# How 8th Graders are Assessed through Tests by Mathematics Teachers?*

**Ahmet Delil****                                              **Bulent Nuri Ozcan**
Manisa Celal Bayar University, TURKEY                    Manisa Celal Bayar University, TURKEY

**Abstract:** Teacher-made tests (TMT) are the most used instruments for assessment and evaluation. This study investigates the cognitive requirements, test construction errors, and item types of TMTs. Content analysis technique is used in order to analyze and classify TMT items based on TIMSS-2019 assessment framework and based on criteria that is constructed to determine test construction errors. The data is consisted of 548 items in 30 exam papers of 18 mathematics teachers from 13 distinct schools. The distribution of TIMSS-2019 cognitive demands of all TMTs indicates that there is a strong emphasis on knowing or applying cognitive domains, with a total percentage of 93. Since 83% of all questions are of multiple choice and 17% are constructed-response type, teachers mostly prefer multiple choice item type. Findings also reveal that except face validity, there are errors concerning test constructions. Consequently, it is suggested that teachers should give more care on preparing items of higher cognitive levels, on tests of mixed type items, and on tests that involve lesser construction errors for more reliable tests. Finally, it is also suggested that measurement and evaluation specialists should be employed in each school or in each local Ministry of National Education Authority at least, in order to support teachers, but if this is not possible in a close time, there must be in-service training programs on measurement and evaluation for teachers to participate in.

**Keywords:** *Teacher-made tests, eighth grade mathematics, TIMSS-2019 assessment framework, test construction errors, item type.*

## Introduction

Assessment and evaluation are crucial concepts in education. The term assessment is defined as "the process of gathering evidence about a student's knowledge of, ability to use, and disposition toward mathematics and of making inferences from that evidence for a variety of purposes" (National Council of Teachers of Mathematics, 1995, p.3). The evidences are gathered in order to monitor student progress, evaluate programs, evaluate student achievement, and make instructional decisions (Van de Walle, Karp, & Bay-Williams, 2010). Also the term evaluation is used for "the process of determining the worth of, or assigning a value to, something on the basis of careful examination and judgment" (National Council of Teachers of Mathematics, 1995, p.3). Perhaps the most used instruments for gathering information for assessment and evaluation are teacher-made tests (TMT). Oescher and Kirby (1990) reported that TMTs dominate the assessments used by teachers, regardless of test purpose, grade level, or subject area.

Marso and Pigge (1991) reported that multiple-choice, matching, and short response item types were most frequently preferred by teachers. The most frequent errors they found were omitting directions, writing incomplete stems, requesting trivial facts rather than big ideas, developing ambiguous questions, and providing clues to test questions. They also found that test items that teachers used were mostly at the knowledge cognitive level which is the lowest cognitive level in hierarchical order. DiDonato-Barnes, Fives and Krause (2014) also stated that teachers experienced difficulty in selecting items at the cognitive level.

There are different classifications of cognitive levels made, and the best known one is the Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956) which classifies cognitive domain under six categories. The one defined by Marzano and Kendall (2008) is also well-known, and unlike Bloom's taxonomy, it is in a hierarchical order. The Trends in International Mathematics and Science Study (TIMSS) assessment framework is one of the most recent and probably the simplest which classifies cognitive domain under three categories as knowing, applying, and reasoning (Mullis & Martin, 2017). First conducted in 1995 on fourth graders only, TIMSS reports the trend data of student achievement

---

worldwide for every four years since then. While Turkish fourth grade students participated in TIMSS in 2011 and 2015 only, eighth-grade students participated in 1999, 2007, 2011 and 2015. The results of TIMSS in 2015 show that the performance of Turkish eighth graders is below the average, which is consistent with the previous performances in 1999, 2007 and 2011 (Mullis, Martin, Foy, & Hooper, 2016).

