ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 238-253

A Diagnostic Comparison of Turkish and Korean Students' Mathematics Performances on the TIMSS 2011 Assessment*

TIMSS 2011 Sınavına Kore ve Türkiye'den Katılan Öğrencilerin Matematik Performanslarının Tanısal Bir Karşılaştırılması

Sedat ŞEN** Muhammet ARICAN***

Abstract

The purpose of the present study was to analyze an international large-scale data set using a cognitive assessment approach. Although some researchers question the usefulness of international large-scale assessments (e.g., TIMSS), participating countries have continued to use the results from these large-scale assessments to improve their curricula and teaching methods. Despite the common reporting practice—single-score—in these large scale assessments gives useful insights about students' overall performances, they still lack diagnostic information. Cognitive diagnosis models (CDMs) were developed to provide more feedback on students' cognitive strengths and weaknesses. This study retrofitted the TIMSS 2011 eighth grade mathematics assessment by applying a specific CDM called the DINA (the deterministic, inputs, noisy, "and" gate) model to data from South Korea and Turkey. Results of the DINA model were used to make a detailed comparison between students of these two countries.

Key words: Mathematics assessment, cognitive diagnosis, DINA model, TIMSS

Öz

Bu çalışmanın amacı büyük ölçekli bir sınavın tanılayıcı değerlendirme yaklaşımlarından biriyle analiz edilmesidir. Bazı araştırmacılar büyük ölçekli sınavların (örn: TIMSS) kullanışlılığını sorguluyor olsa da katılımcı ülkeler bu sınavlardan alınan sonuçları kullanarak müfredatlarında ve öğretim metotlarında geliştirmeler yapmaya devam etmektedir. Bu sınavlarda yaygın olarak kullanılan ve tek bir puan sunmaya dayalı olan uygulamalar öğrencinin genel performansı hakkında bilgi sunsa da tanısal bilgi sunmada yeterli değildir. Öğrencilerin bilişsel olarak güçlü ve zayıf yanlarıyla ilgili daha detaylı bilgi sunabilmek için bilişsel tanı modelleri geliştirilmiştir. Bu çalışmada Kore ve Türkiye veri setleri kullanılarak TIMSS 2011 sekizinci sınıf matematik sorularının bilişsel tanı modellerinden DINA model ile tekrar analizi yapılmıştır. Bu modelden elde edilen sonuçlar kullanılarak iki ülke öğrencilerinin performanslarının karşılaştırılması yapılmıştır.

Anahtar kelimeler: Matematik eğitimi, bilişsel tanı, DINA model, TIMSS

INTRODUCTION

Since the first administration of the Trends in International Mathematics and Science Study (TIMSS) in 1995, the comparison of the relative performances of participating countries has become very helpful for finding out country-level success relative to other countries. Although some researchers question the relationship between student-level (or country-level) achievement and comparison studies based on such international large-scale assessments (Holliday & Holliday, 2003; Wang, 2001),

* This study was presented at the 2015 annual meeting of the American Educational Research Association Conference, Chicago, IL, USA.

** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, e-mail: sedatsen@harran.edu.tr

*** Dr., Ahi Evran University, College of Education, Mathematics Education, Kırşehir-Turkey, e-mail: muhammetarican@gmail.com

Geliş Tarihi: 01.07.2015
Kabul Tarihi: 07.12.2015

participating countries have continued to use the results from these large-scale assessments (e.g., TIMSS) to improve their curricula and teaching methods to fill gaps or to reach excellence.

According to Toker and Green (2012), educational assessment is important since, by means of this assessment, educators evaluate the effects of educational programs and manage these programs. Because outcomes of education necessitate meeting a universally accepted criteria (Toker & Green, 2012), in mathematics education, researchers have been using international comparative studies (e.g., TIMSS, PISA) to evaluate students' achievements and their mastery of curricular instruction (Lee, Park, & Taylan, 2011). However, as discussed by Dogan and Tatsuoka (2008), since the reports on students' achievements and their mastery of curricular instruction rely on total scores and rankings of the participating countries, they do not provide enough information about students' strengths and weaknesses. The common reporting practice in these large scale assessments is to provide a single overall score for each student and report students' averages across their countries. Although the single test scores give useful insights about the overall performances in terms of subject areas, they still lack diagnostic information. The lack of the diagnosity of a single score based on test assessments has frustrated many researchers (Nichols, 2012). Hence, as Leighton and Gierl (2007) stated,

> There is increasing pressure to make assessments more informative about the mental processes they measure in students. In particular, there is increasing pressure to adapt costly large-scale assessments (Organization for Economic Co-operation and Development [OECD], 2004; U.S. Department of Education, 2004) to be informative about students' cognitive strengths and weaknesses. (p. 5)

In order to provide an example to show how diagnostic feedback can be given using real data, this study analyzes TIMSS 2011 data from a cognitive diagnostic assessment (CDA; Leighton & Gierl, 2007) perspective. Over the last two decades, the interest in CDA has increased in order to obtain more information about students' performances on a measurement. This type of assessment classifies students based on their degrees of mastery of specific skills. Thus, examiners and instructors can obtain more information relevant to classroom teaching and learning. Unlike a single-overall test score, CDA-based reports simply show what students know (master) and what they do not know (master) rather than how much they know.

The main purpose of this study is to examine Turkish eight graders' strengths and weaknesses on topics that were covered on the TIMSS 2011 mathematics achievement test. In order to do so, in this study, the relative performances of Turkish students in comparison with South Korean (Korea hereafter) students were assessed. Hence, this CDA-based study examines the following research questions:

1. How do Turkish and Korean eight graders' relative TIMSS 2011 mathematics performances differ?
2. What are the Turkish eight grade students' weaknesses and strengths on TIMSS 2011's mathematics topics in comparison to the Korean eight graders?

Literature Review

Several recent studies (e.g., Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee et al., 2011; Toker & Green, 2012; Lee et al., 2013) have been conducted to compare students' achievements on international large-scale assessments (e.g., TIMSS, PIRLS) using DCMs. These studies have provided useful feedback on the students' performance and skills, the linkage between teachers' instruction and students' performances, and the countries' educational systems and their curricular instructions. For instance, Dogan and Tatsuoka (2008) compared Turkish and American eight-grade students' mathematics performances on the TIMMS-R 1999. Their results indicated that Turkish students were weak in algebra and probability/statistics in comparison to their American peers, and they also "demonstrated poor profiles in skills such as applying rules in algebra, approximation/estimation, solving open-ended problems, recognizing patterns and relationships, and quantitative reading" (Dogan & Tatsuoka, 2008, p. 263). Similarly, Im and Park (2010) compared Korean and American eight-grade students' mathematics performances on the TIMMS 2003. The results showed significant differences in the performances of Korean and American students, especially in "problem restructuring and reasoning, measurement, and geometry" (p. 287). Their results suggested that encouraging students' independent problem solving was the most useful instructional strategy for both Korean and American students. Moreover, American students benefitted from reviewing, re-teaching, and clarifying as well. In addition to the above studies, Lee et al. (2011) compared the performances of fourth-grade students' in Massachusetts and Minnesota to the nationwide results (not including MA and MN) on the TIMSS 2007. Their results demonstrated that students in Massachusetts and Minnesota outperformed students in the US overall. Lee et al. (2011) also provided fine-grained diagnostic information on students' performances, which they suggest could be exactly applied to classroom instruction. For example, by analyzing item parameter estimates (e.g., slipping and guessing) they offered curricular suggestions to the classroom teachers on how to improve students' performances.

