Yılmaz Koğar & Koğar, 2018

Volume 2 Issue 1, pp.29 - 49

Date of Publication: 15th March, 2018

DOI-https://dx.doi.org/10.20319/pijtel.2018.21.2949

This paper can be cited as: Yılmaz Koğar, E., & Koğar, H. (2018). Examination of Dimensionality and Latent Trait Scores on Mixed-Format Tests. PUPIL: International Journal of Teaching, Education and Learning, 2(1), 29-49.

EXAMINATION OF DIMENSIONALITY AND LATENT TRAIT SCORES ON MIXED-FORMAT TESTS

Esin Yılmaz Koğar
Faculty of Education, Ömer Halisdemir University, Niğde, Turkey
esinyilmazz@gmail.com

Hakan Koğar
Faculty of Education, Akdeniz University, Antalya, Turkey
hkogar@gmail.com

ABSTRACT

The aim of the present study is to examine various item types utilized to measure success in mathematics in terms of dimensionality and latent trait scores. The data collection instruments utilized in the present study were the student questionnaires and the mathematics achievement tests developed to measure 4th and 8th grade students' mathematics success in TIMSS 2015. It is assumed in the current study that two different dimensions are formed: the combination of MC and CR items, forming the "math ability" and the CR items, forming the "CR ability". To determine the dimensionality of latent trait of math ability, three different IRT models – unidimensional, within-item and between-items – were used. It was found that the within-item model displayed a better fit, when compared to the unidimensional model. Moreover, the within-item dimensional model showed better fit according to AIC and BIC as well. In the unidimensional and within-item models, the talent parameter predictions were similar. While the effect of the variables of sense of school belonging and students' confidence in mathematics *on the primary trait were significant, the home resources for learning variable also had a significant impact within 8th grade.*

Keywords

Multiple choice items, constructed response items, mixed format tests, multidimensionality, latent dimension score

---

1. Introduction

One of the topics that educational research studies mostly focus on is how student success can be measured more effectively. For this purpose, different item types appropriate for measuring students' academic proficiency levels were tried to be determined. Item types are generally referred to as selected response (SR) and constructed response (CR) items. SR items are multiple choice (MC) items and those that are derivatives of these items. These items necessitate the selection of the correct answer to a question or a situation from among the given alternative responses (Simkin & Kuechler, 2005). The most widely used SR items is the multiple choice item type. On the other hand, the constructed response items require students to construct their own answers (Popham, 2006). Short-answer items, portfolio evaluations, oral exams, science projects and the like can be listed as examples for constructed response items (Hogan & Murphy, 2007). Even musical performances or portfolios formed for visual arts are among CR item types (Pollack, Rock, & Jenkins, 1992).

It is believed that MC item types measure cognitive behaviors that are based on the lower levels of the cognitive classification and, thus, MC items should be designed so as to measure higher level behaviors (Hancock, 1994; Simkin & Kuechler, 2005). However, there is persistent, common belief that CR items measure more complicated and higher level skills (Martinez, 1999). Bennett, Rock and Wang (1991) stated that items consisting of MC items are those that measure simple, conceptual definitions, while tests consisting of CR items enable the measurement of higher order thinking skills. On the other hand, MC items have such positive features as enabling ease in implementation and scoring specifically in wide scale tests (Dufresne, Leonard, & Gerace, 2002), ensuring objectivity in scoring (Becker & Johnston, 1999), enabling numerous questions representing the content coverage to be asked (Saunders & Walstad, 1990), and having the potential to be administered to many people. However, MC items have many disadvantages, which are discussed in the related literature. For example, Brown, Bull and Pendlebury (2013) claimed that when compared to CR items, constructing

MC items, particularly when an item pool was inaccessible, was more difficult. Another disadvantage emphasized by Dufresne et al. (2002) was that poorly written MC items could conceal the test takers' knowledge rather than reveal it. For this reason, it can be maintained that MC items are difficult to prepare and require specialization. Another point is that as there is the possibility of giving correct answers to questions based on MC items by chance, whether or not the student provided a correct answer based on possessing the appropriate knowledge cannot be identified (Hastedt, 2004). Furthermore, the fact that MC items provide individuals with less opportunity to organize, synthesize, discuss and express their knowledge than do CR items is considered as the limitations of MC items (Lukhele, Thissen, & Wainer, 1994; Tuckman, 1993). There are also views that CR items measure students' real life problem solving skills more effectively (Bacon, 2003; Fenna, 2004). However, CR has negative features, such as requiring much time for them to be answered and lowering content validity (Griffo, 2011), entailing the potential impact of subjective claims on scoring as responses are scored by individuals (Downing, 2006; Wainer & Thissen, 1993), and requiring verbal skills to answer them (Haladyna, Downing, & Rodriguez, 2002).

