed information regarding the *S-X*^{2} item level diagnostic statistics of 14 dichotomous items.

Table 2. *S-X*^{2} Item Level Diagnostic Statistics of MC Items

Item	1PL		2PL		3PL	
	X^2	*df*	*X*^{2}	*df*	*X*^{2}	*df*
1	12.98	11	11.51	11	13.10	10
2	21.43	9	18.46	9	5.13	9
4	17.87	10	16.26	10	15.92	9
6	28.82*	10	22.13	10	9.37	10
7	19.54	11	15.59	11	15.26	10
8	19.61	11	19.13	11	20.52	10
10	15.46	11	16.01	11	8.39	10
13	21.03	10	21.41	10	21.34	10
16	8.22	9	9.49	8	2.94	7
18	43.64*	11	27.62*	11	13.38	10
19	15.80	10	15.41	10	11.70	9
20	11.66	11	11.29	10	10.62	9
22	19.22	11	19.52	11	17.62	10
24	21.38	11	20.03	11	17.74	10

*$p< .01$

An investigation of item difficulties help one see how those values change as the model improves. As seen above, fit statistics increase significantly as the parameters are added. When Table 3 is examined it is seen that not only item difficulties but also order of the items based on their difficulty values are changed dramatically. For example while item 16 is the most difficult item when 1PL or 2PL is the model of choice, it is the fourth difficult one in 3PL model.

Considering RMSEA values, it might seem logical not to compare models and conclude that 1PL fits the data considerably well, further analysis of -2LL statistics on model improvement it is seen that not only the 3PL model is preferred over the 1PL and 2PL, it can be concluded that the difficulty values obtained for the first two models are misleading. Recall that difficulty parameter represents the proportion of examinees who respond correctly in 1PL, it represents that proportion after accounting for item-specific discrimination and guessing parameters (Bergan, 2010). After analyzing the MC items, the remaining 11 CR items in the test were analyzed through Muraki's GPCM and Semajima's GRM. As provided in Table 4, the fit statistics based on those two models are similar.

Table 3. Difficulty Parameters and Order of Items Based on Their Difficulties

Item	1PL		2PL		3PL	
	b	*Order of Difficulty*	*b*	*Order of Difficulty*	*b*	*Order of Difficulty*
1	-2.92	12	-2.62	14	-2.84	14
2	0.73	2	0.87	2	1.36	1
4	-3.03	13	-2.58	11	-2.54	11
6	0.13	3	0.17	3	1.12	2
7	-1.99	9	-2.59	13	-2.56	12
8	-2.64	11	-2.14	1	-2.17	10
10	-1.31	6	-1.28	6	-0.31	6
13	-0.65	4	-0.75	4	0.91	5
16	1.04	1	0.88	1	0.94	4
18	-1.06	5	-1.74	8	1.04	3
19	-1.46	7	-1.19	5	-1.15	7
20	-1.77	8	-1.56	7	-1.57	8
22	-2.09	1	-1.80	9	-1.80	9
24	-3.16	14	-2.58	12	-2.71	13

**p*< .01

Table 4. *Comparison of GRM and GPCM*

	GRM	GPCM
RMSEA	.04	.05
Marginal Reliability	.78	.77
-2LL	44509.85	44564.98
Number of Misfitting Items	3	3

Although RMSEA was computed as .05, indicating a good overall model fit, three items had poor fit statistics at .01 level when GPCM was used to conduct the analysis. A reliability value of .77 is considered to be acceptable. When the same 11 CR items were analyzed through Samejima's Graded Response Model (GRM), an RMSEA of .04 and a reliability of .78 indicate a slightly better overall fit than that of the GPCM. Both models had 3 misfitting items. Item statistics are provided in Table 5 below.

As the second step of the IRT analyses, all the MC and CR items were calibrated simultaneously and fit indices were examined to compare different models. The results of those analyses are given below.