In this study, Korea was chosen as a reference country, because Korean eight graders have been regularly placed in the top three in TIMSS mathematics performance. As stated by Mullis, Martin, Foy, and Arrora (2012), 42 countries and 14 benchmarking entities participated in TIMSS 2011. In that assessment, the international TIMSS scale average was set to 500. Among 42 countries, Turkish students had an average score of 452 and were ranked in 24th place. Korean students had an average score of 613 and were ranked in first place on the TIMSS 2011. As explained by Im and Park (2010), several studies investigated which characteristics of Korean education have been contributing to such tremendous performance in mathematics. According to Im and Park (2010), the results of those studies pointed out that factors contributing to Korean students' high achievement could be grouped under social and instructional factors. Social factors included "competitive examination and selection, a regular and metric number system, the serious attitudes of students towards tests, meaningful repetitive learning, and the competence of mathematics teachers (Kim et al., 2008; Park, 2004)" (Im & Park, 2010, p. 288), and instructional factors included "cooperative learning activities (Chung & Son, 2000; House, 2009), the use of constructivist strategies (Fisher & Kim, 1999), and teachers' guidance (Oh, 2005)" (Im & Park, 2010, p. 288). These social and instructional factors also affected our decision to select Korea as the reference country.

Diagnostic Classification Models

A number of cognitive diagnosis models (CDMs), also known as diagnostic classification models (DCMs), have been developed (Rupp, Templin, & Henson, 2010) to apply the CDA approach. For an overview of DCMs, the reader is referred to DiBello, Roussos, and Stout (2007), Fu and Li (2007), Rupp and Templin (2008a), and Rupp et al. (2010). However, it should be noted de la Torre (2011) classified these psychometric models as either general or a specific type based on their characteristics. Specific DCMs include: d*eterministic inputs, noisy "and" gate* (DINA; Haertel, 1989; de la Torre, 2009; Junker & Sijtsma, 2001), *deterministic inputs, noisy* "*or*" *gate* (DINO; Templin & Henson, 2006), *noisy-input, deterministic "and" gate* (NIDA; Junker & Sijtsma, 2001), and the *reduced reparameterized unified model* (R-RUM; Hartz, 2002; Roussos et al., 2007). General DCMs include the log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009), the general diagnostic model (GDM; von Davier, 2005), and the generalized DINA (G-DINA; de la Torre, 2011) model. This study focused on the DINA model. Thus, a brief description of the DINA model is presented below.

The DINA Model

The DINA model is a non-compensatory model with a conjunctive rule (Rupp et al., 2010). Based on the conjunctive nature of the DINA model, a respondent has to master all of the measured attributes of an item in order to get full credit for this item. Respondents get zero credit for an item if they did not master at least one of the measured attributes of this item. Thus, the DINA model divides respondents into two groups for each item: those who mastered all attributes and those who did not master all attributes. This is done with the conjunctive kernel of the DINA model, which is presented as a latent response vector (ξ_{ri}) below (Equation 1). Let X_{ri} be the response of examinee *r* to item *i*, and let

$\boldsymbol{\alpha}_r = \{\alpha_{rk}\}$ be the examinee's binary attributes vector, which is coded as 1 for presence or mastery of attribute *k* on the *k*th element and zero otherwise. Like most of the CDMs, the DINA model requires a Q-matrix (Tatsuoka, 1985) that shows the relationship among items (*i*,...,*I*) and attributes (*k*,...,*K*). A value of 1 for the Q-matrix entry (i.e., $q_{ik} = 1$) indicates that attribute *k* is measured for item *i*. For example, suppose we measure four attributes in an arithmetic test. Let addition, subtraction, division, and multiplication be four attributes coded as Attribute 1, Attribute 2, Attribute 3, and Attribute 4, respectively. Based on this attribute list and the DINA model specification, students have to master both Attribute 1 (addition) and Attribute 3 (division) in order to get full credit ($X_{ri} = 1$) for an item such as $\frac{4+8}{3}$. A student with mastery of addition or division cannot get full credit ($X_{ri} = 0$), as he/she would miss one of the required attributes for this item. The conjunctive kernel of the DINA model can be presented as below:

$$\xi_{ri} = \prod_{k=1}^{K} \alpha_{rk}^{q_{ik}} \qquad \text{(Equation 1),}$$

where ξ_{ri} is the latent variable which is coded as zero or one for respondent *r* and item *i*, and q_{ik} is the Q-matrix entry described above. α_{rk} represents the latent attribute variable indicating whether respondent *r* has mastered attribute *k* ($\alpha_{rk} = 1$) or not ($\alpha_{rk} = 0$). Thus, the latent response vector (ξ_{ri}) can have a value of 1 if respondent *r* masters all the attributes required for item *i* and a value of 0 if the respondent did not master at least one of the measured attributes for item *i*. It is possible that respondents who have mastered all attributes can give a wrong answer to item *i*, while respondents who have missed one of the required attributes can correctly answer item *i*. The former refers to slipping, and the latter refers to a guessing situation in the DINA model specifications. Thus, two parameters are obtained for each item in the DINA model regardless of the number of attributes. Item slipping (s_i) and guessing (g_i) parameters do not change across attributes, because they are item-specific. In the DINA model, these two item parameters are defined as follows:

$$s_i = P(X_{ri} = 0 | \xi_{ri} = 1) \qquad \text{(Equation 2),}$$

$$g_i = P(X_{ri} = 1 | \xi_{ri} = 0) \qquad \text{(Equation 3).}$$

After defining slipping and guessing parameters, the probability of the correct response of a respondent in latent class *c* for item *i* can be computed as below:

$$P(X_{ri} = 1 | \xi_{ri}) = (1 - s_i)^{\xi_{ri}} {g_i}^{1-\xi_{ri}} \qquad \text{(Equation 4).}$$

According to Equation 4, respondents need to master all attributes measured by an item in order to answer this item correctly. DINA model was used in this study, because the DINA model requires an estimation of two parameters for each item, and the number of attributes does not affect the number of estimated parameters in the DINA model. The DINA model is also an appropriate model for equally important items like TIMSS items. The DINA model has been used in analyses of the TIMSS data by several authors, including Lee et al., (2011) and Choi, Lee, and Park (2015).

