JOURNAL OF EDUCATIONAL SCIENCES
www.educational-sciences.com

COGNITIVE DIAGNOSTIC ASSESSMENT OF TIMSS-2007 MATHEMATICS ACHIEVEMENT ITEMS FOR 8TH GRADERS IN TURKEY

Turker TOKER*, Kathy GREEN**

*Unıversıty of Denver, Morgridge College of Education, Research Methods and Statistics,turker.toker@du.edu
** Unıversıty of Denver, Morgridge College of Education, Research Methods and Statistics, Kathy.Green@du.edu

ABSTRACT
This study investigated students' responses to 49 released TIMSS-2007 mathematics items and considered what those responses might show about participants' mastery level of cognitive attributes. There were 4,498 8th-grade students from seven geographical regions of Turkey who responded to these TIMSS items. In this study, an IRT analysis was also conducted to both compare item parameters based on the data from Turkey with international item parameters. As a result of this study, educators in Turkey should note that students have weaker knowledge in geometry, graphics, charts, figures, rule application in algebra, and data management and stronger skills with whole numbers and integers, data, probability, and basic statistics, translating/formulating equations and expressions to solve a problem, judgmental applications of knowledge in arithmetic and geometry, and applying and evaluating mathematical correctness.
The results of the present study can help teachers and curriculum developers in Turkey ensure that the utilized educational policies and methodologies help students to improve their cognitive mastery levels.

Key words: *cognitive assessment, mathematics, Least Squares Distance Method, validation, Rasch model, psychometrics, TIMMS-2007*

ÖZET
Bu çalışma TIMSS-2007 Türkiye evreni çerçevesinde yapılmış olup öğrencilerin bilişsel yeterlilik düzeyleri incelenmiştir. TIMSS-2007 sınavına Türkiye'nin yedi coğrafık bölgesinden 4.498 sekizinci sınıf öğrencisi katılmıştır. Ayrıca bu çalışmada IRT analizi kullanılarak uluslararası madde güçlükleri ile Türkiye evrenine ait madde güçlükleri de karşılaştırılmıştır. Bu çalışma sonucunda öğrencilerin geometri, grafikler, figürler, cebir kuralları, data yönetimi, tam sayılar, olasılık, basit analiz, problem çözümünde formül kullanımı, eleştirel düşünme, uygulamalı matematik ve geometri alanlarında zorluklar yaşadığı yargısına ulaşılmıştır.
Bu çalışmanın sonuçlarının öğretmenlere ve matematik programı hazırlayanlara, Türk eğitim sisteminde uygulanan politikaların ve yöntemlerin öğrencilerin bilişsel yeteneklerini geliştirebilmeleri adına yardımcı olması düsünülmektedir.

Anahtar Sözcükler: *Bilişsel Değerlendirme, Matematik, Kücük Kareler Yöntemi, Geçerlilik, Tek Parametreli Lojistik Model, Psikometri, TIMSS-2007*

INTRODUCTION

Educational assessment has evolved from grading one's achievement level to being diagnostically useful at every step in education (Bolt, 2007). Assessment affects grades, placement, advancement, instruction, curriculum, and in some cases, funding. Assessment is critical to both evaluating the effects of educational programs and also to implementing those programs. Outcomes of education need to meet globally accepted criteria. Students must be able to think wisely and critically, to examine in detail, and to make inferences which are important very important in today's global world (Gierl,2007). Changes in the skills base and the knowledge which our students need, require new learning goals; these new learning goals change the relationship between assessment and instruction with teachers needing to be informed about both.

In this study, data were taken from one of the most respected international exams, the Trends in International Mathematics and Science Study (TIMSS-2007), used to assess cognitive abilities of 8th graders' mathematics achievement. The purpose of this study was to use cognitive attributes as developed by Tatsuoka (Tatsuoka, 1983) and her associates to extend our understanding of the skills and cognitive processes as well as knowledge attributes mastered by students in the Turkish educational system. The study was specific to attribute identification for Turkish students who participated in TIMSS-2007and to provide information about attributes of TIMSS-2007 8th grade mathematics results for Turkey to inform teachers in the Turkish educational system about student mastery of different skills and processes. Even though TIMSS 2007 has extensive content validity support based on MONE's information given to the test administrators (TIMSS-2007 Technical Report, 2008), results show that instructional methods are not adequate as students' scores fall short. This study investigated students' responses to the released mathematics items and then considered what those responses might show about participants' mastery level of cognitive areas. By taking the results to the item level, instructors can see whether the mistakes made by students in the sample are also mistakes made by their own students. Therefore, teachers can focus on those areas in need of more work during their instruction to improve cognitive mastery levels of students.

TIMSS-2007

TIMSS was designed to assess trends in students' mathematics and science achievement. TIMSS-2007 is the fourth in a four-year-cycle of assessments (previously administered in 1995, 1999 and 2003). One purpose of the assessment is to align mathematics and science curricula in the participating countries with the intended TIMMS curricula (TIMSS-2007 Technical Report, 2008). TIMSS results assess the degree to which students have learned mathematics concepts and skills likely to have been taught in school. TIMSS tests put an emphasis on questions and tasks that offer insight into the analytical, problem-solving, and inquiry skills and capabilities of students. In addition, students, teachers, and school principals in each participating country are asked to complete questionnaires concerning the context for learning mathematics and science, so as to provide a resource for interpreting the achievement results and to track changes in instructional practices. TIMSS-2007 assessed the mathematics and science achievement of children in two target populations: fourth grade and eighth grade students (Shen, 2000).