Martinez (1999) stated that no item type is completely appropriate for all purposes and conditions. Even though MC and CR items have their own strengths and weaknesses, using them in combination makes significant contributions to assessment and is a means to strengthening the validity of the assessment (Ercikan et al., 1998). Furthermore, using these two different types of items in combination in a test can compensate for items' weaknesses by synthesizing its strengths (Cao, 2008). Based on these various views, in current scales, different types of items are used in combination in assessment instruments. There are numerous studies on different item types in the related literatüre (Ackerman & Smith, 1988; Bennett et al., 1991; Hancock, 1994; Marengo, Miceli, & Settanni, 2016; Ward, Dupree, & Carlson, 1987). Thissen, Wainer and Wang (1994) have stated that there has been a growing interest in the use of MC and CR items in combination, which has led to the question of whether or not these two types of items measure the same things.

There are discussions over whether or not the items in composite tests in which different item types are used in combination measure the same feature, and the findings obtained from studies regarding the dimensional aspect of these tests are complicated. Some of these studies report that different item types measure that same feature or structure (Bacon, 2003; Bennett et al., 1991; Bridgeman, 1991; Ercikan et al., 1998; Hancock, 1994; Griffo, 2011; Thissen et al., 1994; Wainer & Thissen, 1993). In a study by Wang (2002), it was reported that the differences between MC and CR items did not affect the structure of the test, and that most of the MC and CR items were loaded onto the same factor in single factor analyses. Accordingly, he refused the hypothesis that MC and CR items measured different mathematical competencies. On the other hand, some studies yielded results which pointed to the fact that different item types do not measure the same structure (Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Ercikan et al., 1998; Lissitz, Hou, & Slater, 2012; Sykes, Hou, Hanson, & Wang, 2002; Walker & Beretvas, 2003; Ward, Frederiksen, & Carlson, 1980). In a study by Birenbaum and Tatsuoka (1987), in which 285 8th grade students' arithmetic abilities were measured by means of both MC and CR items, it was determined that the two tests had different structures. Similarly, Traub and MacRury (1990) maintained that the two different item types measured different abilities, but that the nature of these differences were not clear. Kuechler and Simkin (2010) compared individuals' performances in MC and CR items which were designed to assess proficiency in topics requiring the same cognitive level, and they arrived at the conclusion that CR items were in general more complicated than MC items. The findings of some studies indicate that multidimensionality in mixed-format tests that include items with different formats could be originating from the format of the items (Cao, 2008; Kim & Kolen, 2006).

The use of MC and CR in combination in mixed-format tests can lead to some assessment problems (Marengo et al., 2016). Today, item response theory (IRT) is frequently used in developing and evaluating wide scale tests owing to its strong mathematical infrastructure. In operations based on IRT, the dimensionality of the data set needs to be examined, and the selected IRT model should be appropriate to the dimensionality of the data structure. Dimensionality can be defined as the number of latent variables that takes into consideration the correlations among the item responses within a certain data set (Camilli, Wang & Fesq, 1995). Gessaroli and Champlain (2005) define the dimensionality of a test as the function of the interaction between an item set and the test taker group responding to these items. In order not to arrive at erroneous results in IRT predictions, it is important to dimensionality analyses should be done to determine whether or not the test has a single or multiple dimensions. For this reason, showing that the test is predominantly assessing one factor indicates that the unidimensionality is secured (Embretson & Reise, 2000).

Multidimensionality, on the other hand, refers to the items in the test measuring more than one factor (Smith, 2009).

In the present study, basically whether or not different item types in the same test lead to dimensionality is examined. Wang (2002) stated that unlike unidimensional IRT models, multidimensional models have the advantage of discovering or confirming dimensionality. For this reason, the dimensionality of the data set is examined by using both unidimensional and multidimensional IRT models. Between-item models and within-item models, which are multi-dimensional models, can be defined as follows:

The Multidimensional Between-item Models: These tests consist of numerous sub-tests assessing different but related latent dimensions. In these tests, each item belongs only to one sub-scale and there is no common item between the scales (Adams & Wu, 2010). In other words, in this model, there are different unidemensional sub-scales (Wang, 1994).

The Multidimensional Within-item Models: In a test consisting of items assessing more than one latent feature, some items take place in more than one dimension. In such cases, these tests are referred to as within-item multidimensional tests (Adams & Wu, 2010).

The graphical representation of the above information is presented in Figure 1.