METHOD

Subjects and Data

Data sets from the students of two countries (i.e., Korea and Turkey) were compared in this study. Data were taken from the TIMSS 2011 eighth grade mathematics test, which included 28 blocks (14 science and 14 mathematics) and 14 test booklets. Each booklet was composed of four blocks of items: two mathematics and two science blocks. Students responded to different types of questions including multiple-choice (four response options) and constructed responses assessing four content domains: Number (30%); Algebra (30%); Geometry (20%); and Data and Chance (20%). According to the

TIMSS 2011 design, only six of the 14 mathematics assessment blocks were made publicly available. Based on the pairs of released blocks, only four booklets (Booklets 1, 2, 5, and 6) can be obtained for an eighth grade mathematics assessment as administered in the real exam settings. Booklet sample sizes for Korea and Turkey and the number of items for different content domains are presented in Table 1. Each booklet showed different distributions for content domains. The administration of Booklet 2 to Korean and Turkish students was selected for the DINA model analyses in this study due to the following reasons: (a) there were relatively more topics—13—in Booklet 2; (b) the subject areas of the items were distributed evenly—nine items for Numbers, nine items for Algebra, seven items for Data and Chance, and seven items for Geometry; and (c) the cognitive domains among the items were also distributed evenly—10 items required knowing, 13 items required applying, and nine items required reasoning. Booklet 2 was composed of Block 2 and Block 3 with 32 items, including 15 multiple choice and 17 constructed response items. There were 368 Korean students and 488 Turkish students who had taken Booklet 2.

Table 1. Descriptive Characteristics of the TIMSS 2011Mathematics Booklets

Booklets	Blocks	Turkey (*N*)	Korea (*N*)	Number	Algebra	Geometry	Data and Science
Booklet1	M01-M02	503	410	8	9	5	4
Booklet2	M02-M03	488	368	9	9	7	7
Booklet5	M05-M06	490	369	7	9	10	6
Booklet6	M06-M07	494	361	5	12	8	8

Construction of Q-Matrix

Attributes, which are used to define skills required to solve a specific item, were adopted from the Common Core State Standards for Mathematics (CCSSM; Common Core State Standards Initiative, 2010). The CCSSM was developed as a result of recognizing the need for a more focused and coherent mathematics curriculum in the United States to improve the quality of mathematics education and to increase mathematics achievement to the level of high-performing countries (Common Core State Standards Initiative, 2010). Therefore, standards from high-performing countries played a significant role in the development of the CCSSM (Common Core State Standards Initiative, 2014). Thus, in this study, the CCSSM was used to determine our attributes. By means of carefully examining TIMSS items and the standards, a list of 13 attributes (see Table 2) was created. In order to generate attributes that cover all possible skills, some of the two related standards were combined and separated with semi-colons. Using the attribute list in Table 2, 32 items were coded independently by four doctoral students with advance degrees in mathematics education at one large public university in the Southeast. An attribute was included in our Q-matrix if at least two coders agreed that an item measured that attribute (see Table 3).

The attributes in the Q-matrix are independently generated by considering the required steps to solve each item. For example, in Item 6, students were given a picture of a rectangular garden that had a (*x + 4*)-meter width and an *x*-meter height (see Figure 1). The garden consisted of two small rectangular gardens and one rectangular path. The path was 1 meter wide and was between the two small gardens. Students were asked to calculate the total area of the two small rectangular gardens, which were shaded, in m^2. In order to solve this problem, students need to master three attributes (Attributes 4, 5, and 11). First, they must understand the concept of area and relate area to multiplication—Attribute 11. Second, they need to multiply width and height for the big rectangular garden and for the rectangular path to calculate their areas. These two multiplication operations involve using algebraic expressions and require applying previous knowledge of arithmetic to algebra—Attribute 4. Third, they must know the distribution property, which also requires applying previous knowledge of arithmetic to algebra, and understand that the equivalent expressions of $x * (x + 4)$ and $x * 1$ are $x^2 + 4x$ and x—Attribute 5. In the last step, they can obtain the area of the shaded garden as $(x^2 + 3x)\,\mathrm{m}^2$ by subtracting *x* from $x^2 + 4x$. This last step also requires mastery of Attribute 4, since students who master Attribute 4 can apply arithmetic operations to algebraic equations. A student can solve this problem also by subtracting 1-meter from (*x + 4*)-meter and multiplying (*x + 3*)-meter by *x*-meter. Students also need to master Attributes 4, 5, and 11 to use this method. Note that one item

(Item M052503A) was dropped when constructing the Q-matrix, because Item M052503A and Item M052503B were identical in the original 32-item list. Thus, only 31 items were used to create our Q-matrix (see Table 3).

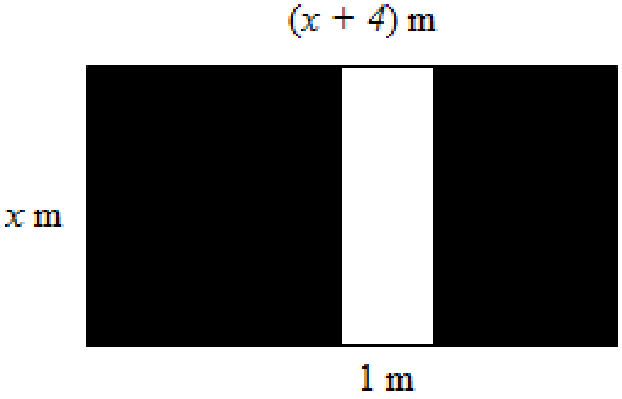

Figure 1. Item 6 (M052173) From the TIMSS 2011 Fourth Grade Mathematics

Table 2. Attributes Adopted from the Common Core State Standards Initiative (2010)

Content domain	Attribute description	Frequency
Numbers	A1-Possesses understanding of fraction equivalence and ordering; uses equivalent fractions as a strategy to add and subtract fractions.	5
	A2-Understands decimal notation for fractions, and compares decimal fractions; performs operations with decimals.	5
	A3-Understands ratio concepts, and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100.	4
Algebra	A4-Applies and extends previous understandings of arithmetic to algebraic expressions; solves real-life and mathematical problems using numerical and algebraic expressions and equations.	8
	A5-Reasons about and solves one-variable equations and inequalities; uses properties of operations to generate equivalent expressions.	4
	A6-Analyzes and solves linear equations and pairs of simultaneous linear equations.	1
	A7-Uses the four operations with whole numbers to solve problems; identifies and explains patterns in arithmetic.	3
Geometry	A8-Draws, constructs, and describes geometrical figures, and describes the relationships between them.	6
	A9-Solves real-life and mathematical problems involving angle measure, area, surface area, and volume.	5
	A10-Understands congruence and similarity using physical models, transparencies, or geometry software.	3
	A11-Recognizes perimeter, understands concepts of area, and relates area to multiplication and addition.	2
Data and Chance	A12-Represents and interprets data; draws informal comparative inferences about two populations.	3
	A13-Investigates chance processes and develops, uses, and evaluates probability models.	4

Data Analysis

As outlined in the TIMSS 2011 assessment framework, the TIMSS items were assessed using a three-parameter logistic item response theory (3PL IRT) model. This comparative study attempted to analyze TIMSS data sets for Korea and Turkey using a DINA model in order to present an application of a CDA-based analysis. As de la Torre and Lee (2008) showed, the results of the DINA model are consistent with that of the IRT models for the same data.