Table 6 shows that the data have acceptable RMSEA and marginal reliability statistics in all combined models. The 1PL, 2PL, and 3PL models combined with GRM and GPCM fit the data well based on the RMSEA statistics. That is, when dichotomous and polytomous items are analyzed together in the current achievement test, both GRM and GPCM can be chosen.

Table 5. *S-X*^{2} Item Level Diagnostic Statistics for Polytomous Items

Item	GRM		GPCM	
	X^2	*df*	X^2	*df*
3	47.11	34	38.76	33
5	60.61*	32	54.41*	31
9	38.43	35	39.80	34
11	34.47	33	42.06	32
12	42.63	32	66.51*	31
14	33.73	32	39.97	32
15	37.00	32	45.54	33
17	55.07*	32	48.00	32
21	46.42	32	44.05	32
23	68.39*	34	87.18*	34
25	41.61	35	43.68	34

**p*< .01

A close look at the differences in -2LL statistics revealed that, as more parameters are added to the model, fit gets better. The -2LL difference between the 1PL and the 2PL (χ^2(13)= 545.05, *p*< .05) was significant; however, the difference between the 2PL and the 3PL was not (χ^2(14)= 13.62, *p*> .05) if GRM is used for the CR items. Similarly, -2LL statistics difference between the 1PL and the 2PL (χ^2(13)= 287.95, *p*< .05) was significant; however, between the 2PL and the 3PL (χ^2(14)= 13.42, *p*> .05), the difference was not significant when GPCM is used for the CR items. These preliminary results suggest that when dichotomous and polytomous models are combined in the same test, GRM and GPCM produce similar results. That is, considering the overall model fit statistics, after one decides which polytomous model will be used; s/he can choose the 2PL or 3PL model for the dichotomously scored items. Yet, one should take the item statistics in consideration before making the final decision regarding the model. Table 7 provides item-level fit values for all combined models.

As Table 7 displays, out of 25 items, 8 items misfit the 1PL, 4 items misfit the 2PL models, and 3 items misfit the 3 PL model when GRM is the model of choice for polytomous items. On the other hand, 8 out of 25 items displayed misfit when the 1PL is applied to the dichotomously scored items when GPCM is the model of choice for the polytomous ones. This number went down to 6 in the 2PL and to 4 in the 3 PL model with the combination of GPCM. When the item diagnostics regarding the MC items are examined, it is seen that items 2 and 7 do not fit under any combined models. Item 6 fits all the models except when the GRM or the GPCM is combined with the 1 PL. Item 13 fits all the models except when the GRM is combined with the 1PL, and the item 18 fits all the models except when the GRM or the GPCM is combined with 3PL. There are three CR items displaying misfit under different models. The fit statistics of the item 12 appear to be acceptable only when the GRM is combined with the 2PL or the 3PL. Item 14 is considered as misfitting like item 18 when the GRM is combined with the 1PL. Item 23 does not fit when the GPCM is combined with the 2PL or the 3PL. Based on the item level statistics, it can be concluded that the data have best fit statistics when the 3PL and GRM models are combined.

Table 6. *Model Fit Statistics of Combined Models*

	1PL & GRM	1PL & GPCM	2PL & GRM	2PL & GPCM	3PL & GRM	3PL & GPCM
RMSEA	.05	.04	.04	.04	.04	.04
Marginal Reliability	.80	.80	.82	.82	.82	.82
-2LL	77330.80	77112.34	76785.75	76824.39	76772.13	76810.97
Number of Misfitting Items	7	8	4	6	3	4