Reliability and validity were concerns of TIMSS developers. The organization has procedures to ensure that the tests are reliable. Because reliability is not enough for good quality measurement, an effort also is directed to assessing the validity of the tests. The validity of test items includes unified agreement in mathematics and science for both 4th- and 8th grade students. Agreement means that the items included in the tests measure those agreed-upon elements of mathematics and science (TIMSS-2007 Technical Report, 2008). For extensive information regarding the reliability and validity of the TIMSS-2007, see the TIMSS-2007 Technical Report.

Cognitive Diagnostic Assessment (CDA)

One of the most important concerns in education systems is summative assessment. In order to have powerful, effective, and meaningful summative assessment, evaluation should also be formative, which means it has to support teaching and learning processes with results (DiBello&Stout, 2007). An ideal assessment would not only be able to meet precise psychometric standards, but would also be able to provide specific feedback about how students learn and what attributes they need to achieve goals. Cognitive diagnostic assessment is used to examine the cognitive processes necessary for successful task completion, because it supplies specific information about each attribute students need to master instead of a single score result (McGlohen, 2004).

The method used in this study was developed by Tatsuoka and her associates (1983), and comprises two parts. The first part involves determining the relation between the items of a test and the attributes they are assessing. The description of which attributes are essential for each item is shown in a Q-matrix. The Q-matrix is a [K x n] binary matrix, where K is the number of attributes to be measured and n is the number of items on the test. For a given element of the Q-matrix in the kth row and the ith column, a value of one indicates that item i does indeed measure attribute k and a zero indicates it does not. Figure 1 provides an example of a 3x3 Qmatrix:

Figure 1.Sample Q-matrix

Q =

	i1	**i2**	**i3**
k1	1	1	0
k2	0	0	0
k3	1	0	1

The first item in the above Q-matrix shows that students who mastered the first and third attributes should have correct responses to item one. For a correct response to item two, students should master the first attribute. A correct response to item three means they should have mastery of attribute three. Attributes which are not necessary for a correct response to the item are shown by zeros meaning that students do not need to master those attributes to correctly respond to that item. Expert consultation in the content area is a way to build the structure of the Q-matrix to determine if an item measures a specific attribute (see Appendix). Some other ways to build the structure of the Q-matrices can be borrowing from the test blueprint or subjectively evaluating each item to draw conclusions about which attributes are being measured (Tatsuoka, 1983). The second part of the rule space method deals with interpretation of the relationship between test items and the Q-matrix. The reason for using the rule space method is to obtain diagnostic information on students' mastery levels of specified cognitive skills to improve interpretations of test scores on the item level for students, teachers, and administrators.

For the purpose of this study, a relatively new approach to cognitive diagnostic analysis, called the Least Squares Distance Method (LSDM), was used to explore the mastery of cognitive attributes on the released TIMSS-2007 mathematics test items. This approach has not been used previously with TIMSS items.

Least Squares Distance Method

The least squares distance method (Dimitrov, 2007) uses the Rasch item position parameters of binary test items to (a) validate cognitive attributes that underlie item responses and (b) assess the probability of correct processing of such attributes across levels of the scale continuum. The LSDM is a conjunctive model in which a correct answer on a test item requires mastery of all cognitive attributes associated with that item. The cognitive attributes for all test items are outlined in a Q-Matrix, where a “1” shows an attribute is needed for an item and a “0” means an attribute is not needed. The basic assumption with LSDM is that, theoretically, the probability of correct item response is equal to the probability that all required attributes are correctly applied.

The graphical image of this probability across ability levels shows the probability curve for cognitive attribute A_k. The Rasch item characteristic curve (ICC) is represented by the recovered item probabilities (*Prec*), or the recovery curve. The ICC recovery compared to LSDM provides information about how well the required attributes describe the item across ability levels. The mean absolute difference (MAD) between the LSDM curve and the ICC provides for validation of attributes for each item across ability levels. A MAD equal to 0.0 would indicate perfect ICC recovery. According to Dimitrov(2007), a classification for level of ICC recovery was developed: “(a) very good (0.00 ≤MAD < 0.02), (b) good (0.02 ≤ MAD < 0.05), (c) somewhat good (0.05 ≤ MAD < 0.10),(d) somewhat poor (0.10 ≤ MAD < 0.15), (e) poor (0.15 ≤ MAD < 0.20), and (f) very poor (MAD ≥ 0.20).” (p. 373). These criteria were applied in this study.

Dimitrov (2007) pointed out an interpretation of LSDM results with respect to heuristic criteria for validation of cognitive attributes: “(1) The smaller the LSD…, the better the cognitive attributes hold together (jointly for all items) at this ability level; (2) The attribute probability curves (APCs) should exhibit logical and substantively meaningful behavior in terms of monotonicity, relative difficulty, and discrimination; (3) The better the ICC recovery for an item, the better the required attributes explain the item” (pp. 372-373).