Table 3. Q-Matrix for the Eighth Grade TIMSS Mathematics Test

		Attributes												
Item	Item ID	1	2	3	4	5	6	7	8	9	10	11	12	13
1	M052216	1	1	0	0	0	0	0	0	0	0	0	0	0
2	M052231	0	1	0	0	0	0	0	0	0	0	0	0	0
3	M052061	0	0	0	1	0	0	1	0	0	0	0	0	0
4	M052228	1	0	0	0	0	0	0	0	0	0	0	0	0
5	M052214	1	1	1	0	0	0	0	0	0	0	0	0	0
6	M052173	0	0	0	1	1	0	0	0	0	0	1	0	0
7	M052302	0	0	0	1	0	0	0	0	0	0	0	0	0
8	M052002	0	0	0	1	1	0	0	0	0	0	0	0	0
9	M052362	0	0	0	0	0	0	0	1	1	1	0	0	0
10	M052408	0	0	0	0	0	0	0	1	1	1	0	0	0
11	M052084	0	0	0	0	0	0	0	0	1	0	1	0	0
12	M052206	0	0	0	0	0	0	0	1	1	0	0	0	0
13	M052429	0	0	0	0	0	0	0	0	0	0	0	0	1
14	M052503B	0	0	0	0	0	0	0	0	0	0	0	1	0
15	M042032	1	1	0	0	0	0	0	0	0	0	0	0	0
16	M042031	1	1	0	0	0	0	0	0	0	0	0	0	0
17	M042186	0	0	0	0	0	0	1	0	0	0	0	0	0
18	M042059	0	0	1	0	0	0	0	0	0	0	0	0	1
19	M042236	0	0	0	1	1	0	0	0	0	0	0	0	0
20	M042226	0	0	0	1	0	0	0	0	0	0	0	0	0
21	M042103	0	0	0	1	1	0	0	0	0	0	0	0	0
22	M042086	0	0	0	1	0	0	0	0	0	0	0	0	0
23	M042228	0	0	0	0	0	0	1	0	0	0	0	0	0
24	M042245	0	0	0	0	0	1	0	0	0	0	0	0	0
25	M042270	0	0	0	0	0	0	0	1	0	0	0	0	0
26	M042201	0	0	0	0	0	0	0	1	1	0	0	0	0
27	M042152	0	0	0	0	0	0	0	1	0	1	0	0	0
28	M042269	0	0	0	0	0	0	0	0	0	0	0	1	0
29	M042179	0	0	0	0	0	0	0	0	0	0	0	0	1
30	M042177	0	0	1	0	0	0	0	0	0	0	0	0	1
31	M042207	0	0	1	0	0	0	0	0	0	0	0	1	0

In addition to responses from Korean and Turkish students to the TIMSS eight grade mathematics assessment, the attributes (see Table 2) and Q-matrix (see Table 3) were also inputted into a DINA model. Since the TIMSS mathematics items included multiple choice and constructed responses, we dichotomized (0 = wrong answer, 1 = correct answer) those items for use with the dichotomous DINA model in this study. The DINA model parameters were estimated using maximum likelihood estimation with an expectation-maximization (EM) algorithm. All analyses were conducted using an object-oriented software package called OxEdit (Doornik, 2003) in order to obtain DINA model estimations using expectation-maximization (EM) algorithm. This program was chosen for analyses because it was a free software unlike other commercial software packages. The codes for the DINA model were requested from de la Torre (personal communication, February, 2014). The results of the two countries were compared in order to identify the weaknesses and strengths of the students of each country. Item parameter estimates and attribute mastery prevalence estimates are presented in the Results section. In addition, 3PL IRT model estimations were obtained using maximum likelihood estimation method for comparison purpose.

RESULTS

As presented above, the DINA model provides one slipping and one guessing parameter per item. These two parameters are equal across attributes. The DINA-based discrimination index (de la Torre, 2008) can also be calculated using slipping and guessing parameters for each item (i.e., $\delta = 1 - g - s$). The item discrimination index refers to the probability of correctly solving an item without the effect of guessing and slipping parameters. Put differently, it is the difference in probabilities of a correct response between $\xi = 0$ and $\xi = 1$. Slipping, guessing and discrimination parameter estimates for Korean and Turkish samples are presented in Table 6. Sixty-two item parameter estimates (31 guessing and 31 slipping parameters) were obtained for each sample. In addition to item parameters, in total 2^{13} = 8,192 attribute profile parameters were estimated for the 13 attributes listed in Table 2. Fit statistics for DINA model analyses are presented in Table 4. Since IEA (The International Association for the Evaluation of Educational Achievement) used 3PL IRT model for TIMSS analyses, results of 3PL IRT model were also provided for two samples before presenting main DINA model results (see Table 5).

Table 4. Fit Statistics for DINA Model Analyses

Country	Log-Likelihood	AIC	BIC
Korea	-4201.82	24909.65	57163.05
Turkey	-7552.31	31610.63	66193.30

Note. AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion.

Item Parameters

Table 6 presents item parameter estimates for slipping, guessing and the discrimination index for both countries. The small slipping and guessing parameter estimates indicate that examinees who master the measured attributes are able to apply the attributes correctly. As shown in Table 6, Items 4, 24, and 25 (the three with the lowest guessing and slipping parameter estimates) are the most informative items for Korean and Turkish samples. For example, for a Korean respondent who mastered Attribute 1, there is less than a 1% chance (s_4 = .009) that Item 4 is answered incorrectly. In contrast, a respondent who has not mastered Attribute 1 has no chance (g_4 = .000) of answering this item correctly. On the other hand, a Korean student has a 93% chance of answering Item 15 correctly even if he/she lacks at least one of two attributes (i.e., Attribute 1 or Attribute 2). It is desirable for a DINA model to have small guessing and slipping parameter estimates for a good model-data fit (Rupp et al., 2010). Higher values of item guessing and slipping parameters could be an indication of item-specific model misfit (Rupp & Templin, 2008b). DINA model item parameter estimates with high guessing values can be an indication item-specific misfit for Items 1, 7, 9, 15, 19, 29 and 31 in Korean data set while high slipping parameter estimates indicates possible misfits for Items 12, 14 and 21 in Turkey data set.

The mean values for item guessing, slipping parameters and the discrimination index are presented in the last row of Table 4. As can be seen in Table 6, Korean students had higher guessing parameter estimates and lower slipping parameter estimates than Turkish students for most of the items. The mean item discrimination index for the Korean sample was lower ($\bar{\delta}$ = .525) than that ($\bar{\delta}$ = .619) for the Turkish sample (see Table 6). Both samples had high discrimination indices for most of the items. A high discrimination index indicates a greater difference of probabilities of correct responses between $\xi = 0$ and $\xi = 1$. For most of the items, the item discrimination index was lower for the Korean sample than for the Turkish sample due to the higher guessing parameter estimates for the Korean sample. Among the 31 items, Items 1 (requires Attributes 1 and 2; Numbers), and 15 (requires Attributes 1 and 2; Numbers) had the lowest discrimination indices for the Korean sample due to their high guessing and low slipping parameter estimates. It should be noted that the item discrimination index for Item 24 was found to be very high (.999) for both the Korean and Turkish samples, indicating that Item 24 was very informative. This item appeared to discriminate probabilities of correct responses between $\xi = 0$ and $\xi = 1$ very well.