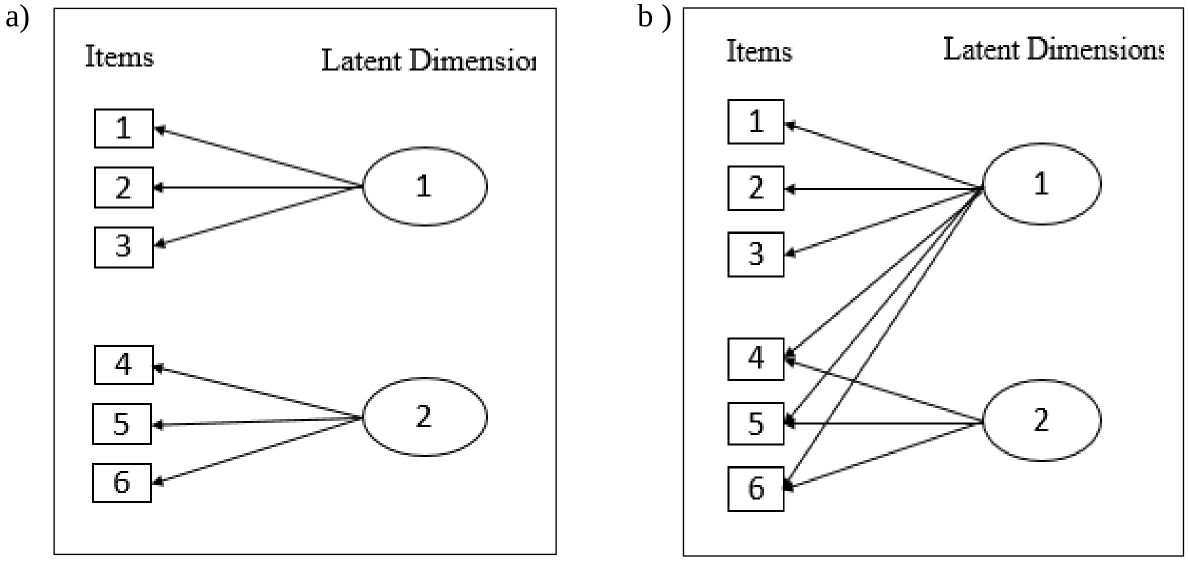

Figure 1: *A Graphical Representation of Between-item (a) and Within-item (b) Multidimensionality*

1.1 The Aim and Significance of the Study

In the present study, whether the available data set were more fitted with the unidimensional IRT or the multidimensional between-item and within-item models was examined to answer the question, “Does item type used in mixed-format tests consisting of different item types result in dimensionality?” In addition, some variables that could affect these features, such as gender, were selected to examine their impact on math ability as a primary feature and CR ability as a secondary feature.

Another aim of the study was to investigate the impact of a multidimensional format on a unidimensional model in grouping students according to different proficiency levels. The fit between ability distributions based on a unidimensional model and the ability distributions for the main dimension (MC+CR) based on the multidimensional model was examined. It was aimed to draw attention to the necessity to review the process of ability identification in situations where classification was found to entail a high level of incompatibility.

Thus, the present study focused on whether or not two item types assessed the same structure. The focus of the study is considered to be significant among education studies as there is an increase in the use of different item types, which, as reported in education studies, assess not only different cognitive structures, but also different structures (Nickerson, 1989). It is beleived that the current study will provide teachers and test developers and implementers with beneficial information in choice of item type, choice of model, test development, test calibration and reporting of the scores.

2. Methodology

2.1 Participants and Procedure

The sample of the current study was determined within the scope of TIMSS, which is a scanning research, held every four years by the International Association for the Evaluation of Educational Achievement, based on assessing information and skills that 4th and 8th grade students have learned in content areas of mathematics and sciences. A total of 61 countries participated in the TIMSS 2015 implementation. The 4th and 8th grade students participating in this implementation from Turkey were selected randomly. A total of 6456 (49.2% female ve 50.8% male students) 4th grade students from 260 schools, and a total of 6079 (48.4% female and 51.6% male) 8th grade students from 238 schols from Turkey participated in TIMSS 2015.

Thus, analyses were made on a total of 462 (48.92% female and 51.08% male) 4th grade students (1 student from the sample of 463 students was eliminated from the data set for leaving most of the questions unanswered) to whom the 14th booklet was administered and a total of 435 (46.9% female and 53.1% male) 8th grade students to whom the 2nd booklet was administered.

2.2 The Data Collection Instruments

In the present study, as a data collection instrument, achievement tests in the TIMSS assessment entailing items based on assessing 4th and 8th grade students' performances in mathematics and student questionnaires, which provided various types of information about the students, were utilized.

Mathematics Achievement Test

The distribution of all the items in the 4th and 8th grade mathematics assessment in TIMSS is presented in the following tables according to learning domains and item format. It can be observed that different content domains and a different number of items were used in the 4th and 8th grade mathematics achievement tests. These tests include MC and CR items designed to assess students' cognitive domain (knowing, applying and reasoning) on multiple math contents domains. MC items have four options and the item has only one correct answer. The correct answers for these items are coded as 1 and the incorrect ones are coded as 0. However, some compound multiple-choice items are worth two score points (2 score points: fully correct, 1 score point: partially correct), and in CR items, students respond to the items by writing or drawing. These items are scored by using scoring rubrics developed for each item. Some of the CR items are coded as 1-0, similar to MC items, while in others, incorrect responses are coded as 0, partially correct responses are coded as 1 and completely correct answers are coded as 2 (Martin, Mullis, & Hooper, 2016). In the TIMSS 2015 assessment, there were 14 mathematics and 14 science blocks. While six of these 14 blocks made up the new assessment block, the remaining eight were from the TIMSS 2011 assessment (IEA, 2016). These blocks were distributed to 14 test booklets, they have four blocks which consists of two in math and two in science (Yıldırım et al., 2016).

TIMSS technical report about the achievement data expresses that for the analysis of responses is used the IRT scaling. Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the TIMSS 2015 assessment data - a three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for constructed response items with just two response options, which also were scored as correct or incorrect and a partial credit model was used with polytomous constructed response (Martin et al., 2016).