METHOD

Participants

Released items and Turkish students’ responses to TIMSS-2007 in mathematics administered in 2007 were used in this study. There were 4,498 8th-grade students from seven geographical regions of Turkey. There were 14 booklets which were randomly assigned to students with between 314 and 331 students receiving each booklet. Booklets contained at least one released item. Since there are limited numbers of released items which were repeated in different booklets, all booklets were used in this study. The study used data from 2,093 female students with a mean age of 13.96, and 2,405 male students with a mean age of 14.09.

A Q-matrix was developed by two Turkish speaking mathematics teachers and the first author of this paper. All of the experts had bachelor’s degrees in teaching and were male. Two of them worked as mathematics teachers in the U.S. The ages of experts were 27, 30 and 35. The first author of this paper had three years of elementary school teaching experience in Turkish schools. The other two experts had three and five years of math teaching experience.

Instrument

The TIMSS-2007 for 8th grade consisted of 179 questions which included 96multiple-choice questions. There were only 51 multiple-choice items released. No information was found about why TIMSS administrators released those items. Two of the items were dropped since they did not provide any variation at all in student responses, i.e. all students answered correctly or all answered incorrectly. The final dataset was based on 49 items; see Appendix for sample items. For the Q-matrix, only released multiple-choice items were used. This test covered content domains of numbers, geometric shapes and measures, and data display. The cognitive domains included in the exam included knowing, reasoning, and applying. Two examples of items from this test can be seen in Figures 2 and 3.

Analysis

A Q-matrix of cognitive attributes (Appendix and Table 1) of each item was developed based on Tatsuoka and her associates' classification of cognitive attributes (Tatsuoka, Corter, & Tatsuoka, 2004). The Q-matrix includes 27 attributes divided into the three categories of content, process, and skill. Cognitive attributes which are represented by all items or not represented by all items were not used since they would provide no variation. The final version of the Q-matrix included 20 attributes. To ensure the consistency of the subjective judgment of attributes, the cognitive attribute matrix was developed based on the independent identifications of three experts mentioned in the previous pages.

A correct answer on an item means that a student has mastered all attributes required, within a margin of error. A linear regression analysis was conducted to see if the Q-matrix could explain item difficulty. Using the responses of Turkish students to the released math items of TIMSS-2007, an IRT analysis was conducted to both compare results with international item parameters and to use for LSDM analysis. Item parameters were correlated and a scatterplot constructed to identify items that had distinctly different positions for the Turkish and international data. Logit item difficulties for those items were compared using a significance test.

The probability of correct response was calculated via the WINSTEPS program for each item for Turkish students (Linacre, 2007). The individual attribute probabilities and the average LSDs for the 49 items across ability levels were estimated using the MATLAB computer program (TheMathWorks, Inc., 2005). Finally, the mean absolute difference between the ICC and the LSDM recovery curve for each item was calculated to validate the cognitive attributes.

RESULTS

Item Parameter Comparison

It was hypothesized that the item parameters would be invariant across Turkish students and the international data. To test this hypothesis for the released 49 items, WINSTEPS results for Turkish students and item parameters from the official website of TIMSS were used. There were three items which were statistically significantly different at $p < .01$ in logit position: items 1(t =5.69), 25 (t =7.35), and 37 (t =8.45). These three items assessed basic concepts and operations in fractions and decimals (C2), computational applications of knowledge in arithmetic and geometry (P2), basic concepts and operations in elementary algebra (C3).

A Pearson correlation between these two sets of item logit positions produced a correlation of r = .82, and so strong consistency in item logit position. Both overall item parameters and Turkish students' item parameters are given in Table 1.

Cognitive Attributes Matrix

The overall agreement level among the three experts on the Q-matrix elements was 68%. After a final meeting, the judges agreed on the final version of the Q-matrix. Seven attributes (P5, P6, S4, S5, S8, and S10) which were in the original classification of cognitive attributes were deleted because no items required the attributes. One more attribute (S11) was excluded because it was present in all items. The last version of the cognitive attribute matrix had 6 content, 8 process, and 6 skill attributes, or 20 total attributes (see Table 1).

There were some statistically significant relationships between attributes. The relationship between attributes C1-Basic concepts and operations in whole numbers and integers, and P9-Management of data and procedures was statistically significant (r =.43, p < .05). Additionally, the correlation between C1-Basic concepts and operations in whole numbers and integers, and C4-Basic concepts and operations in two-dimensional geometry was also significant (r = .62, p < .05). For attributes P7-Generating, visualizing, and reading figures and graphs and S3-Using figures, tables, charts, and graphs the correlation was high and statistically significant (r = .89,p < .05). From these results, we can conclude that some attributes overlap. We can also say in order to master one attribute, one needs to master another attribute.

A multiple regression analysis showed that most of the variance in item difficulties was accounted for by the identified cognitive attributes; however none of the individual attributes was statistically significant. The estimates of R^2 and adjusted R^2 were .65 and .42, respectively.

LSDM Analysis

The LSDM was conducted across 17 ability levels in the interval from −4.0 to 4.0, with increments of 0.5 on the logit ability scale. The 20 attributes were applied more accurately and consistently by higher ability examinees. The attribute probability curves (APCs) monotonically increased across the ability levels and provided information about the relative difficulty and discrimination of the 20 attributes. C4-Basic concepts and operations in two-dimensional geometry was the most difficult attribute because its APC was consistently below the other APCs across all ability levels. Other attributes can be put in increasing difficulty order as follows: C1, C5, P1, P3, P8, P10, S1, S9, C3, C2, S2, P9, C6, S7, P2, P7, S6, P4, S3, and C4.