Table 7. *S-X^2 Item Level Diagnostic Statistics for All Items*

Items	1PL&GRM		2PL&GRM		3PL&GRM		1PL&GPCM		2PL&GPCM		3PL&GPCM	
	X^2	*df*	X^2	*df*	X^2	*df*	X^2	*df*	X^2	*df*	X^2	*df*
1	35.83	26	28.69	25	28.65	24	31.15	26	28.73	25	28.69	24
2	89.85*	24	48.06*	26	48.04*	25	75.22*	26	48.19*	26	48.18*	25
3	68.77	49	68.17	49	67.63	49	53.85	47	53.99	47	53.55	47
4	36.66	26	36.73	26	36.73	25	34.70	25	36.74	26	36.76	25
5	83.45*	49	76.67*	47	75.38*	46	76.33*	48	70.04*	45	69.92	45
6	97.52*	25	42.55	28	33.74	27	68.36*	25	42.39	28	33.86	27
7	76.48*	27	50.72*	28	50.48*	27	60.94*	27	50.94*	28	50.70*	27
8	38.01	26	26.75	25	26.55	24	49.48*	26	26.67	24	26.45	23
9	52.33	47	52.06	47	51.91	47	55.57	48	51.32	47	51.26	47
10	37.28	27	28.39	28	28.40	27	28.53	27	28.44	28	28.51	27
11	73.48	48	46.78	45	46.60	45	59.63	46	53.24	46	53.11	46
12	94.30*	51	66.12	47	66.22	47	95.18*	49	87.58*	45	87.77*	45
13	64.41*	26	30.75	28	30.73	27	45.39	26	30.77	28	30.75	27
14	78.14*	48	42.36	43	42.05	43	69.66	45	53.50	44	53.16	44
15	68.54	49	40.65	44	39.63	43	48.04	46	47.08	45	47.19	45
16	30.63	24	29.25	24	29.28	24	30.24	25	29.54	24	30.27	24
17	68.61	44	64.22	43	63.97	43	61.76	44	64.07	44	63.84	44
18	136.53*	27	51.17*	28	41.38	27	102.20*	27	51.03*	28	41.37	27
19	25.00	27	21.04	27	21.10	26	21.10	27	21.09	27	21.15	26
20	43.58	27	36.47	28	36.43	27	37.43	28	36.56	28	36.51	27
21	64.76	46	58.70	44	59.00	44	62.03	44	61.46	44	61.62	44
22	28.45	26	27.89	26	27.91	25	31.87	27	27.90	26	27.91	25
23	67.58	52	61.73	48	61.61	48	77.30*	47	75.05*	47	74.95*	47
24	26.71	25	25.77	25	25.76	24	25.63	25	25.78	25	25.76	24
25	60.69	50	63.48	52	63.78	52	73.98	49	68.48	52	66.35	51

$^*p< .01$

Since the GRM and the GPCM are not nested models, traditional model comparison statistics, such as comparing -2LL differences, are not appropriate to decide whether a combination of the 3PL and GRM or the 3PL and GPCM models provide better fit for the data used in this study. On the other hand, it is possible to use Akaike's Information Criterion (AIC: Akaike, 1974) and Schwarz's Bayesian Information Criterion (BIC: Schwarz, 1978) for this purpose (Kang, Cohen, & Sung, 2005). As both GRM and GPCM models have the same number of parameters (Bartolucci, Bacci, & Gnaldi, 2015), it is logical to compare them utilizing AIC and BIC. Although significance tests are not available with these statistics, they provide estimates of the relative differences between the two options.

AIC and BIC statistics were computed as 76960.97 and 77393.23 respectively for the combination of the 3PL with the GPCM; on the other hand, an AIC of 76920.13 and a BIC of 77346.63 were obtained when the GRM was selected with the 3PL model for the dichotomous items, which can be considered as a sign that supports the conclusion that the combination of the 3PL and GRM models has a