4th Grade Data Set: In this study, test boooklet no 14 was used for 4th grade students. Booklet no 14 is comprised of a total of 25 items scored in the mathematics achievement test.

Of these items, 15 are MC, while 13 are CR items. Only one of CR items is scored partially.The reason why booklet no 14 was chosen is based on the fact that the number of MC and CR items was close to each other and that only one of the CR items was scored based on multiple categories. Thus, with the assumption that the information loss in the information function of the test when CR item was scored as 1 or 0 would not be very signficant, the completely correct answers to this item was coded as 1 and those responses that were partially correct or totally incorrect were recoded as 0.

8th Grade Data Set: In the present study, the second test booklet was chosen for 8th grade students. In booklet no. 2, the mathematics achievement test is comprised of 29 items. Of these, 16 are MC and 13 are CR items. Only one of the CR items has partial scoring. The reason underlying the selection of this booklet is based on the fact that the number of MC and CR items was close to each other and that only one of the CR items was scored based on multiple categories. Thus, with the assumption that the information loss in the information function of the test when CR item was scored as 1 or 0 would not be very signficant, the completely correct answers to this item was coded as 1 and those responses that were partially correct or totally incorrect were recoded as 0.

Student Questionnaire

In addition to assessing students’ mathematics and science performances, TIMSS obtains information regarding students’ contexts with the aid of wide-range questionnaires, one of which is the student questionnaire. The student questionnaire is filled out by the students participating in the implementation. The student questionnaire is comprised of questions addressing students’ home and school life, their self-perceptions, their attitudes towards mathematics and science lessons, homework and out of school activitives, their use of the computer, the educational tools and resources they have in their home and their personal information. Various scales are formed using the IRT scaling methods, particularly the Rasch partial scoring method, based on the responses given to the questionnaire items by the students (Martin et al., 2016). The indices used this study are as follows:

Students Sense of School Belonging Scale (SSB): SSB expresses the students' feelings towards their own school and their sense of school belonging. The items in this scale consists of seven items, which require the respondent to indicate his/her level of agreement on a 4-level scale: "completely agree, partially agree, partially disagree, completely disagree".

Student Bullying Scale (SB): SB was formed by asking students to what extent they experienced the eight bullying they behaviours defined in the questionnaire based on a scale of frequency: "never, a couple of times a year, once or twice a month, at least once a week."

The Students Like Learning Mathematics Scale (SLM): SLM is based on students' levels of agreement (completely agree, partially agree, partially disagree, completely disagree) with the given nine statements.

The Students' Views on Engaging Teaching in Mathematics Lessons Scale (EML): EML was formed as a result of students' agreement levels (completely agree, partially agree, partially disagree, completely disagree) with ten items.

Students Confident in Mathematics Scale (SCM): SCM measures to what extent students are self-confident in their math ability. For this purpose, students are asked to indicate whether they "completely agree, partially agree, partially disagree, completely agree" with the nine items in the scale (seven of the items are from TIMSS 2011 implementation and two items are new).

The Home Resources for Learning Scale (HER): HER is an index score derived from the answers that the 8th grade students had given to three items on the scale.

The Students Value Mathematics (SVM): SVM is an index score comprised of the answers 8th grade students had given to three items on the scale. Students were asked to "completely agree, partially agree, partially disagree, partially disagree" with the given statements.

In the present study, gender of the students, a categoric variable, was also addressed as an independent variable for both grade levels.

2.3 Analysis of Data

Dimensionality Analyses: Initially a unidimensional model was established by modelling as if all the items were based on one latent dimension, and this model was analyzed by utilizing the Rasch model. Subsequently, multidimensional within-item models were established. First, a multi-dimensional between-item model in which CR items were placed as a secondary dimension was established, and then a multi-dimensional between-item model in which MC and CR items were in one dimension (math ability) and CR items were in a second dimension was formed. These models were analyzed utilizing a multi-dimensional Rasch-type item response model. In the related literature, the advantages of using multi-dimensional models are reported as compensating for the weaknesses of the linear factor analysis, providing the opportunity to reveal or confirm dimensionality, and providing stronger evidence for the debate over the mixed-format with exploratory and confirmatory approaches (Wang, 2002). The analyses run for all three alternative models were realized using ConQuest 4 software (Adams, Wu, Macaskill, Haldane, & Sun, 2015). ConQuest offers three approximation methods for computing the integrals that must be computed in marginal maximum likelihood estimation (MML): quadrature (Bock/Aitken quadrature), gauss (Gauss-Hermite quadrature) and Monte Carlo. In the present study, gauss was employed as an estimation method because this method is generally the preferred approach for problems of three or fewer dimensions (Adams, et al., 2015, p.36).

The model data fit for all three models was analyzed by comparing the -2log likelihood-ratio, the G2 statistics test, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the adjusted Bayesian information criterion values obtained from each model.