The APCs obtained with these probabilities are graphed in Figure 4. C4 was the most difficult attribute (the lowest curve), followed in decreasing difficulty by S3, P4, P2, P7, S6, S7, C6, P9, S2, and C2. All the other attributes have similar difficulty values. For example, the probability of correct performance on C4 at the ability level y=0 was .654. That is, the likelihood of examinees with ability at the origin of the logit scale to correctly process geometrical operations in two dimensional geometry (C4), was .654. For the same examinees, the likelihood to correctly process algebra operations in elementary algebra (C3) was higher (.743).

For each item, the mean absolute differences for the ICC recovery across ability levels are given in Table 3. Graphically, the ICC recovery is presented (Figure 5) for four examples of items (47, 40, 19, and 17) with differing MAD levels. According to Dimitrov's criteria, item 47 is in the category of very good with a MAD of .017, item 40 is in the category of good with a MAD of 0.036, item 19 is in the category of somewhat good with a MAD of .060, and item 17 is in the category of somewhat poor with a MAD of .14. With the conventional rule for degree of ICC recovery described earlier in the LSDM section of this paper, the examination of all 49 graphs for ICC recovery and their MAD values revealed that the ICC recovery was very good for four items (20, 36, 41, and 47), good for 16 items (1, 2, 5, 11, 13, 16, 24, 25, 28, 31, 34, 35, 39, 40, 45, and 46), somewhat good for 15 items (4, 7, 10, 18, 19, 21, 22, 29, 30, 33, 37, 38, 42, 48, and 49), somewhat poor for eight items (6, 8, 9, 12, 17, 26, and 27), poor for four items (3, 14, 43, and 44) and very poor for two items (15 and 23). Generally speaking, such diagnostic information on ICC recovery can be particularly useful in validating math sub-skills for students.

Overall, the findings indicated that the 20 attributes related to difficulties in mathematical skills of students (Mean MAD = .075).The mean MAD value suggests that overall item recovery is somewhat good based on the Dimitrov's criteria. For this reason, the APCs of the 20 attributes provide valuable information in terms of their difficulty (Figure 4). But the results for ICC recovery suggest that there is room for improvement regarding the set of attributes and their links to items in the Q-matrix. Indeed, using Dimitrov's criteria compared to the MAD values of items in Table 3, 35 items have very good, good, or somewhat good ICC recovery while 14 items do not. According to these results, it might be said that the Q-matrix can be improved to get better results since 14items were not well-recovered; for example, see the location of Items 47 and 17 in Figure 5.

DISCUSSION

Cognitive diagnostic assessment is a useful way to examine mastery of skills and processes needed to answer test items (Tatsuoka, Corter, &Tatsuoka, 2004). This study examined cognitive attributes using item position parameters from the Rasch model and attributes identified by three experts. A benefit of this approach is that student performance on individual attributes is obtained. With information about student performance on an individual attribute in hand, instruction can be tailored to the individual attribute level and then to overall success on item solution. With the LSDM, the present study investigated cognitive attributes on TIMSS-2007 items for Turkish students. First, items were evaluated using the 27 initially-identified attributes. Once independent responses of experts were collected, a lack of variability was found resulting in deletion of seven of the attributes. In addition to this, two of the items were deleted due to a lack of response variability. The degree of the LSDM recovery of ICC was then assessed for the 49 items and item parameters for both international and Turkish students were analyzed in order to assess the correlation between those parameters and so the generalizability of results.

The results showed the pattern of mastery of cognitive attributes for the test with respect to the 20 revised attributes. The monotonic decrease of LSDs across ability levels demonstrated that it was unlikely that different cognitive strategies had been employed by students with different ability levels, and that the higher ability students applied the attributes more accurately and consistently. The APCs generally indicated attribute difficulties relative to each other and clear discriminations. The LSDM recovery of ICCs showed that on average, 71% of the items were recovered well by the attribute probabilities, revealing that most of the 49 items could be substantially explained by the identified attributes. But the LSDM recovery also suggested that there is need for modifying attributes for some items such as Items 3, 6, 8, 9, 12, 14, 15, 17, 23, 26, 27, 43,and 44, because they were not explained very well by the attributes.

With respect to the three categories of attributes, the present research generated some useful results. It was found that identifying the content attributes was easier than identifying the cognitive process and skill attributes. The MAD values also showed that the content attributes accounted better for the items than the other two kinds of attributes and so, the LSDM recovery of ICCs was more accurate for content attributes. Although all attributes together contributed to the correct response for an item, their overlap might lead to unclear results. Some of the attributes included the same mastery areas. For example P7 and S3 overlapped on usage of figures and graphs. This shows that it might not be reasonable to combine different types of attributes in one analysis. The most difficult attribute for students to master was C4-Basic concepts and operations in two-dimensional geometry. This indicated that Turkish students' level of mastery was not sufficient in geometry.

Item logit positions did not differ significantly between the international and Turkish indices, with the exception of three items. These items were item 1, item 25 and item 37.