Attribute Probability and Attribute Prevalence

In addition to item parameter estimates, the DINA model provides respondent parameters estimates (attribute probability and attribute prevalence). Attribute probability assigns respondents to any of the *C* (2^A where *A* denotes the number of attributes) latent classes. As mentioned above, 8,192 classes exist for 13 attributes in our TIMSS example. The attribute prevalence estimate is obtained by summing the probabilities across all latent classes requiring that specific attribute. Attribute prevalence estimates are presented in Table 7 for the Korean and Turkish samples. For all of the 13 attributes, the

Korean sample had a higher attribute prevalence than the Turkish sample (see Table 7). These results indicate that Korean students are more likely to master all of the attributes. The probability of Turkish students mastering some attributes is also high (e.g., Attributes 3 and 11). Attribute 6 had the lowest probability value for the Turkish sample (.320) and the Korean sample (.609). Thus, Attribute 6, analyzes and solves linear equations and pairs of simultaneous linear equations, was difficult to master by eighth grade students. Besides Attribute 6, Turkish students also had difficulty in mastering Attributes 13, 7, 4, and 1.

Table 5. 3PL IRT Model Item Parameter Estimates for the Korean and Turkish Samples

	Korea			Turkey		
	Guessing	Difficulty	Discrimination	Guessing	Difficulty	Discrimination
Item 1	0.000	-2.560	1.928	0.252	0.888	2.629
Item 2	0.000	-2.350	1.092	0.002	0.222	1.120
Item 3	0.368	-0.240	2.378	0.027	1.111	1.690
Item 4	0.214	-1.349	1.654	0.111	0.882	3.239
Item 5	0.110	-0.949	0.878	0.269	1.436	4.200
Item 6	0.066	0.199	4.025	0.101	1.603	5.450
Item 7	0.000	-2.296	1.357	0.000	-0.365	1.602
Item 8	0.000	0.265	2.131	0.006	1.414	4.495
Item 9	0.000	-2.004	1.695	0.000	1.006	1.511
Item 10	0.356	-0.851	3.366	0.040	0.677	1.895
Item 11	0.000	-1.194	1.898	0.173	0.619	2.461
Item 12	0.109	-0.375	1.901	0.045	1.458	3.580
Item 13	0.000	-1.143	1.710	0.193	0.779	4.996
Item 14	0.275	-0.922	1.531	0.000	1.396	1.056
Item 15	0.950	-0.054	6.516	0.194	0.237	2.398
Item 16	0.478	-0.767	4.488	0.259	1.065	3.458
Item 17	0.000	-0.838	1.742	0.000	0.767	1.724
Item 18	0.000	-0.744	2.426	0.032	0.850	2.228
Item 19	0.400	-0.828	4.344	0.264	0.765	3.080
Item 20	0.000	-1.068	3.072	0.000	0.694	3.590
Item 21	0.000	-0.260	2.348	0.000	1.789	2.503
Item 22	0.000	-0.640	2.401	0.000	0.805	2.613
Item 23	0.386	-0.495	1.743	0.011	0.179	1.981
Item 24	0.000	-0.355	2.531	0.151	1.190	3.012
Item 25	0.000	-1.412	2.025	0.000	0.588	1.714
Item 26	0.000	-0.906	2.549	0.085	0.854	4.351
Item 27	0.441	-0.662	1.804	0.295	2.008	2.471
Item 28	0.000	-1.761	1.137	0.351	0.877	1.177
Item 29	0.624	-1.124	2.536	0.000	-0.335	1.359
Item 30	0.147	-0.972	2.064	0.112	0.276	1.626
Item 31	0.344	-1.085	1.979	0.000	0.539	1.180

By means of examining the attribute prevalence estimates (see Table 7), it was concluded that Turkish students were particularly weak in mastering Attributes 1, 8, and 13 when compared to their Korean peers. These three attributes had the highest prevalence estimate differences for Korea and Turkey. Hence, while most of the Korean students mastered these three attributes, many Turkish students had difficulty in mastering them. On the contrary, Attributes 3, 11, and 5 had the lowest prevalence estimate differences for Korea and Turkey. That means the probability of Turkish students' mastery of those three attributes were close enough to the probability of Korean students' mastery. However, it should be noted that these three lowest prevalence estimate differences mainly occurred because of the increments on the probability of the Turkish students' mastery on those attributes, not because of the decrements on the probability of the Korean students' mastery.

Table 6. Item Parameter Estimates for the Korean and Turkish Samples

	Korea			Turkey		
Item	Guessing	Slipping	Discrimination	Guessing	Slipping	Discrimination
1	.878	.004	.118	.306	.121	.573
2	.586	.034	.381	.144	.134	.722
3	.470	.113	.417	.115	.347	.538
4	.000	.009	.991	.000	.107	.893
5	.517	.187	.297	.269	.323	.408
6	.087	.209	.704	.086	.411	.503
7	.778	.027	.195	.441	.077	.482
8	.031	.317	.652	.011	.384	.605
9	.768	.009	.224	.151	.234	.615
10	.581	.000	.419	.203	.018	.779
11	.470	.024	.507	.190	.031	.779
12	.232	.184	.583	.044	.526	.430
13	.186	.055	.759	.172	.115	.714
14	.365	.024	.611	.079	.540	.381
15	.928	.007	.065	.442	.078	.480
16	.527	.000	.473	.265	.146	.589
17	.218	.067	.715	.072	.210	.718
18	.248	.025	.727	.113	.210	.678
19	.682	.000	.319	.350	.000	.651
20	.462	.032	.506	.021	.197	.782
21	.177	.137	.685	.006	.557	.437
22	.138	.078	.784	.030	.265	.706
23	.475	.099	.426	.266	.116	.618
24	.000	.001	.999	.000	.001	.999
25	.003	.033	.964	.007	.160	.832
26	.243	.016	.741	.103	.137	.760
27	.419	.043	.538	.262	.421	.317
28	.521	.044	.435	.340	.161	.499
29	.719	.011	.270	.443	.088	.469
30	.504	.041	.455	.357	.101	.542
31	.668	.011	.321	.172	.127	.701
Mean	.415	.059	.525	.176	.204	.619

The top five highest attribute class profiles with the highest probability estimates are presented in Table 8. These classes with highest probabilities were selected from 8,192 possible latent classes. The probability estimates listed in Table 8 can be interpreted as percentages, as the sum of probabilities for 8,192 different latent class profiles are equal to unity. For example, a probability value of .013 for a latent class profile indicates that only 13% of the respondents were assigned to this specific latent class. As shown in Table 8, 44% of Korean students mastered all of the attributes (attribute class of 1111111111111), while only almost 13% of Turkish students mastered all of the attributes. Other latent class profiles with the highest probability values showed that Attributes 5, 6, 7, and 12 were difficult to master for Korean students (see bolded zeros in Table 8). Less than one percent (p = .0016) of Korean respondents appeared to master none of the attributes (attribute class of 0000000000000). The second largest latent class for Turkish students was the mastery of all attributes except for Attributes 5 and 6. The posterior probability of this latent class profile was .026. Therefore, Attribute 6 appeared to be difficult to master by Turkish students as it was not mastered by most of the Turkish students (see bolded zeros for Attribute 6 in Table 8). Furthermore, 1.0% of the Turkish students could not master any of the attributes, while another 1.0% of Turkish sample only mastered Attribute 3 (understands ratio concepts and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100).