Considerable research had been done on cognitive levels of mathematics questions in Turkey. Among them Delil (2019) examined Turkish fifth-grade bursary examination mathematics items all of which are of multiple choice from years 1999 to 2018 based on TIMSS-2019 cognitive and content domains, and he concluded that Turkish fifth-grade mathematics items require upper cognitive demands comparing TIMSS target percentages. Similarly Basol, Balgalmis, Karli and Oz (2016) examined transition from elementary to secondary education system examination (known briefly as TEOG in Turkish) items from 2013 to 2016 in terms of TIMSS-2015 cognitive domain, and they found that the majority of the mathematics items were in applying, and very few were in reasoning. In their study Delil and Yolcu Tetik (2015) analyzed Turkish eighth grade high stakes mathematics examination questions based on TIMSS-2015 framework. They reported that of all eighth grade high stakes mathematics examination items, 48% were in applying and 25% were in reasoning which means that Turkish students face questions that require higher cognitive levels than those in TIMSS. Incikabi, Kurnaz and Pektas (2013) investigated mathematics and science questions in entrance examinations for middle schools in Turkey in terms of their cognitive requirements, and found that both science and mathematics examination items were mostly in applying, and neglected the reasoning cognitive domain. Incikabi (2012) examined cognitive requirements of level determination test (known as SBS in Turkish) mathematics items from 2009 to 2011 in terms of TIMSS-2007 cognitive domain, and found that 55% of the items were in applying, 33% were in knowing and the rest 12% were in reasoning. He also reported that 100% of the items were of multiple-choice.

Cakan (2004) compared the elementary and secondary school teachers' assessment practices, and concluded that while most of elementary school teachers use multiple choice items most frequently, secondary school teachers prefer using essay tests most often than any other item type. Ozmen, Taskin and Guven (2012) examined types of problems used by 7th grade mathematics teachers, and found that they mostly used short response word problems. Tastekinoglu and Aydin (2014), examined 4th graders mathematics examination questions, and they found that of all questions while 67% were in knowing, 18% were in applying and the rest 15% were in reasoning cognitive domains; 80% were in numbers, 16% were in geometric shapes and measures, and 4% were in data display content domains. They also reported that teachers mostly prefer mixed type questions.

The quality of TMTs depends not only on cognitive requirements and on item types, but also depends on the amount of the test item construction errors they involve. According to Oescher and Kirby (1990), formatting was a problem in more than 70% of the tests they examined. Common examples of this deficiency were inconsistent style and margins, and lack of space for answers. They also reported that none of the tests contained a written explanation determining the point value of any test or item. It is well known that the lack of test validity and reliability leads to incomplete, inaccurate and undependable inferences about student learning (Gareis & Grant, 2015, p.11-12).

Literature reveals that the way teachers assess their students is important and problematic most of the times. In this study we aim to examine TMTs; and the main problem is that "what are the cognitive requirements, test construction errors, and item types of teacher-made tests?". By answering this question we will be able to classify the cognitive levels of the test items in terms of TIMSS-2019 cognitive domain, to see test construction errors, and to find out the item types that are most used. By this effort we will be able to compare TIMSS questions and teacher made questions, which may shed light on reasons why Turkish students demonstrate poor performance in TIMSS studies over years. This study may also contribute to the literature as it may give a picture on how teachers assess students in classroom practices.

## Methodology

This is a qualitative study that uses content analysis. According to Krippendorff (2004, p.18), "content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use". In this study, the content analysis technique is used in order to analyze and classify TMT items based on TIMSS-2019 assessment framework and based on Table 3 which is constructed to determine test construction errors. In the following we explain how the documents subject to the study were obtained, reliability, the coding scheme, and classification of the items.

*Data Collection*

The data is consisted of 548 items in 30 exam papers of 18 mathematics teachers from 13 distinct schools of Demirci town of Manisa city in Turkey. Demirci has a reputation that the students in this rural area are the most successful in high school entrance exams, considering their average within Manisa (F.Yuksel†, personal communication, April 29, 2016), and this is the reason why Demirci is chosen for data collection. The written permission of the local National Education Authority was granted in order to be able to collect the 1st and 3rd examination papers during the first

---

† Fevzi YUKSEL, Co-director of Manisa Ministry of National Education Authority.

semester of 2015-2016 academic calendar. Since the central exam scores were considered as the 2nd examination scores in schools, there are no 2nd exam papers included in the data.

*The Coding Scheme and Classification of the Items*

In order to examine the cognitive requirements of the exam papers, TIMSS-2019 cognitive domain is considered. The characteristics of cognitive processes in the cognitive domain, namely knowing, applying, and reasoning, are given in Figure 1. The written permission from TIMSS Science Coordinator was granted to be able to use the coding scheme that is derived from their assessment framework. Also, the coding scheme and the findings for the test construction errors are presented in Table 3.