This study was well developed with a large sample of students and the findings represented most of the released items in the mathematics test for the eighth grade in Turkey. But some caveats should be considered. The first limitation is the selection of the attributes. With respect to the cognitive model, if the same attributes are shown as relevant for two different items, the item difficulties of the two items should be close. Therefore, large differences in item difficulties would signify the misspecification or inadequacy of the chosen attributes for the items. Because of this, items 2 (logit position = .42) and 3 (logit position = -1.08), items 23 (logit position = -1.89) and 28 (logit position = .20), and items 27 (logit position = -1.45) and 33 (logit position = 1.13) were flagged for potential problems in the specification of the attributes since similar attributes were identified but the logit difficulties were very different.

Moreover, no guessing or omissions were taken into account in the LSDM (Dimitrov, 2007). In the current study, since the results from the IRT model showed that most of the items were beyond the students' abilities, making guessing likely. These problems with the LSDM were not taken into account in this study. Future studies might investigate effects of guessing with the LSDM.

REFERENCES

Bolt, D. (2007).The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, *44*(4), 377-383.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285–291.

Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement, 31*(5), 367-387.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*, 325–340.

Linacre, J. M. (2007). Winsteps (Version 3.64.2) [Computer Software]. Chicago: Winsteps.com.

Martinez, J. (2001) Exploring, inventing, and discovering mathematics: A pedagogical response to TIMSS. *Mathematics Teaching in Middle School, 7(2),* 114-120.

MEB-EARGED (2003). Üçüncü Uluslararası Fen ve Matematik Çalışması (TIMSS 1999) Ulusal Rapor [Third international mathematics and science study (TIMSS 1999) national report]. Ankara: Turkey.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). *TIMSS 2007 Technical Report.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center, 2008; http://timss.bc.edu/TIMSS2007/techreport.html.

McGlohen, M. (2004).*The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment.* Unpublished Doctoral Dissertation, University of Texas at Austin, TX.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20*(4), 901-926.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901-926.

The MathWorks, Inc. (2005). MATLAB (Version 7.1) [Computer software]. Natick, MA: The MathWorks, Inc.

SUMMARY

This study was well developed with a large sample of students and the findings represented most of the released items in the mathematics test for the eighth grade in Turkey. But some caveats should be considered. The first limitation is the selection of the attributes. With respect to the cognitive model, if the same attributes are shown as relevant for two different items, the item difficulties of the two items should be close. Therefore, large differences in item difficulties would signify the misspecification or inadequacy of the chosen attributes for the items. Because of this, items 2 (logit position = .42) and 3 (logit position = -1.08), items 23 (logit position = -1.89) and 28 (logit position = .20), and items 27 (logit position = -1.45) and 33 (logit position = 1.13) were flagged for potential problems in the specification of the attributes since similar attributes were identified but the logit difficulties were very different.

To assess the stability of results, it is suggested that in future studies, an LSDM analysis be run with a random Q-matrix and with Q-matrices of multiple experts run separately. In addition to this, an aggregate Q-matrix can be iteratively revised for poor items to decrease the MAD and increase the multiple correlation between attributes and item difficulties. If an item could be written to uniquely address a specific attribute, these results would provide concurrent validation evidence and could be the best scenario to see how well students are doing on certain attributes. To clearly see the relationship between items and attributes, simple attributes are better than complex ones. It is feasible that items might be constructed for content attributes and less so for process and skill attributes, which are conveyed via a content item. Thus process and skill attributes would be overlaid on content.

Industrialization of the country is very important to supply more jobs to unemployed members of the younger generation. During its development process, educational policies must shape Turkey's steps. This study shows that students have weak levels of mastery in certain areas, some of which are related to engineering and physical sciences. The results indicated that the administrators of educational systems in Turkey should consider changes in teaching geometry, because geometry may enable teaching of important mathematical thinking skills which are needed in physical science and engineering. These skills are of course very important in maintaining technological development in today's global industries (Attributes C4, P2, and S6).

Furthermore, students are not good at reading figures, graphs and tables. Also, the results are not good enough in basic algebra for a country that is on its way to membership in the European Union. Teachers might choose to focus on cognitive process attributes more because they were four of the seven most difficult attributes. Improving students' processing levels could produce better results at the international level. For example, to improve mastery of attribute C4, teachers could consider student-centered learning theories. Instead of saying what a triangle is, they might give examples from students' lives. Students could participate in several activities during their classes, which is possible if classrooms are not overcrowded. [The number of students in one class should not be more than 24 students (EU Criteria).] In addition, although the average number of students per teacher for Turkish 8th grade classrooms is lower than in many European countries, the Turkish government should balance it throughout the country. There are some areas where there are around 50-55 students in a classroom. There are also some areas where there are only 10-12 students in a classroom. The government should build new schools in more crowded areas of the country to allow greater use of class activities. Districts must also supply class activity materials to schools. Although eight years of education in elementary schools is compulsory and free for all students, schools are struggling with economic hardships. The ministry of national education might assume a stronger financial role in compulsory education system. Districts might also assume the role of supplying funding that schools can use for materials based on the number of students attending.

A country's development is highly connected to its curricula. Better quality outcomes from the educational system can only be supplied by better quality curricula. The results of the present study could help teachers and curriculum developers ensure that the utilized educational policies and methodologies would help students to improve their cognitive mastery levels.