Table 7. Estimates of Attribute Prevalence

	Attribute Prevalence	
Attribute	Korea	Turkey
1	0.866	0.392
2	0.803	0.464
3	0.753	0.611
4	0.708	0.382
5	0.728	0.522
6	0.609	0.320
7	0.702	0.361
8	0.882	0.445
9	0.768	0.543
10	0.806	0.543
11	0.768	0.581
12	0.714	0.462
13	0.790	0.354

Table 8. *Top Five Attribute Class Profiles for the Korean and Turkish Samples*

Korea		Turkey	
Attribute Profile	Probability	Attribute Profile	Probability
1111111111111	0.443	1111111111111	0.128
11111**0**1111111	0.039	1111**00**1111111	0.026
11111111111**0**1	0.019	11111**0**1111111	0.017
1111**000**111111	0.012	**00**1**0000000000**	0.010
1111**000**1111**0**1	0.013	**0000000000000**	0.010

DISCUSSION

This study showed the application of a CDA-based assessment for a large-scale test data set, which has been originally analyzed with a traditional IRT model (3PL). CDM approach was selected, because it is possible to report a more detailed evaluation of students' performances on specific skills. Korea (the top performing country) and Turkey (the focus of the study) were selected for analyses in this study to show how a DINA model can be used to obtain fine-grained information about the performances of the students from these two countries. There are several advantages of the DINA model over traditional IRT models. For example, analyses based on IRT models provide a single overall score based on invariant item and ability parameters. Unlike IRT models, CDMs (e.g., the DINA model) are used to obtain qualitative information in addition to quantitative information. The qualitative part of the CDMs comes from a latent class based structure. Using this property, it was tried to show which skill profiles both Korean and Turkish students were assigned. This specific respondent information could be very useful for instructors and educational policy makers for demonstrating the mastery of each student on each attribute, which is important feedback for instructors. In addition to attribute masteries, a number of item parameter estimates can be estimated with DCMs, like item guessing, slipping, and discrimination parameters.

The results of the DINA model for the Korea and Turkey data sets provided different patterns for the strengths and weaknesses of the two countries. As in the original 3PL IRT analysis, the Korean sample showed a higher performance than the Turkish sample in this study. As expected, the posterior probability of mastering all of the attributes (i.e., 1111111111111) for Korean students was higher than that of Turkish students. Additionally, one percent of the Turkish sample mastered none of the attributes, while this percentage was less than one percent for the Korean sample. In addition, six percent of Turkish respondents mastered only one of the thirteen attributes. These findings were very crucial for diagnosing the most problematic attributes (or skills) for the Turkish sample. Another attribute related finding showed that attribute prevalence estimates were higher than .70 for all items except for Attribute 6 in the Korean sample. However, all of the attribute prevalence estimates were less than .70 for the Turkish sample.

As a result of examining the estimates provided in Table 7, it was decided that Turkish students had difficulties mastering Attributes 4, 6, and 7. Because these three attributes were classified in the Algebra content domain, it was suggested that Turkish educators should pay more attention to eight graders' understanding of Algebra topics. They should especially focus on students' understanding of analyzing and solving linear equations and applying previous understandings of arithmetic to algebraic expressions. This result was consistent with the findings from Dogan and Tatsuoka (2008) who also stated Turkish students' weaknesses in algebra content domain when compared to American students. Furthermore, when compared to their Korean peers, Turkish students were particularly weak in mastering Attributes 1, 8, and 13. These three attributes had the highest prevalence estimate differences for Korea and Turkey. Hence, while most of the Korean students mastered these three attributes, many Turkish students had difficulties in mastering them. The items in which the mastery of Attribute 1 was required were all fractions and decimals items. Therefore, the results indicate Turkish students' weaknesses in the fractions and decimals subject area—especially with understanding fraction equivalence and ordering—compared to their Korean peers. Similarly, the mastery of Attribute 8 was required in solving geometry items; so, compared to their Korean peers, Turkish students did not perform well on the geometry items that involved drawing, constructing, and describing geometrical figures and the relationships between them. Additionally, except for one item, the mastery of Attribute 13 was necessitated in solving data and chance items. Hence, in comparison to Korean students, as in the Dogan and Tatsuoka (2008) study, Turkish students also did not perform well on the data and chance problems that investigated the chance process and using and evaluating probability models. On the contrary, the three lowest prevalence estimate differences between Korea and Turkey were obtained for Attributes 3, 11, and 5. Thus, it can be concluded that Turkish students performed relatively well on items that involved understanding ratio concepts and using ratio reasoning; recognizing perimeter and understanding concepts of area; and reasoning about and solving one-variable equations and inequalities.

It should be noted that model-data fit and item fit statistics may have an effect on the interpretations of item parameter estimates obtained from a DINA model. More appropriate conclusions can be made based on models with better fit. It is obvious that DINA models in this study did not show perfect fit to two data sets. Assuming that we have enough model-data fit, we can make several conclusions based on DINA model results. Under this condition, item parameter estimates from the DINA model can provide feedback for students from the two countries. Apparently, Korean students were less likely to slip and were more likely to guess correct answers. However, the Turkish sample yielded lower guessing parameter estimates and higher slipping parameter estimates, indicating possible problems with content knowledge or testing strategies. Item parameter estimates can also be used for improving measurement instruments. Results of item parameter estimates showed problems with several items. For example, Items 6 and 8 yielded higher slipping parameter estimates for both samples. Both items were classified under the algebra content domain, and Item 6 was a multiple choice item, whereas Item 8 was a constructed response item. Turkish students were also most likely to slip on Items 21, 14, 12, and 27. Considering Items 21, 14, and 12 were also constructed response items, it can be concluded that Turkish students were more likely to slip on constructed response items. Dogan and Tatsuoka (2008) also stated a similar weakness of the Turkish students' in their study. They observed that Turkish students did not perform well on the open-ended items and had difficulty constructing answers in comparison to selecting an answer from given alternatives. Therefore, the findings of this study suggest that Turkish educators and policy makers should pay more attention to teaching students how to deal with constructed response items instead of teaching test skills to solve multiple choice items. To accomplish this, teachers should encourage students through verbal and written expressions of their mathematical understandings. In addition, item discrimination indices may also be useful for identifying poor items. For instance, Items 1 and 15 had the highest guessing parameters and lowest discrimination indices for the Korean sample. Hence, these two items were not very informative and required improvements.