In order to work out the coding procedure, the coding panel (the researchers) first read Mathematics Cognitive Domains section of TIMSS-2019 Mathematics Framework independently (Mullis & Martin, 2017, p.22). Then they met and read the framework together. They trained on some released eighth grade mathematics item samples of TIMSS-2011 (Foy, Arora, & Stanco, 2013). Then they coded the 20 questions of an examination independently. After the initial coding, the inter-rater reliability coefficient was found as 0.85 which was found large enough (Neuendorf, 2002, p.143) to make valid inferences. They then coded the rest of the questions independently with the most care. Some disagreements occurred and they were discussed until an agreement was reached based on TIMSS-2019 Mathematics Framework. Examples for classifications of TMT items and the eighth grade TIMSS-2011 released items in terms of their cognitive requirements are given in appendices.

According to TIMSS-2019 assessment framework, problem solving abilities depend on the knowledge that student is able to recall and concepts that student understand. So, "knowing" cognitive domain includes recalling, recognizing, classifying/ordering, computing, retrieving and measuring subdomains. "Applying" domain involves the application of mathematics in some contexts, and this domain includes determining, representing/modelling and implementing subdomains. Finally, "reasoning" domain involves logical and systematic thinking, and it includes analyzing, integrating/synthesizing, evaluating, drawing conclusions, generalizing and justifying subdomains. Each subdomain is described in Figure 1 (Mullis & Martin, 2017, p.23-24).

| (1) KNOWING | (2) APPLYING | (3) REASONING |
|---|---|---|
| **1.1 Recall** | **2.1 Determine** | **3.1 Analyze** |
| Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation (e.g., a × b = ab, a + a + a = 3a). | Determine efficient/appropriate operations, strategies, and tools for solving problems for which there are commonly used methods of solution. | Determine, describe, or use relationships among numbers, expressions, quantities, and shapes. |
| **1.2 Recognize** | **2.2 Represent/Model** | **3.2 Integrate/Synthesize** |
| Recognize numbers, expressions, quantities, and shapes. Recognize entities that are mathematically equivalent (e.g., equivalent familiar fractions, decimals, and percents; different orientations of simple geometric figures). | Display data in tables or graphs; create equations, inequalities, geometric figures, or diagrams that model problem situations; and generate equivalent representations for a given mathematical entity or relationship. | Link different elements of knowledge, related representations, and procedures to solve problems. |
| **1.3 Classify/Order** | | **3.3 Evaluate** |
| Classify numbers, expressions, quantities, and shapes by common properties. | | Evaluate alternative problem solving strategies and solutions. |
| **1.4 Compute** | **2.3 Implement** | **3.4 Draw Conclusions** |
| Carry out algorithmic procedures for +, –, ×, ÷, or a combination of these with whole numbers, fractions, decimals, and integers. Carry out straightforward algebraic procedures. | Implement strategies and operations to solve problems involving familiar mathematical concepts and procedures. | Make valid inferences on the basis of information and evidence. |
| **1.5 Retrieve** | | **3.5 Generalize** |
| Retrieve information from graphs, tables, texts, or other sources. | | Make statements that represent relationships in more general and more widely applicable terms. |
| **1.6 Measure** | | **3.6 Justify** |
| Use measuring instruments; and choose appropriate units of measurement. | | Provide mathematical arguments to support a strategy or solution. |

*Figure 1: The coding scheme of the study (Mullis & Martin, 2017).*

**Findings**

The cognitive requirements of all TMT items are summarized in Table 1. It is seen that of all items, about 50% are in knowing, 43% are in applying and 7% are in reasoning cognitive domains. Also, reasoning cognitive domain seems to be the least represented in each examination.