Table 1. Item Difficulties and the 20 Initial Attributes for the 49 Items

Item	Item Diff. Inter.	Item Diff. Turkey	Content Attributes						Cognitive Process Attributes								Skill Attributes					
			C1	C2	C3	C4	C5	C6	P1	P2	P3	P4	P7	P8	P9	P10	S1	S2	S3	S6	S7	S9
1*	0.12	-0.45	1	1	0	0	0	1	0	1	1	0	0	0	1	1	0	0	1	1	1	1
2	0.95	0.42	1	1	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	0	0	1
3	-0.35	-1.08	1	1	1	0	0	0	1	1	1	0	1	1	0	0	1	1	1	1	0	1
4	0.55	-0.17	1	1	1	0	1	0	1	1	0	1	1	1	0	0	1	0	0	0	0	1
5	0.63	0.63	0	0	1	1	0	1	0	1	1	0	0	0	1	1	1	1	0	0	1	1
6	0.097	-0.4	1	1	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	0	0	0
7	-	-0.95	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
8	-	-0.56	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
9	-	0.99	0	1	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
10	-	0.70	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	1
11	-	-0.59	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
12	-	1.09	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	1	0
13	-	0.14	0	0	1	1	0	1	1	1	1	0	0	0	1	1	0	1	0	0	1	1
14	-	-0.86	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	1	1	0	1	0
15	-	-1.89	1	0	1	0	1	0	0	1	1	0	1	0	1	0	0	1	0	1	1	0
16	-0.19	0.26	0	1	1	0	1	0	0	0	1	1	1	0	0	1	1	0	0	0	1	0
17	-0.50	-1.11	1	0	1	0	1	0	0	0	1	0	1	0	0	1	1	1	1	0	0	1
18	-0.49	0.01	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	0	1	0	0
19	0.70	0.68	0	0	0	1	0	1	0	0	1	0	1	1	1	0	1	0	1	1	1	0
20	0.20	0.30	0	1	1	1	0	1	1	1	0	0	1	1	1	0	0	0	0	1	1	1
21	-0.68	-0.90	0	0	0	0	1	1	0	1	0	0	1	1	1	1	1	0	1	0	1	0
22	1.13	0.91	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	1
23	-0.93	-1.89	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1	0	1
24	-0.01	-0.68	1	1	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0	1	0	1
25*	0.31	-0.48	1	0	1	0	0	0	0	1	1	0	0	0	0	1	0	1	0	1	0	0
26	-0.15	-0.63	0	0	1	1	0	1	0	1	1	0	0	0	1	1	1	0	0	0	1	1
27	-	-1.45	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0
28	-	0.20	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
29	-	0.82	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
30	-	0.28	0	0	1	0	0	1	0	0	1	0	1	0	1	0	1	0	0	1	0	0
31	-	0.39	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0
32	-	1.53	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
33	-	1.13	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
34	-	0.11	0	0	0	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0
35	-	0.64	0	0	1	1	0	1	1	1	1	0	0	0	1	1	0	1	0	0	1	1
36	-	0.03	1	0	1	0	1	0	0	0	1	0	1	0	0	1	1	1	1	0	0	1
37*	1.05	-0.43	0	1	1	0	0	0	0	1	1	0	0	1	1	0	1	0	1	1	0	1
38	0.64	0.79	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
39	0.03	0.37	0	0	1	1	0	1	0	1	1	0	1	0	1	0	1	1	1	0	1	0
40	0.64	0.64	0	1	1	0	1	0	0	0	1	1	1	0	0	1	1	0	1	0	1	0
41	0.89	0.25	0	1	1	0	0	0	0	1	1	1	0	0	0	1	1	1	0	1	0	0
42	1.23	1.09	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	1
43	-0.23	-1.17	0	0	1	0	0	0	1	1	1	1	0	0	1	1	1	0	0	1	1	1
44	0.05	-1.11	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
45	0.53	-0.09	0	0	1	0	0	0	1	1	1	1	0	0	1	0	1	0	0	0	0	1
46	1.30	0.43	0	0	1	1	0	1	1	1	0	1	0	1	0	1	1	0	1	0	1	0
47	0.51	0.81	0	0	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	1

48	-0.15	0.03	0	0	1	0	0	0	0	1	1	1	0	1	1	0	0	1	1	0	1	0
49	0.91	1.22	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	1	1	1	0	1

*Item logit positions differ at $p < .05$ between Turkey/international.

Figure 2. Sample item (TIMSS-2007 Technical Report, 2008)