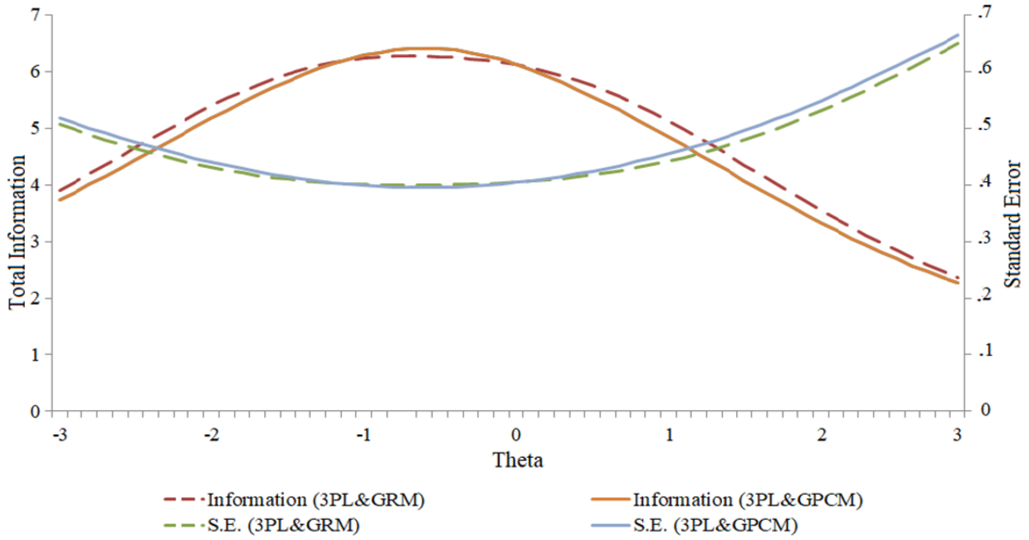

Figure 1. Total Information Functions and Standard Errors Obtained Through Final Models

better model fit than that of the 3PL and the GPCM. A further analysis of total information functions and standard errors would show the difference between the two competing choices.

Above graph compares the test information functions and corresponding standard errors. Combination of 3PL and GRM models provide higher information with lower standard errors as the ability of test takers get closer to the lower end and higher end of the theta distribution. On the other hand, the combination of 3PL and GPCM models provides slightly more information for the students with ability level close to the mean.

General Discussion

The goal of this study was to assess the changes in fit statistics when dichotomous and polytomous items were calibrated separately and concurrently. The 1PL, 2PL, and 3PL IRT models were applied to dichotomously coded MC items, and it was seen that, in general, as the parameters are added to the model, fit statistics get better. When the GPCM and GRM models are compared, the GRM is the model of choice for the analyzed data due to higher reliability and lower RMSEA and -2LL statistics. The results show that multiple choice and constructed response items can effectively be used in the same test when the data are analyzed through IRT models.

It is seen that 1PL&GRM and 1PL&GPCM have the same number of misfitting items; however, 2PL&GPCM has more misfitting items than 2PL&GRM. In addition, 3PL&GPCM has more misfitting items than 3PL&GRM. RMSEA statistics are (.04) the same for all combinations except for the 1PL&GRM (.05). Reliabilities are the same (.82) for all the combined models except for 1PL&GRM and 1PL&GPCM (.80).

Considering the reliability statistics, the change in the number of misfitting items and RMSEA statistics, the most promising combination is 3PL&GRM for the data utilized in this research. The findings support the conclusions reached by Sykes and Yen (2000), who reported substantially more items not fitted when the 1PL is combined with polytomous response models than 3PL. On the other hand, the findings of current study do not fully confirm the findings of Chon, Lee and Ansley (2007), who stated that the 3PLM and GPCM models tended to fit the mixed format data best.

This study serves as a promising step in the utilization of combined models in elementary school tests. More studies are needed to discover the applicability of such analyses in different subjects, such as literacy and mathematics. As indicated previously, the data used in this study are unidimensional. In real situations, it is likely to have a multidimensional data set. Therefore, further studies should be conducted on such data sets. Although misfitting items are determined, the reason for the misfit is out of the scope of the current study. Further studies using effect sizes to quantify the misfits and exploring the reasons for the misfit are encouraged.