*Table 1. The cognitive requirements of TMT items*

| Exam ID | Number of Items | | | Total |
| | Knowing | Applying | Reasoning | |
|---|---|---|---|---|
| 1 | 11 | 8 | 1 | 20 |
| 2 | 13 | 5 | 0 | 18 |
| 3 | 5 | 9 | 1 | 15 |
| 4 | 10 | 10 | 0 | 20 |
| 5 | 7 | 3 | 1 | 11 |
| 6 | 2 | 6 | 2 | 10 |
| 7 | 7 | 7 | 0 | 14 |
| 8 | 7 | 13 | 0 | 20 |
| 9 | 14 | 6 | 0 | 20 |
| 10 | 12 | 6 | 0 | 18 |
| 11 | 9 | 10 | 1 | 20 |
| 12 | 10 | 10 | 0 | 20 |
| 13 | 7 | 11 | 2 | 20 |
| 14 | 13 | 6 | 0 | 19 |
| 15 | 5 | 10 | 5 | 20 |
| 16 | 15 | 4 | 1 | 20 |
| 17 | 16 | 3 | 1 | 20 |
| 18 | 3 | 14 | 3 | 20 |
| 19 | 10 | 8 | 2 | 20 |
| 20 | 7 | 11 | 3 | 21 |
| 21 | 4 | 3 | 2 | 9 |
| 22 | 5 | 4 | 2 | 11 |
| 23 | 19 | 5 | 0 | 24 |
| 24 | 4 | 13 | 2 | 19 |
| 25 | 16 | 3 | 1 | 20 |
| 26 | 4 | 14 | 1 | 19 |
| 27 | 14 | 6 | 0 | 20 |
| 28 | 3 | 14 | 3 | 20 |
| 29 | 16 | 3 | 1 | 20 |
| 30 | 4 | 13 | 3 | 20 |
| **Total** | **272** | **238** | **38** | **548** |
| | **(49.64%)** | **(43.43%)** | **(6.93%)** | **(100%)** |

When it comes to examining the item types presented in Table 2, it is found that of all 548 items, there are 455 (83%) multiple choice, 2 (0%) True/False and 91 (17%) constructed-response type questions. Also, out of 30 examinations 11 (37%) are mixed type and the rest are of multiple choice.

*Table 2. Item types of TMTs*

| | Multiple Choice | True/False | Constructed Response |
|---|---|---|---|
| TIMSS Target | 50% | - | 50% |
| TMT Items | 83% (455) | 0% (2) | 17% (91) |

In order to find out the test construction errors that TMTs involve, we use the criteria in the following table. There are questions under the titles "Face validity", "Directions", "Typing", "Spacing", "Ambiguous words/phrases" and "General look".

*Table 3. Test construction errors of TMTs*

| Title | Criteria | | Yes/Good | No/Bad |
|---|---|---|---|---|
| (1) Face validity | Is there a title? | | 30 | 0 |
| | Does the exam paper have face validity? | | 30 | 0 |
| | | **Total** | **60** | **0** |
| (2) Directions | Directions given at the beginning of the exam sheet? | | 0 | 30 |
| | Duration of the exam is stated? | | 12 | 18 |
| | Point values for each item is stated? | | 21 | 9 |
| | Rules are clearly stated? | | 0 | 30 |
| | Directions for answer key is given? | | 0 | 30 |
| | Correction rule is given? | | 0 | 30 |
| | Avoiding directions for taking wrong exam is given? (For example teacher's name is given) | | 16 | 14 |
| | | **Total** | **49** | **161** |
| (3) Typing | Exam items are not taken as copy/paste pictures? | | 4 | 26 |
| | Font types/sizes are the same throughout the exam sheet? | | 4 | 26 |
| | Each item number is given and is in order? | | 27 | 3 |
| | In negative items, the negative words are underlined? | | 14 | 6 |
| | There is no punctuation or writing error? | | 19 | 11 |
| | | **Total** | **68** | **72** |
| (4) Spacing | There is space between items to distinguish them? | | 17 | 13 |
| | There is enough space for solutions/answers? | | 8 | 22 |
| | | **Total** | **25** | **35** |
| (5) Ambiguous words/phrases | Apart from items are there not any ambiguous words/phrases/ pictures/advices etc. on the exam sheet? | | 8 | 22 |
| | | **Total** | **8** | **22** |
| (6) General look | Is there space left for student identity info? | | 29 | 1 |
| | For mixed type examinations, is there any instruction for where different type of questions start and finish? | | 1 | 10 |
| | For examinations that include multiple choice items, is there A/B groups? | | 1 | 27 |
| | | **Total** | **31** | **38** |
| | | **Grand Total** | **241** | **328** |

It is seen from the Table 3 that out of 569 points, there are 241 points that are affirmative. For example, concerning "face validity" title, it is seen that each exam paper seem to be about mathematics. While for "directions", there are 49 points out of 210 points that are affirmative, there are 68 points out of 140 points related with "typing" that are affirmative. Also for "spacing", "ambiguous words/phrases" and "general look" titles there 25 out of 60, 8 out of 30 and 31 out of 69 points that are affirmative respectively which may mean that in most of the cases TMTs involve test construction errors.