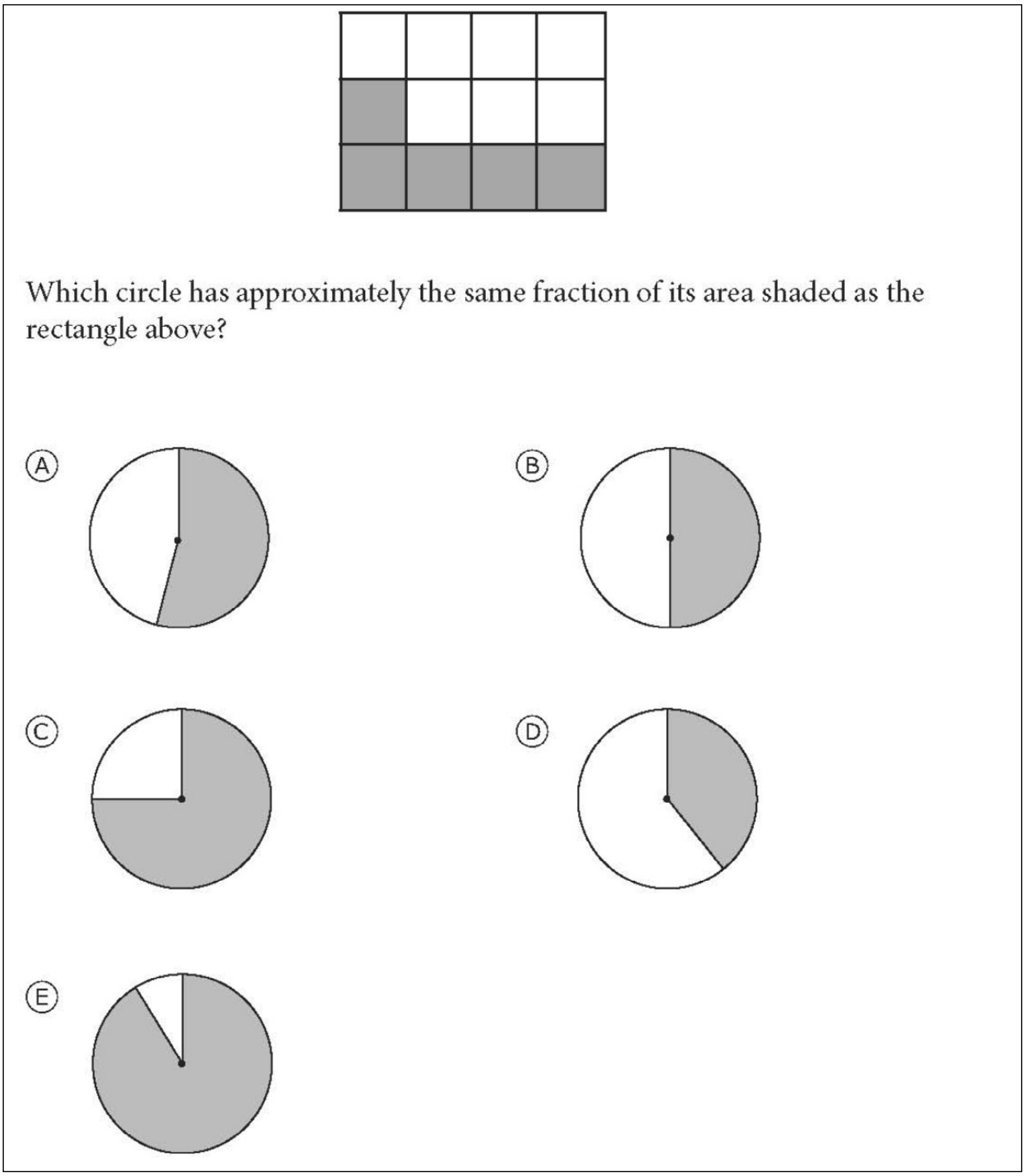


Figure 3. Sample item (TIMSS-2007 Technical Report, 2008)

A bowl contains 36 colored beads all of the same size: some blue, some green, some red, and the rest yellow. A bead is drawn from the bowl without looking. The probability that it is blue is $\frac{4}{9}$. How many blue beads are in the bowl?