Group Ability Differences: To predict the individuals' abilities in each dimension, the expected-a-posteriori (EAP) estimation method was utilized. In order to measure the impact of group variation on abilities, separate multiple regression analyses were run with data sets for 4th and 8th grade students. Among the variables, the only the "gender" variable was a categoric variable. Hence, by using the dummy coding structure for this variable this variable, gender was coded as 0 and 1 for males and females, respectively. All the other group variables used in the study are index scores and are continuous values.

Classification Analyses: With the aim of comparing classification analyses belonging to individuals' abilities obtained from unidimensional and multi-dimensional models, the obtained individual ability scores belonging to math ability were grouped into four based on mean scores and standard deviations. The following ability cut-points were used: (1) Poor: ability < –1 SD; (2) Low achieving: –1 SD ≤ ability < 0; (3) Proficient: 0 ≤ ability < +1.00 SD; (4) Highly proficient: +1 SD ≤ ability. The degree of agreement between the two classifications was then evaluated using Cohen's kappa coefficient.

3. Results

The model data fit values of the 4th and 8th grade students' levels of mathematics achievement in TIMSS 2015 obtained from the unidimensional, between-item and within-item models are presented in Table 1. It was found that in all the -2LL, AIC, BIC and adjusted BIC values for 4th Grade and 8th Grade, the within-item (bidimensional) model had a higher level of fit when compared to the other models. When the significance of the G^2 log-likelihood test and the -2LL chi-square values were examined, the within-item model was identified to have a significantly higher model data fit, with a significance level of .01, compared to the unidimensional model. The model data fit values for within-item and between-item models were very close to each other.

Table 1: *Model fit statistics for the unidimensional and multidimensional models*

	4th Grade			**8th Grade**		
Model	**Unidimensional Model**	**Between-item Model**	**Within-item Model**	**Unidimensional Model**	**Between-item Model**	**Within-item Model**
-2LL	12695.59	12673.97	12673.24	12913.08	12774.29	12774.09
Number of parameters	26	28	28	30	32	32
G^2 LR Test	χ^2 (2) = 33.65, p<.01			χ^2 (2) = 239.84, p<.01		
AIC	12747.59	12729.97	12729.24	12973.08	12838.29	12838.09
BIC	12764.87	12748.58	12747.85	12992.23	12858.72	12858.52
Adjusted BIC	12696.88	12675.26	12674.53	12914.34	12775.55	12775.35

Notes: LL = Log-likelihood; AIC = Akaike's information criterion; BIC = Bayesian information criterion.

Based on the item difficulty and fit statistics in Table 2, it was found that the minimum and mean values of item difficulty for 4th Grade were close to each other, while the maximum values of the unidimensional model were higher. It was also observed that the range of the item difficulty values for the unidimensional model for 8th Grade were higher than the item difficulty values obtained from the within-item model. According to Bond and Fox (2007), a statistical value of fit ranging between .7 and 1.3 indicates a positive fit. While there are outfit items with values above 1.3 in both models, the means of the fit values obtained from both models for 4th Grade and 8th Grade are within the value limits.

Table 2: *Item diffuculty estimates and fit statistics for two models*

		Unidimensional Model			**Within-item Model**		
		Difficulty	**Infit**	**Outfit**	**Difficulty**	**Infit**	**Outfit**
4th Grade	Minimum	-2.28	.79	,71	-2.17	.82	.67
	Maximum	2.85	1.21	1.87	1.49	1.29	2.38
	Mean	.02	1.00	1.06	-.03	1.00	1.09
	Std. Deviation	1.13	.12	.29	.92	.13	.35
8th Grade	Minimum	-.60	.66	.33	.54	.65	.28

	Maximum	3.31	1.27	1.46	1.96	1.37	1.71
	Mean	1.02	.98	.98	.69	1.00	1.00
	Std. Deviation	.98	.16	.30	.60	.14	.29

The mean value for ability predictions obtained from the unidimensional and multidimensional models were determined as .00. It was found that the ability predictions for the unidimensional model had a higher range and standard deviation. With the effect of CR ability, the difference between the minimum and maximum values of the ability predictions belonging to the within-item model and the unidimensional model for 8th Grade were predicted to be higher. The values for the CR ability, which was examined as a secondary feature, were between .76 and .81 for 4th Grade, and between 1.62 and 2.57 for 8th Grade. The range and standard deviation values of CR ability for 8th Grade were found to be higher.

Table 3: *Descriptive Statistics about Ablity Parameters*

	Models	**Min**	**Max**	**M**	**SD**
4th Grade	Unidimensional Model	-3.15	3.09	-.01	1.13
	Within-item Model: Primary Dimension (Math Ability)	-2.75	2.87	.00	1.00
	Within-item Model: Secondary Dimension (CR Ability)	-.76	.81	.00	.28
8th Grade	Unidimensional Model	-2.47	3.08	.00	1.10
	Within-item Model: Primary Dimension (Math Ability)	-1.93	2.82	.00	.87
	Within-item Model: Secondary Dimension (CR Ability)	-1.62	2.57	.00	.76

The standard values of the ability parameters obtained from the unidimensional and within-item models were categorized at four different levels, and the accuracy of the categorization was analyzed using Cohen's kappa statistic. The kappa value for 4th Grade was found to be .89. The percentage for the total amount of inaccurate categorizations for 4th Grade turned out to be 8.01%, 3.25% being higher than the unidimensional model and 4.76% being higher than the within-item model. The kappa value for 8th Grade was found to be .95. The percentage for the total amount of inaccurate categorizations for 8th Grade turned out to be 3.22%, 1.15% being higher than the unidimensional model and 2.07% being higher than the within-item model.