## Discussion and Recommendations

The distribution of cognitive demands of all TMTs indicates that there is a strong emphasis on knowing or applying cognitive domains, with a total percentage of 93 (Table 1). Since the target percentages of knowing, applying and reasoning cognitive domains of TIMSS-2019 study are 35, 40 and 25 respectively, this result is inconsistent with the target percentages of TIMSS mathematics framework which shows that 75 percentages of TIMSS mathematics items require knowing or applying in total (Figure 2). So, it means that TMT items require lower cognitive demands than those in TIMSS, and this may be a reason behind the undesired results of Turkish eighth graders in TIMSS assessments over the years. See Figure 2 for a comparison of TEOG, TMT and TIMSS questions in terms of their cognitive domains.
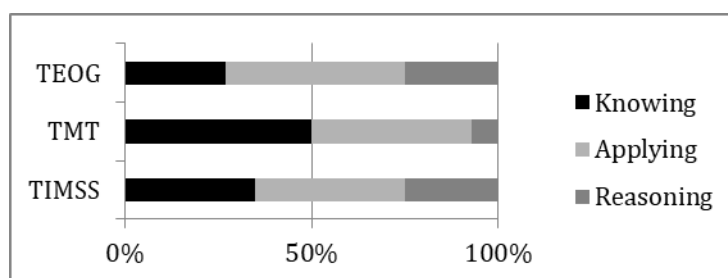


*Figure 2. TEOG, TMT and TIMSS questions percentages in terms of their cognitive domains*

Also, TEOG examination questions' cognitive domains were previously reported to be 27% knowing, 48% applying and 25% reasoning by Delil and Yolcu Tetik (2015). It can be interpreted as TMT items require lower cognitive demands than those in TEOG examination questions, too. The reason why TMT items require lower cognitive levels comparing TEOG and TIMSS items may be because teachers are doing assessments mostly for learning, and another reason is because students' individual averages of their school courses effect their TEOG scores, teachers tend ask questions of lower cognitive levels in order to keep student mathematics course scores higher. So, teachers are recommended to give more attention to prepare items of higher cognitive levels.

There are dramatic changes from teacher to teacher in terms of their questions' cognitive domains (Table 1). For example, out of 24 items in exam paper with ID #23, while there are 19 items in knowing, there are 5 in applying and there aren't any in reasoning cognitive domain. But, in exam paper with ID #28, there are 3 in knowing, 14 in applying and 3 in reasoning cognitive domain. Considering their cognitive domains, an ideal example of a TMT seems to be the one with ID #15 which has 5 items in knowing, 10 in applying and 5 in reasoning. This result indicates that teachers don't consider the cognitive domains which are closely related with the quality of their exams. It is also observed that teachers have many questions in common, which may mean that questions are copy-pasted from the internet or other resources. Thinking about students' abilities in using technology and getting the questions, this may negatively affect the reliability of their assessment which is not desired. Consequently, it is suggested that teachers should not use tests or questions directly from the internet or other resources.

Since 83% of all questions are of multiple choices, 17% are constructed-response type and 0% are true-false, teachers seem to prefer multiple choice item type. This may be understandable since they face a big pressure from the families and school management about student success in high stakes examinations that are fully multiple choice type in Turkey. But, it is stated in Turkish mathematics curriculum that in order for students to present their potentials, it is important to enrich assessment tools (Ministry of National Education, 2013). According to Ben-Simon and Cohen (2004), being familiar with different kinds of item formats like multiple choice, constructed response and performance assessment tasks may positively affect student performance. One reason of the low performance of Turkish students in international surveys like TIMSS may be the students' unfamiliarity with items of mixed types during their classroom assessments. So, teachers are suggested to use more mixed type tests.

Another quality indicator of the TMTs is the lesser amount of test construction errors that they involve. Findings reveal that except "face validity" there are problems concerning test constructions of teachers. For example, only for "directions" of the tests alone there are 161 errors made out of 210 criteria. So, it is seen that teachers are omitting directions in their tests. "Typing" is also a problem since out of 140 criteria there are 72 errors made. As for "spacing", "ambiguous words/phrases" and "general look" there are 25, 8 and 31 errors out of 60, 30 and 69 criteria respectively. Test construction errors are very important as they may reduce the reliability of TMTs. Considering the data collected from Demirci that have a good reputation in success of high stakes tests, TMTs in the rest of Manisa (or even Turkey) may be in a worse situation. According to Cakan (2004), most of the teachers in both elementary and secondary schools find themselves incapable of using assessment and measurement techniques. Consequently, as a remedy, measurement and evaluation specialists can be employed in each school or in each local Ministry of National Education Authority at least, in order to support teachers, and this is not nonsense since a similar action is taken for psychological counseling and guidance specialists in schools all over Turkey. If this is not possible, in-service training programs on measurement and evaluation for teachers to participate in are vital.