(A) 4

(B) 8

(C) 16

(D) 18

(E) 20

Figure 4. Attribute Probability Curves

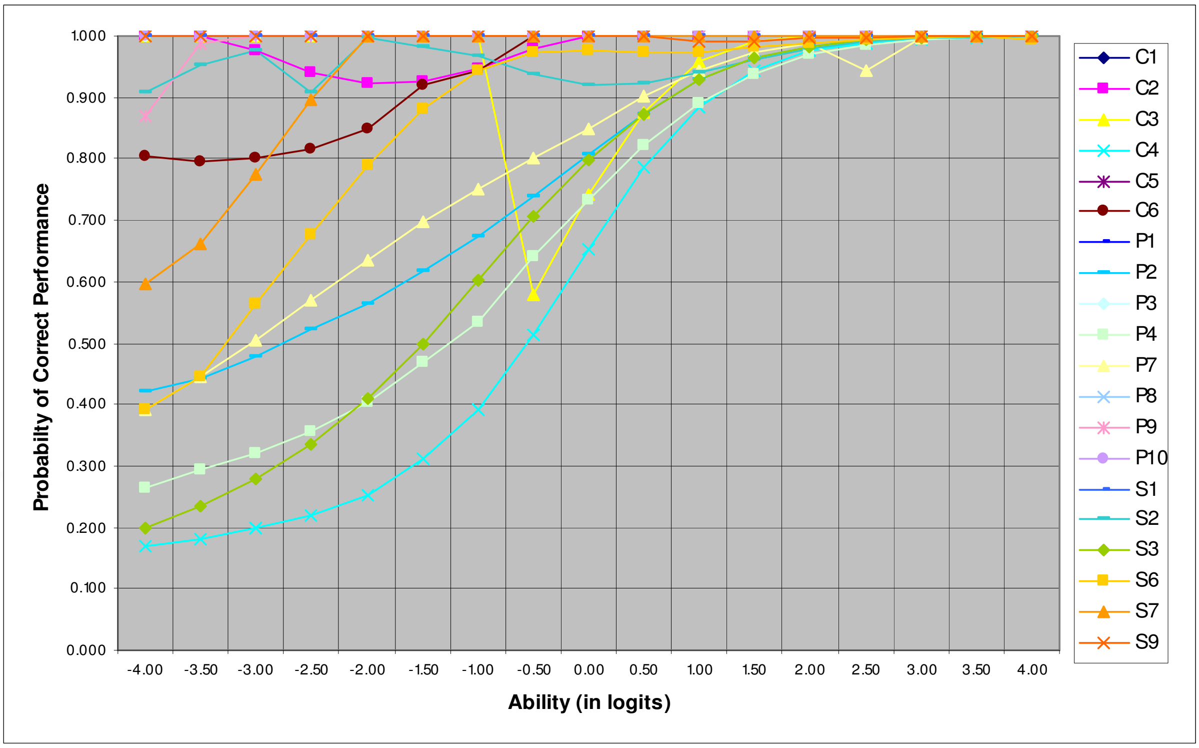
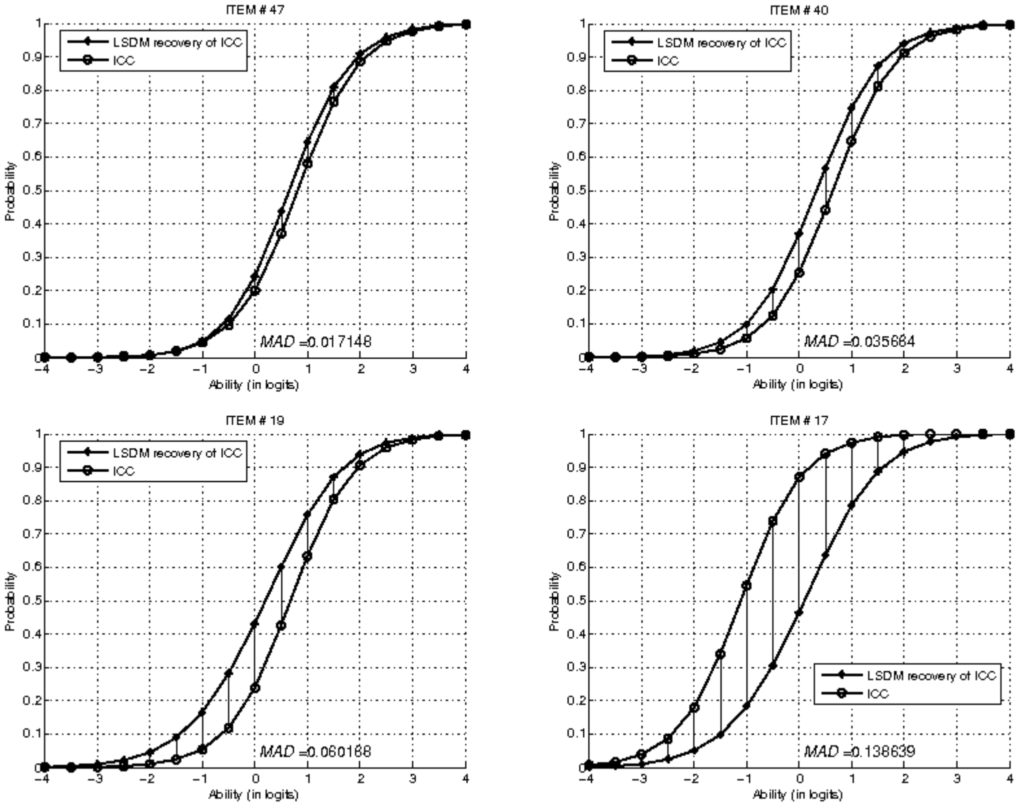


Figure 5.Item Characteristic Curve Recovery with the LSDM for Four Sample Items



Appendix

The Content, Cognitive Process and Skill Attributes for the TIMSS-R (1999)

Content attributes

C1 Basic concepts and operations in whole numbers and integers
C2 Basic concepts and operations in fractions and decimals
C3 Basic concepts and operations in elementary algebra
C4 Basic concepts and operations in two-dimensional geometry
C5 Data, probability, and basic statistics
C6 Measuring or estimating: length, time, angle, temperature, etc.

Cognitive Process attributes

P1 Translate/formulate equations and expressions to solve a problem
P2 Computational applications of knowledge in arithmetic and geometry
P3 Judgmental applications of knowledge in arithmetic and geometry
P4 Applying rules in algebra
P5 Logical reasoning-includes case reasoning, deductive thinking skills, if-then, necessary and sufficient, generalization skills
P6 Problem search; analytic thinking, problem restructuring; inductive thinking
P7 Generating, visualizing, and reading figures and graphs

P8 Applying and evaluating mathematical correctness
P9 Management of data and procedures
P10 Quantitative and logical reading

Skill (item type) attributes
S1 Unit conversion
S2 Apply number properties and relationships; number sense/number line
S3 Using figures, tables, charts, and graphs
S4 Approximation/estimation
S5 Evaluate/verify/check options
S6 Patterns and relationships (inductive thinking skills)
S7 Using proportional reasoning
S8 Solving novel or unfamiliar problems
S9 Comparison of two/or more entities
S10 Open-ended items, in which an answer is not given
S11 Understanding verbally posed questions