In sum, various factors might have affected Korean and Turkish eight-grade students' performances on the TIMSS assessment. As previously discussed, Im and Park (2010) attributed Korean students' high achievement to the social and instructional factors. In a similar vein, when compared Chinese Taipei and Turkey on the TIMSS 2007 eight-grade science items, Ozturk and Ucar (2010) found that socio-economics, parents' education level, and quality of schooling contributed to Turkish students' relatively low academic performance. In this study, our results identify situations for instructors where current curriculum may be improved to help students master some lacking attributes based on CDM-based feedback. As Leighton and Gierl (2007) stated, recent CDM studies have been applied for post-hoc analyses and item analyses rather than constructing the tests (Chapter 7). Although, our study demonstrated that retrofitting of a CDM via the DINA model can be very useful for the TIMSS assessment, it is evident that more benefit can be obtained from CDM-based analyses when tests are designed using CDMs in advance.

REFERENCES

Choi, M. K., Lee, Y-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1563–1577.

Chung, Y.L., & Son, D.H. (2000). Effects of cooperative learning strategy on achievement and science learning attitudes in middle school biology. *Journal of the Korean Association for Research in Science Education, 20,* 611–623.

Common Core State Standards Initiative (2010). *The common core state standards for mathematics*. Washington, D.C.: Author.

Common Core State Standards Initiative (2014). Myths vs. facts. Retrieved from http://www.corestandards.org/about-the-standards/myths-vs-facts/

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.

de la Torre, J., & Lee, Y. S. (2008, March). *Relationships between cognitive diagnosis, CTT and IRT indices: An empirical investigation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, *Psychometrics*) (pp. 979–1027). Amsterdam: Elsevier.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68*(3), 263–272.

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox (version 3.1)* [Computer software]. London: Timberlake Consultants Press.

Fisher, D. L., & Kim, H. B. (1999, April). *Constructivist learning environments in science classes in Korea.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Fu, J., & Li, Y. (2007, April). *An integrated review of cognitively diagnostic psychometric models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–323.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74,* 191–210.

Holliday, W. G., & Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Education Forum, 67*, 250–257.

House, J.D. (2009). Classroom instructional strategies and science career interest for adolescent students in Korea: Results from the TIMSS 2003 assessment. *Journal of Instructional Psychology, 36*, 13–19.

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation, 16*(3), 287–301.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kim, K., Kim, S., Kim, N., Park, S., Kim, J., Park, H., & Jung, S. (2008). *Characteristics of achievement trend in Korea's middle and high school students from International Achievement Assessment (TIMSS/PISA) (KICE Research report, RRE-2008-3-1).* Seoul, Korea: Korea Institute of Curriculum and Evaluation.

Lee, Y.-S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.

Lee, Y.-S., Johnson, M., Park, J. Y., Sachdeva, R., Zhang, J., & Waldman, M. (2013, April). *A multidimensional scaling (mds) approach for investigating students' cognitive weakness and strength on the TIMSS 2007 mathematics assessment.* Paper presented at the 2013 Annual Meeting of the American Educational Research Association Conference in San Francisco, CA.

Leighton, J., & Gierl, M. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (2012). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

Oh, S. (2005). Discursive roles of the teacher during class sessions for students presenting their science investigation. *International Journal of Science Education, 27,* 1825–1851.

Ozturk, D., & Ucar, S. (2010). By using TIMSS data, determination and comparison of the factors that affects science achievement of 8 grade students from Taiwan and Turkey. *Cukurova University Journal of Social Sciences, 19*(3), 241–256.

Park, K.M. (2004, July). *Mathematics teacher education in East Asian countries: From the perspective of pedagogical content knowledge.* Paper presented at the 10th International Congress on Mathematical Education, Copenhagen, Denmark.

Roussos, L., DiBello, L. V., Stout, W., Hartz, S., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & Gierl, M. J. (Ed.), *Cognitively diagnostic assessment for education: Theory and practice.* (pp. 275–318). Thousand Oaks, CA: SAGE.

Rupp, A. A., & Templin, J. L. (2008a). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262.

Rupp, A. A., & Templin, J. (2008b). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–96.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York: Guilford Press.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12,* 55–73.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Toker, T., & Green, K. (2012, April). *An application of cognitive diagnostic assessment on TIMMS-2007 8th Grade Mathematics items.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report RR-05-16.

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30,* 17–21.

GENİŞ ÖZET

Giriş

TIMSS (Trends in International Mathematics and Science Study) sınavı 1995 yılındaki ilk uygulamasından beri 4. ve 8. sınıf fen bilgisi ve matematik derslerinde katılımcı ülkelerin kendi öğrencilerinin performanslarını diğer katılımcı ülke öğrencilerinin performanslarıyla karşılaştırmalarına yardımcı olmuştur. Her ne kadar TIMSS tarzındaki uluslararası büyük ölçekli sınavların bu tür karşılaştırmalar için kullanılması eleştiriliyor olsa da (Holliday ve Holliday, 2003; Wang, 2001), katılımcı ülkeler bu sınavlardan alınan sonuçlara göre kendi öğretim sistemlerinde ve müfredatlarında düzenlemelere gitmişlerdir. Genel olarak bakıldığında bu büyük ölçekli uluslararası sınavlar toplam skora dayalı bir değerlendirme sistemi içermekte ve her ülkenin öğrencilerine toplam puanlar atayarak ülkeler bazında elde edilen ortalama puanlara göre ülkelerin kendi yerleri hakkında karşılaştırma yapmalarına imkan sağlamaktadır. Tek bir puana dayalı değerlendirme yaklaşımları öğrenci performansları açısından çok detaylı bilgi sunmadığı gerekçesiyle eleştirilmiş (Nichols, 2012; Leighton ve Gierl, 2007) ve bunların yerine daha detaylı değerlendirmeye olanak sağlayan bilişsel tanı modelleri geliştirilmiştir (Rupp, Templin, ve Henson, 2010). Bilişsel tanı modellerine ait detaylar Rupp, Templin ve Henson (2010) ve DiBello, Roussos ve Stout (2007) çalışmalarında bulunabilir. Bu bilişsel tanı modellerinden en yaygın olarak kullanılanlardan bir tanesi olan DINA (*deterministic inputs, noisy "and" gate,* Haertel, 1989; de la Torre, 2009; Junker ve Sijtsma, 2001) modeli bu çalışmada kullanılmıştır. Temel olarak DINA modeli bir maddenin doğru cevaplanabilmesi için o madde için gerekli olan özellikler neler ise cevaplayıcının bu özelliklerde yeterlilik kazanmasını şart koşar. Her madde için madde kayması (item slipping) ve madde tahmini (item guessing) olmak üzere iki parametre sonucu elde etmemizi sağlar.