References

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

Andrich, D. (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms?. *Medical Care, 42*, 7-16.

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). Retreived from http://pareonline.net/getvn.asp?v=18&n=5.

Baker, F. (2001). *The basics of Item Response Theory*. Portsmouth, NH: Heinemann.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2015). *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Boca Raton, FL: Chapman and Hall/CRC.

Bergan, J. R. (2010). *Assessing the relative fit of alternative item response theory models to the data*. Retrieved from http://www.ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf

Brown, C., Templin, J. & Cohen, A. (2015). Comparing two- and three-parameter logistic models via likelihood ratio tests: a commonly misunderstood problem. *Applied Psychological Measurement 39*(5), 335-348.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). Irtpro (Version 2.1) [Computer software]. Retrieved from http://www.ssicentral.com.

Chon, K.H., Lee, W. & Ansley, T.N. (2007). *Assessing IRT model-data fit for mixed format tests* (CASMA Report No. 26). Iowa City, Iowa.

De Ayala, R.J. (2009). *The theory and practice of Item Response Theory*. NY: Guilford.

Florida Department of Education (2015). *Florida standards assessments technical report (4).* Tallahassee, Florida. Retrieved from www.fldoe.org/core/fileparse.php/5663/urlt/1415TechV4FSA.pdf

Green, B. F., Bock, R. D., Humphreys, L.G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton, R.K., Swaminathan, H.& Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* California: SAGE.

Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An Investigation of Item Type in a Standards-Based Assessment. *Practical Assessment, Research & Evaluation, 12*(8). Retreived from http://pareonline.net/getvn.asp?v=12&n=18

Hu, L.T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Jiao, H., Liu1, J., Haynie, K., Woo, A., & Gorham, J. (2011). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement, 20*(10), 1-17.

Kang, T., Cohen, A.S., & Sung, H. J. (2005, March). IRT model selection methods for polytomous items. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kang, T., Cohen, A.S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33(*7), 499-518.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item-fit index for polytomous IRT models. *Journal of Educational Measurement, 45*, 391-406.

Kang, T., & Chen, T. T. (2010). Performance of the generalized S-X^2 item fit index for the Graded Response Model. *Asia Pacific Education Review, 12*(1), 89-96.

Kim, S., Walker, M.E. & McHale, F. (2008). *Equating of mixed-format tests in large-scale assessments*. ETS. Princeton: NJ.

Kinsey, T.L. (2003). *A comparison of IRT and Rasch procedures in a mixed-item format test* (Unpublished Ph.D. thesis). Retrieved from University of North Texas Theses and Dissertations Collections database (OCLC: 53783278).

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Linacre J.M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions, 19*(3), 1032. Retreived from https://www.rasch.org/rmt/rmt193h.htm

Lissitz, R. W., Hou, X. & Cadman Slater, S. (2014). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact. *The Journal of Applied Testing Technology 13*(3), 1-50. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48366

Muthen, L.K. & Muthen, B.O. (2012). Mplus (Version 7) [Computer software]. Retrieved from http://statmodel.com.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (1993). Information functions of the Generalized Partial Credit Model. *Aplied Psychological Measurement, 17*(4), 351-363.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous Item Response Theory models. *Applied Psychological Measurement, 27*, 289–298.

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

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

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling, *Personality and Individual Differences, 42*(5), 893-98.

Tang, K.L (1996). *Polytomous Item Response Theory models and their application in large-scale testing programs: Review of the literature* (ETS Report No. RM-96-08). Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/1996/ibtw.

Tang, K.L & Eignor, D.R. (1997). *Concurrent calibration of dichotomously and polytomously scored Toefl items using IRT models* (ETS Report No. RM-97-6) Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/1997/hxvc.

Warm, T. A. (1978). *A primer of Item Response Theory*. Oklahoma City, OK: U.S. Coast Guard Institute.