Table 4: *Classification Table*

			Unidimensional Model			
			Level-1	**Level-2**	**Level-3**	**Level-4**
Within-item Model	4th Grade	Level-1	67	0	0	0
			(14.50%)	(.00%)	(.00%)	(.00%)
		Level-2	19	133	0	0
			(4.11%)	(28.79%)	(.00%)	(.00%)
		Level-3	0	3	167	15
			(.00%)	(.65%)	(36.15%)	(3.25%)
		Level-4	0	0	0	58
			(.00%)	(.00%)	(.00%)	(12.55%)
	8th Grade	Level-1	64	0	0	0
			(14.71%)	(.00%)	(.00%)	(.00%)
		Level-2	3	168	5	0
			(.69%)	(38.62%)	(1.15%)	(.00%)
		Level-3	0	5	124	0
			(.00%)	(1.15%)	(28.51%)	(.00%)
		Level-4	0	0	1	65
			(.00%)	(.00%)	(0.23%)	(14.94%)

The correlation coefficients of the mathematics ability obtained from the within-item and unidimensional models were .999 and .995 for 4th Grade and 8th Grade, respectively. The reliability coefficients of the ability parameter in the unidimensional model are higher than in the other model. The reliability coefficients range between .83 - .86.

Table 5: *Correlation Table*

	Correlation between Primary Ability		**Reliability Coefficient**
4th Grade	Unidimensional model	.999	.86
	Within-item model		.85
8th Grade	Unidimensional model	.995	.85
	Within-item model		.83

The regression analysis findings across some variables that could have an impact on the primary and secondary abilities are presented in Table 6. In 4th Grade, the variables of sense of belonging and students' confidence in mathematics have an impact on both primary and secondary features at a significance degree of .01; however, despite being significant, the effect size is small. In 8th Grade, the variables of home resources for learning and students' confidence in mathematics have an impact at a significance degree of .01. The variable of sense of belonging is signficant at .05. All these values are significant, but the effect size is small. While in 4th Grade, a high sense of school belonging has a high effect on the scores obtained from the primary and secondary dimensions, in 8th Grade, a low sense of school belonging has a high effect on the scores obtained from the primary and secondary dimensions. Even though gender does not have an impact on both grade levels, it was found that the predictions of math and CR ability for female students were lower, when compared to those made for males. In 4th Grade, six independent variables account for 19.2% of the variance in math ability and 18.8% of the variance in CR ability. In 8th Grade, eight independent variables account for 33.4% of the variance in math ability and 33.2% of the variance in CR ability.

Table 6: *Multiple Regression*

4th Grade						
	Primary Dimension (Math Ability)			**Secondary Dimension (CR Ability)**		
	B	SE	β	B	SE	β
Gender	-.11	.09	-.06	-.03	.02	-.05
SSB	.08**	.03	.15	.02**	.01	.15
SB	.02	.02	.04	.00	.01	.03
SLM	.00	.03	.00	.00	.01	.00
EML	-.01	.03	-.02	.00	.01	-.02
SCM	.19**	.03	.38	.05**	.01	.37
8th Grade						
	Primary Dimension (Math)			**Secondary Dimension (CR Ability)**		
	B	SE	β	B	SE	β
Gender	-.04	.07	-.02	-.04	.06	-.03
HER	.12**	.02	.26	.11**	.02	.26
SSB	-.04*	.02	-.10	-.04*	.02	-.10
SB	.02	.02	.04	.02	.02	.04
SLM	-.02	.03	-.05	-.02	.03	-.05
EML	.02	.02	.04	.01	.02	.04
SCM	.20**	.02	.52	.17**	.02	.52
SVM	-.03	.02	-.06	-.02	.02	-.06

* $p < .05$ ** $p < .01$

4. Discussions

The fundamental aim of the present research study was to determine how dimensionality is affected by item types in cognitive tests in which various item types are used. With this aim, the TIMSS 2015 mathematics achievement tests for 4th Grade and 8th Grade were analyzed based on three different models: unidimensional, between-item and within-item models. When the model data fits were examined, the within-item model was found to yield better results for the mixed-format test. In the study, the item difficulty values, the item fit statistics, ability predictions, reliability coefficients, correlation values for ability predictions and categorization accuracy values obtained from the unidimensional and within-item models were also examined. However, as the statistics turned out to be very close to each other and the relationship between the math ability values obtained from the unidimensional model and those from the within-item model were perfect (.995 - .999), it can be concluded that the effect of CR ability on math basic ability is limited. The item fit values for the items in the unidimensional model display a high level of similarity with those for the items in the within-item model. There were outfit items in both models, which may have stemmed from the relatively small sample size used in the study (N<500). When these results are taken into consideration, it can be concluded that the hypothesis of the study – CR ability is a secondary dimension – was not completely verified. This conclusion shows consistency with the findings of a study conducted by Marengo et al. (2016) on 8th grade mathematics results reported by Italian National Institute for the Evaluation of the Education System (INVALSI). Similarly, in a study conducted by Wang (2002), whether or not format differences in mathematics assessment resulted in multidimensionality was examined and found that MC and CR items loaded highly on the same factors. In an older study by Traub and Fisher (1977), weak evidence was found to support that CR items are secondary dimensions, and hence, it was reported that there was a small effect of format in math assessments.