## References

Basol, G., Balgalmis, E., Karli, M. G., & Oz, F. B. (2016). Content analysis of TEOG mathematics items based on MoNE attainments, TIMSS levels, and reformed Bloom Taxonomy. *Journal of Human Sciences, 13*(3), 5945-5967. doi:10.14687/jhs.v13i3.4326

Ben-Simon, A., & Cohen, Y. (2004, June). *International assessments: merits and pitfalls.* Paper presented at the 30th Annual Conference of the International Association for Educational Assessment, Philadelphia, USA.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives, the classification of educational goals. Handbook I: Cognitive domain.* New York, NY: David McKay.

Cakan, M. (2004). Comparison of elementary and secondary school teachers in terms of their assessment practices and perceptions toward their qualification levels. *Ankara University, Journal of Faculty of Educational Sciences, 37*(2), 99-114.

Delil, A. (2019). How fifth graders are assessed through central exams in Turkey: A comparison with TIMSS 2019 Assessment Framework. *International Online Journal of Educational Sciences, 3*(11), 222-234.

Delil, A., & Yolcu Tetik, B. (2015). An Analysis of Turkish Eighth Grade High Stakes Mathematics Examination Questions Based on TIMSS-2015 Framework. *Celal Bayar University Journal of Social Sciences, 13*(4), 165-184. doi:10.18026/cbusos.87313

DiDonato-Barnes, N., Fives, H., & Krause , E. S. (2014). Using a table of specifications to improve teacher-constructed traditional tests: an experimental design. *Assessment in Education: Principles, Policy & Practice, 21*(1), 90–108. doi:10.1080/0969594X.2013.808173

Foy, P., Arora, A., & Stanco, G. M. (Eds.) (2013). *TIMSS 2011 User Guide for the International Database: Released Items, Mathematics-Eighth Grade.* Boston, MA: TIMSS & PIRLS International Study Center.

Gareis, C. R., & Grant, L. W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (2nd ed.). New York, NY: Routledge.

Incikabi, L. (2012). After the reform in Turkey: a content analysis of SBS and TIMSS assessment in terms of mathematics content, cognitive domains, and item types. *Education as Change, 16*(2), 301-312.

Incikabi, L., Kurnaz, M. A., & Pektas, M. (2013). An Investigation of Mathematics and Science Questions in Entrance Examinations for Secondary Education Institutions in Turkey. *Journal of Baltic Science Education, 12*(3), 352-364.

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Marso, R. N., & Pigge, F. L. (1991). An analysis of teacher-made tests: item types, cognitive demands, and item construction errors. *Contemporary Educational Psychology, 16*(3), 279-286.

Marzano, R. J., & Kendall, J. S. (2008). *Designing and assessing educational objectives: Applying the new taxonomy.* Thausand Oaks, CA. Corwin Press.

Ministry of National Education. (2013). *Ortaokul matematik dersi ogretim programi* [Secondary school mathematics curriculum]. Ankara: Ministry of National Education.

Mullis, I. V., & Martin, M. O. (2017). *TIMSS-2019 Assessment Frameworks.* Boston: Chestnut Hill, MA. Retrieved 08 17, 2016, from http://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I. V., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* TIMSS & PIRLS International Study Center. Retrieved 04 25, 2017, from http://timssandpirls.bc.edu/timss2015/international-results/

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thausand Oaks,CA: Sage Publications.

Oescher, J., & Kirby, P. C. (1990, April). *Assessing teacher-made tests in secondary math and science classrooms.* Paper presented at Annual Meeting of the National Council on Measurement in Education, Boston, MA, USA.

Ozmen, Z. M., Taskin, D., & Guven, B. (2012). Determining the types of problems used by 7th grade maths teachers. *Education and Science, 37*(165), 246-261.