Son zamanlarda çeşitli çalışmalar bilişsel tanı modellerini kullanarak öğrencilerin TIMSS, PISA, ve PIRLS gibi uluslararası büyük ölçekli sınavlardaki başarılarını karşılaştırmışlardır. Bu çalışmalar öğrencilerin performansları ve becerileri, öğretmenlerin öğretim yöntemleri ve öğrencilerin performansları arasındaki ilişki, ve katılımcı ülkelerin eğitim sistemleri ile müfredatları hakkında çok kullanışlı bilgi edinme imkanı sağlamıştır. Örneğin, Dogan ve Tatsuoka (2008) Türk ve Amerikan sekizinci sınıf öğrencilerinin TIMMS-R 1999 sınavındaki matematik performanslarını karşılaştırmışlardır. Bu çalışmaya göre Türk öğrencilerin cebir ve olasılık/istatistik gibi sınavlarda Amerikan öğrencilere göre daha düşük performans sergiledikleri ortaya çıkmıştır. Benzer şekilde, Im ve Park (2010) Güney Kore ve Amerikan sekizinci sınıf öğrencilerinin TIMMS 2003 sınavındaki matematik performanslarını karşılaştırmışlardır. Çalışmanın bulguları Güney Kore ve Amerikan öğrencilerinin performansları arasında çok önemli farklılıklar olduğunu göz önüne çıkarmıştır. Bu farklılıklar özellikle problemlerin yeniden yapılandırılması ve akıl yürütme ile ölçme ve geometri konularında önemli değişiklikler göstermiştir.

Bu çalışmanın asıl amacı TIMSS 2011 matematik sınavındaki konular açısından sekizinci sınıf Türk öğrencilerinin güçlü ve zayıf yanlarını incelemektir. Bu amacı gerçekleştirmek için bu çalışmada Türk ve Güney Kore'li öğrencilerin göreceli performansları karşılaştırılmıştır. Güney Kore'li öğrencilerin düzenli olarak TIMSS matematik sınavında ilk üç sırada yer almaları Kore'yi referans ülke olarak almamızda temel neden olmuştur. Bu doğrultuda bu çalışma tanılayıcı değerlendirme yaklaşımını kullanarak aşağıdaki iki araştırma sorusunu cevaplamaya çalışmaktadır:

1) Türk ve Kore'li sekizinci sınıf öğrencilerinin TIMSS 2011 matematik performansları göreceli olarak nasıl farklılıklar göstermektedir?

2) Kore'li öğrencilerle karşılaştırdığında, Türk öğrencilerinin TIMSS 2011'deki matematik konularında güçlü ve zayıf yanları nelerdir?

Yöntem

Bu çalışmada Türkiye ve Güney Kore ülkelerine ait sekizinci sınıf TIMSS 2011 matematik veri setleri kullanılmıştır. Seçilen örneklemlerde 368 Güney Kore'li ve 488 Türkiye'li öğrenci bulunmaktadır. TIMSS 2011'de uygulanan 14 test kitapçığından 2 numaralı kitapçık bu çalışmadaki analizler için seçilmiştir. Kitapçık 2'de 15 çoktan seçmeli ve 17 kısa yanıtlı madde bulunmaktadır. Seçilen bu 2 numaralı kitapçık bilişsel tanı modellerinden olan DINA model kullanılarak analiz edilmiştir. İki kategorili DINA model kullanıldığı için çoktan seçmeli test maddeleri ve kısa yanıtlı maddeler 0 (yanlış cevap) ve 1 (doğru cevap) şeklinde kodlanmıştır. DINA model analizleri için gerekli olan Q-matris maddeleri doğru cevaplamak için gerekli olan özellikler göz önünde bulundurularak dört matematik eğitimcisi tarafından bağımsız bir şekilde kodlanmıştır. Bu dört eğitimcinin görüşlerine göre toplamda 13 tane özellik Common Core State Standards for Mathematics (CCSSM; Common Core State Standards Initiative, 2010) müfredatı kullanılarak oluşturulmuştur. Q-matris ve öğrenci cevaplarının 1-0 şeklinde kodlanmış olduğu veri setleri kullanılarak DINA model analizleri yapılmıştır. İki ülkeye ait veriler OxEdit programı kullanılarak maksimum olabilirlik yöntemi vasıtasıyla analiz edilmiştir. TIMSS 2011'in uygulayıcı kurum (IEA) tarafından üç parametreli madde tepki kuramı (MTK) ile analiz edilmiş olmasından dolayı üç parametreli MTK modelinden elde edilen iki ülkeye ait sonuçlar da karşılaştırma amaçlı sunulmuştur.

Sonuç ve Tartışma

DINA modeli kullanılarak analiz edilen iki ülke veri setine ait 31 madde için elde ettiğimiz madde parametreleri madde kayma (slipping), madde tahmin (guessing) ve bu iki parametre kullanılarak hesaplanan madde ayırt ediciliği (discrimination) değerleri şeklinde ayrı ayrı rapor edilmiştir. Maddeleri çözmek için gerekli olan özelliklere ait olarak da özellik yaygınlığı (attribute prevalence) yüzdelikler şeklinde sunulmuştur. Türk öğrenciler hem madde parametreleri hem de özellik parametreleri açısından Kore'li öğrencilerden farklılık göstermişlerdir. Genel olarak Kore verisinden elde edilen madde parametreleri yüksek tahmin (guessing) ve düşük kayma (slipping) değerleri içerirken bu durum Türk öğrenciler için tam tersi olarak gözlenmiştir. Koreli öğrenciler testteki maddeleri çözmek için gerekli olan özelliklerin hemen hemen hepsinde yeterlilik kazanmışken Türk öğrenciler çoğu özellikte yeterlilik kazanamamakla beraber en çok Özellik 1, 8 ve 13'te düşük yeterlilik göstermişlerdir.

Dogan ve Tatsuoka (2008) çalışmasına benzer olarak, Türk öğrenciler Cebir, Data Analizi ve Şans konularında Kore'li öğrencilere göre düşük performans sergilemişlerdir. Bu konuların yanında Türk öğrenciler ayrıca Geometri konusunda da Kore'li öğrencilere göre daha az başarılı olmuşlardır. Yine Dogan ve Tatsuoka (2008) çalışmasına benzer olarak Türk öğrenciler açık-uçlu sorularda yeterli başarıyı gösterememişlerdir. Türk öğrencilerin açık-uçlu soruları cevaplamaktaki yetersizliği Türkiye'nin çoktan seçmeli testler üzerine dayalı olan eğitim sisteminin bir neticesi olarak yorumlanabilir. Test sisteminin yanı sıra, Ozturk ve Ucar (2010)'ın da bahsettiği üzere Türk öğrencilerinin düşük performanslarında sosyo-ekonomik nedenler ile, ailelerin eğitim durumları, ve okullardaki öğretimin kalitesi gibi faktörler de etkili olmuş olabilir. Benzer şekilde, Im ve Park (2010) Güney Kore'li öğrencilerin yüksek başarısının sosyal faktörler ve öğretim ile ilgili faktörlere bağlı olduğunu belirtmiştir. Bu çalışma, içerinde sunulan bulguların Türk eğitimcilerine matematik müfredatının nasıl geliştirebileceğine dair bilişsel tanı modeline dayalı geri dönütler vermesi açısından kayda değerdir.