Studies conducted in areas other than maths, namely reading compehension (Ward et al., 1987), computer science (Bennett et al.,1991), analytical reasoning (Bridgeman & Rock, 1993), and chemistry (Thissen et al., 1994) reported that different item formats showed a higher level of fit when compared to the one-factor model. In relation to item format, Griffo (2010), who conducted a study on the NAEP Reading Assessment data, reported that two dimensional models yielded higher fit values than unidimensional models, but as there was a 93% correlation between MC and CR items, it was stated that there was no need for having CR items as a second dimension. Accordingly, the findings of the present study are consistent with those reported by other studies in the related literature. When the proficiency classifications of students based on the ability predictions obtained from the unidimensional model and the primary dimension of the within-item model are examined, by considering the high Kappa statistic and the low percentages of inaccurate classifications, it can be claimed that both models can make similar ability classifications. The high level of fit of the results can indicate that analyses can also be made without taking CR ability as a second dimension.

The effect sizes of the variables that are thought to have an impact on the predictions for individuals' ability levels were examined for both primary math ability and CR ability. Since the study implemented the TIMSS student questionnaire, it can be considered as a rich source of data. In this study for 4th grades, the variables of gender, sense of school belonging, student bullying, students like learning mathematics, students' views on engaging teaching in mathematics lessons, students confidence in mathematics and for 8th grades, all these variables together with home resources for learning and students value mathematics variables were addressed. It was determined that for 4th Grade, the variables of sense of school belonging and students' confidence in mathematics had a significant effect on both dimensions. As for 8th Grade, the variables of home resources for learning, sense of school belonging and students' confidence in mathematics were found to have a significant effect on primary and secondary ability. The fact that the same variables have an impact on both dimensions in both grade levels could again be deriving from the small effect size of the secondary dimension. The gender variable was found to have no effect on the dimensions. However, even though some studies reported that female students were more successful on CR items (Bible, Simkin, & Kuechler, 2008), other studies reported that the gender variable had no impact on the achievement levels of students in different item types (Bacon, 2003; Chan & Kennedy, 2002).

It can be recommended that the present study can be replicated by creating different simulation conditions. Moreover, dimensionality can be examined by using dimension determining techniques other than the ones used in the current study.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12*(2), 117-128. DOI: 10.1177/014662168801200202

Adams, R., & Wu, M. (2010). Multidimensional models. *ConQuest Tutorial*.

Adams, R., Wu, M., Macaskill, G., Haldane, S., & Xun Sun, X. (2015). *ConQuest*. University of California, Berkley: Australian Council for Educational Research.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36. DOI: 10.1177/0273475302250570

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, *75*(4), 348-357. DOI: 10.1111/j.1475-4932.1999.tb02571.x

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free- response and multiple- choice items. *Journal of Educational Measurement*, *28*(1), 77-92. DOI: 10.1111/j.1745-3984.1991.tb00345.x

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education: An International Journal*, *17 (Supplement)*, 55-68. DOI: 10.1080/09639280802009249

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*(4), 385-395. Retrieved from https://conservancy.umn.edu/bitstream/handle/11299/104073/v11n4p385.pdf?sequence=1 DOI: 10.1177/014662168701100404

Bond, T. G., & C. M. Fox. 2007. *Applying the Rasch model* (2nd Edition). Mahwah, N.J.: Lawrence Erlbaum Associates.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education*, *32*, 319-332. DOI: 10.1007/BF00992895

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, *30*, 313-329. DOI: 10.1111/j.1745-3984.1993.tb00429.x

Brown, G. A., Bull, J., & Pendlebury, M. (2013). *Assessing student learning in higher education*. Routledge.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*(1), 79-96. DOI: 10.1111/j.1745-3984.1995.tb00457.x

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets*. University of Maryland, College Park. Retrieved from https://drum.lib.umd.edu/bitstream/handle/1903/8843/umi-umd-5871.pdf;sequence=1

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957-971. DOI: 10.2307/1061503

Downing, S. M. (2006). Selected-response item formats in test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development*, (p. 287-301). London: Lawrence Erlbaum Associates.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Making sense of students' answers to multiple-choice questions. *The Physics Teacher*, *40*(3), 174-180. DOI: 10.1119/1.1466554

Embretson, S. E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and Scoring of Tests With Multiple- Choice and Constructed- Response Item Types. *Journal of Educational Measurement*, *35*(2), 137-154. DOI: 10.1111/j.1745-3984.1998.tb00531.x