Tastekinoglu, E., & Aydin, G. (2014). A comparison of 4th grade mathematics exam items with TIMSS 2011 cognitive domain and mathematics curriculum aqusitions. In *Proceedings of 3rd Turkish Congress of Graduate Studies* (pp. 247-264). Sakarya, Turkey: Sakarya University.

Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2010). *Elementary & Middle School Mathematics: Teaching Developmentally* (7th ed.). Boston, MA: Pearson Allyn & Bacon.

**Appendices**

**Appendix 1:** Examples for classifications of TMT items and the eighth grade TIMSS-2011 released items are given below. While the classification of the TMT items was done by the researchers, the classification of the TIMSS released items has been done by TIMSS developers.

*Example 1:* (TMT item #17 in the 1st exam paper)



The numbers in the figure above are the distances in meters of a squirrel away from the trees. Which tree is the squirrel closest to?

(A) Quince     (B) Almond     (C) Walnut     (D) Mulberry

The item above is classified in "analyze" of *reasoning* cognitive domain.

*Example 2:* (TMT item #15 in the 1st exam paper)

What is the half of the number $4^{20}$?

(A) $2^{10}$     (B) $2^{39}$     (C) $2^{15}$     (D) $4^{10}$

The cognitive requirement of the item above is "determine" of *applying* cognitive domain.

*Example 3:* (TMT item #1 in the 1st exam paper)

Choose the correct value of $\underbrace{3 \cdot 3 \cdots\cdots 3}_{6 \; times}$ in the following.

(A) $6^3$     (B) $3 \cdot 6$     (C) $3^6$     (D) $3 + 6$

The question above is attained in "recall" of *knowing* cognitive domain.

In the following, examples of TIMSS-2011 released items that cognitive domains are classified by TIMSS developers are given (Foy, Arora, & Stanco, 2013).

*Example 4:* (TIMSS-2011 Released Items: Page: 23, ID: M052216)

---

Which number is equal to $\frac{3}{5}$?

(A)  0.8   (B)  0.6   (C)  0.53   (D)  0.35

Cognitive domain attained: *Knowing*

---

*Example 5:* (TIMSS-2011 Released Items: Page: 26, ID: M052061)

---

Kim is packing eggs into boxes. Each box holds 6 eggs. She has 94 eggs. What is the smallest number of boxes she needs to pack all eggs?

Answer: -------------------- eggs

Cognitive domain attained: *Applying*

---

*Example 6:* (TIMSS-2011 Released Items: Page: 48, ID: M042186)

---

Here is a pattern:

3-3=0

3-2=1

3-1=2

3-0=3


What will the next line in the pattern be?

Answer: --------------------

Cognitive domain attained: *Reasoning*

---

**Appendix 2:** In the following we give examples of phrases/advices/pictures placed on exam sheets, which may distract students and reduce the reliability of test scores.

*Example 1:* In the following, there is a picture of an ambulance placed at the left bottom of an exam sheet, which is not related with the questions asked.



*Figure 3: A picture of an ambulance in TMTs.*

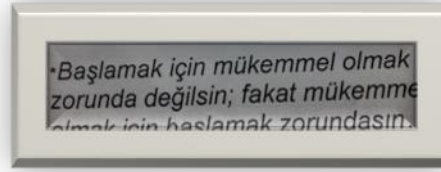***Example 2:*** In the following, there is a phrase written on the bottom of an exam paper.



*Figure 4: A philosophical phrase on an exam paper: "You don't have to be perfect in order to start, but you have to start in order to be perfect".*

***Example 3:*** In the following example, a pray is placed at the bottom of the exam sheet in order to make students smile and motivate them, perhaps.
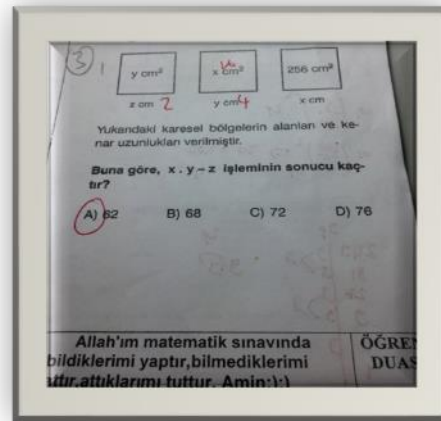


*Figure 5: A prayer on an exam paper: "STUDENT PRAYER: Oh God! Let me do what I know in maths exam, let me take a wild guess on what I don't know, let my guess be true. Amen :) :)"*