Fenna, D. S. (2004) Assessment of foundation knowledge: are students confident in their ability? *European Journal of Engineering Education*, *29*(2), 307-312, DOI: 10.1080/0304379032000157277

Gessaroli, M. E., & Champlain, A. F. (2005). Test dimensionality: Assessment of. *Wiley StatsRef: Statistics Reference Online*. DOI: 10.1002/9781118445112.stat06371/full

Griffo, V. B. (2011). *Examining NAEP: The effect of ıtem format on struggling 4th graders' reading comprehension*. University of California, Berkeley.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333. DOI: 10.1207/S15324818AME1503_5

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of experimental education*, *62*(2), 143-157. DOI: 10.1080/00220973.1994.9943836

Hastedt, D. (2004). Differences between multiple-choice and constructed response items in PIRLS 2001. In *Proceedings of the IEA International Research Conference*.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*(4), 427-441. DOI: 10.1080/08957340701580736

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381. DOI: 10.1207/s15324818ame1904_7

Kuechler, W. L., & Simkin, M. G. (2010). Why ıs performance on multiple- choice tests and constructed- response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55-73. DOI:10.1111/j.1540-4609.2009.00243.x/full

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact. *Journal of Applied Testing Technology*, *13*(3), 1-50.

Lukhele, R., Thissen, D., & Wainer, H. (1993). On the relative value of multiple- choice, constructed- response, and examinee- selected items on two achievement tests. *ETS Research Report Series, Technıcal Report No. 93-28.* DOI: 10.1002/j.2333-8504.1993.tb01517.x

Marengo, D., Miceli, R., & Settanni, M. (2016). Do mixed ıtem formats threaten test unidimensionality? Results from a standardized math achievement test. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *23*(1), 25-36. DOI:10.4473/TPM23.1.2

Martin, M. O., Mullis, I. V. S., & Hooper M. (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207-218. DOI: 10.1207/s15326985ep3404_2

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7. Retrieved from http://www.jstor.org/stable/pdf/1176712.pdf https://doi.org/10.3102/0013189X018009003

Pollack, J. M., Rock, D. A., & Jenkins, F. (1992). Advantages and disadvantages of constructed-response item formats in large-scale surveys. İn *Annual Meeting of the American Educational Research Association, San Francisco*.

Saunders, P., & Walstad, W. B. (1998). Research on teaching college economics. İn *Teaching undergraduate economics: A handbook for instructors*. Boston, MA: İrwin/McGraw Hill, 141-166.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple- choice tests and student understanding: what is the connection?. *Decision Sciences Journal of İnnovative Education*, *3*(1), 73-98. Retrieved from http://www.coba.unr.edu/faculty/kuechler/cv/DSJ?E.3.1.05.pdf ; https://doi.org/10.1111/j.1540-4609.2005.00053.x

Smith, J. (2009). *Some issues in item response theory: Dimensionality assessment and models for guessing*. Unpublished Doctoral Dissertation. University of South California.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). Multidimensionality and the equating of a mixed-format math examination. Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002). Retrieved from https://files.eric.ed.gov/fulltext/ED469163.pdf

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple- choice and free- response items necessarily less unidimensional than multiple- choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113-123. Retrieved from http://www.jstor.org/stable/pdf/1435171.pdf ; https://doi.org/10.1111/j.1745-3984.1994.tb00437.x

Traub, R. E. & Fisher, C. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*(3), 355-369. DOI:10.1177/014662167700100304

Traub, R. E., & MacRury, K. A. (1990). *Multiple-choice vs. free-response in the testing of scholastic achievement*. Ontario Institute for Studies in Education.

Tuckman, B. W. (1993). The essay test: A look at the advantages and disadvantages. *Nassp Bulletin*, *77*(555), 20-26. DOI: 10.1177/019263659307755504

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118. DOI: 10.1207/s15324818ame0602_1

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275. Retrieved from http://www.jstor.org/stable/pdf/1435130.pdf; https://doi.org/10.1111/j.1745-3984.2003.tb01107.x

Wang, W. C. (1994). *Implementation and application of the multidimensional random coefficients multinomial logit model*. University of California, Berkeley.

Wang, Z. (2002). *Comparison of different item types in terms of latent trait in mathematics assessment.* Doctoral dissertation, University of British Columbia.

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). A comparıson of free- response and multiple- choice questions in the assessment of reading comprehension. *ETS Research Report Series*. DOI:10.1002/j.2330-8516.1987.tb00224.x/pdf

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1978). Construct validity of free- response and machine- scorable versions of a test of scientific thinking. *ETS Research Report Series, 1978*(2). DOI:10.1002/j.2333-8504.1978.tb01162.x/pdf

Yıldırım, A., Özgürlük, B., Parlak, B., Gönen, E., & Polat, M. (2016). TIMSS uluslararası matematik ve fen eğilimleri araştırması: TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. Sınıflar. Ankara: T.C. Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Zhang, L., & Manon, J. (2000) *Gender and achievement-understanding gender differences and similarities in mathematics assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, April 2000 (pp. 24-28). New Orleans, LA.