THE INVESTIGATION OF COGNITIVE PROCESSES IN MATHEMATICS LEARNING WITH ITEM RESPONSE THEORY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SELCEN ÖZKAYA SEÇİL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
SECONDARY SCIENCE AND MATHEMATICS EDUCATION

SEPTEMBER 2009

Approval of the thesis

THE INVESTIGATION OF COGNITIVE PROCESSES IN MATHEMATICS LEARNING WITH ITEM RESPONSE THEORY

submitted by **SELCEN ÖZKAYA SEÇİL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Secondary Science and Mathematics Education, Middle East Technical University** by,

Prof. Dr. Canan Özgen ____________________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ömer Geban ____________________
Head of Department, **Secondary Science and Mathematics Education**

Prof. Dr. Giray Berberoğlu ____________________
Supervisor, **Secondary Science and Mathematics Education**

Examining Committee Members

Prof. Dr. Petek Aşkar ______________
Computer Education & Instructional Technologies Dept. HU

Prof. Dr. Giray Berberoğlu ______________
Secondary Science and Mathematics Education Dept., METU

Prof. Dr. Ömer Geban ______________
Secondary Science and Mathematics Education Dept., METU

Assoc. Prof. Dr. Oya Yerin Güneri ______________
Elementary Education Department, METU

Inst. Dr. Omer Faruk Ozdemir ______________
Secondary Science and Mathematics Education Dept., METU

Date: 09.09.2009

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Selcen ÖZKAYA SEÇİL
Signature :

ABSTRACT

THE INVESTIGATION OF COGNITIVE PROCESSES IN MATHEMATICS LEARNING WITH ITEM RESPONSE THEORY

Özkaya Seçil, Selcen
Ph.D., Department of Secondary Science and Mathematics Education
Supervisor: Prof. Dr. Giray Berberoğlu

September 2009, 112 pages

The importance of learning mathematics and using it in daily life is obvious. On the other hand, the results from many national and international assessment studies show that the achievement of Turkish students are very far away from the bare minimum performance. However, in the measurement and evaluation procedures of both primary and secondary educational system, there is a lack of identification of this "bare minimum" or qualitative and clear descriptors for performance levels. A great importance is dedicated to the national exam results expressed in percentage terms of the correct responses, or in total score points in weighted scale scores, but there is still no system of presenting to students their scores with descriptions of these scores in terms of levels of skills that they did or did not reach.

Therefore, this study has aimed to identify the knowledge and skills required for different performance levels defined by setting cut points for the results of a 4th grade mathematics achievement test. The test was conducted in 2007-2008 academic year with 269 fourth grade students in eight different private primary schools in Istanbul. Then, in 2008-2009 academic year, a group of ten teachers of mathematics and assessment experts took part in the study for identifying the performance level descriptors for 4th grade mathematics performance. Two different methods of standard setting were used. One of the methods was based on the one-parameter model of Item Response Theory (IRT) and mostly named as Bookmark Method. The method depended on the statistical identification of the cut points on the scale for performance levels such as Below Basic, Basic, Proficient, and Advanced. The other method was a judgmental method which required the participant teachers to classify the item as carrying the characteristics of performance levels, again, as Below Basic, Basic, Proficient, and Advanced.

The study revealed that the item mappings from two methods were congruent to each other. There was a hierarchical ordering in terms of skills among the performance levels. Also, the results demonstrated that understanding and computation skills were heavily characteristics of Below Basic and Basic levels, whereas, problem solving skill was reached by the students of Proficient and Advanced levels.

Keywords: Mathematics performance, Standard Setting, Item Response Theory, Bookmark Method, Judgmental Method

ÖZ

MATEMATİK ÖĞRENİMİNDEKİ BİLİŞSEL SÜREÇLERİN MADE TEPKİ KURAMIYLA İNCELENMESİ

Özkaya Seçil, Selcen
Doktora, Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü
Tez Yöneticisi: Prof. Dr. Giray Berberoğlu

Eylül 2009, 112 sayfa

Matematik öğrenmenin ve onu günlük yaşam içinde kullanabilmenin önemi açıktır. Öte yandan bir çok ulusal ve uluslararası değerlendirme çalışmasının sonuçları gösteriyor ki, Türk öğrencilerin başarıları, yeterli minimum performanstan çok uzak görünmektedir. Bununla birlikte, hem ilköğretim hem de ortaöğretim ölçme ve değerlendirme sistemlerinde, bu "yeterli minimum" un tanımlanmasında veya performans düzeylerinin açık ve nitel tanımlanmalarında bir eksiklik de yer almaktadır. Ulusal sınavların, doğru cevap yüzdeleri ya da ağırlıklı bölüm puanlarının toplamı olarak ilan edilen sonuçlarına büyük önem adledilmesine rağmen, bu puanların öğrencilere, ulaştıkları ya da ulaşamadıkları beceri düzeylerinin tanımları olarak sunulduğu bir sistem bulunmamaktadır.

Bu nedenle, bu çalışma, bir dördüncü sınıf matematik sınavı sonuçlarının, kesim noktaları belirlenmek suretiyle oluşturulan farklı performans düzeylerinin gerektirdiği bilgi ve becerilerin tanımlanmasını amaçlamaktadır. Sınav, 2007-2008 eğitim öğretim yılında, İstanbul'daki sekiz özel ilköğretim okulunda okuyan 269 dördüncü sınıf öğrencisine uygulanmıştır. 2008-2009 eğitim öğretim yılında ise, 10 matematik öğretmeni ve ölçme değerlendirme uzmanından oluşan bir grup, dördüncü sınıf matematik performans düzeyi tanımlayıcılarının belirlenmesi çalışmasında yer almışlardır. İki farklı standart belirleme yöntemi kullanılmıştır. Bunlardan ilki, Ayraç Yöntemi de denen ve tek parametreli Madde Tepki Kuramı'na dayanan bir yöntemdir. Yöntem, Basit, Temel, Yetkin ve İleri olarak adlandırılan performans düzeylerinin ölçek üzerinde kesim noktalarının istatistiksel olarak tanımlanmasına dayanmaktadır. Diğer yöntem ise, yine Basit, Temel, Yetkin ve İleri performans düzeylerinin, katılımcı öğretmenlerin soruları, bu düzeylerin özelliklerine uygunluğuna göre kategorize etmelerine dayanmaktadır.

Çalışma, iki yöntemler elde edilen madde haritalarının birbirleriyle uyumlu olduğunu göstermiştir. Performans düzeyleri arasında, beceriler açısından hiyerarşik bir sıralama da oluşmuştur. Ayrıca, sonuçlar göstermiştir ki, anlama ve işlem becerileri daha çok Basit ve Temel düzeylerin; problem çözme becerisi ise Yetkin ve İleri düzeylerin karakteristik özelliği olmaktadır.

Anahtar Kelimeler: Matematik başarısı, Standart Belirleme, Madde Tepki Kuramı, Ayraç Yöntemi, Uzman Kanısı Yöntemi

This thesis is dedicated to

My parents Tülay and Cengiz Özkaya

My sister Ayça Özkaya

and to my husband Güray Seçil

ACKNOWLEDGEMENTS

I firstly want to express my gratitude to my supervisor Prof. Dr. Giray Berberoğlu for his continuous support and help. He has always been kind, sincere, and patient.

I want to acknowledge and thank to my colleagues Esin Keskin, Remzi Harman, Emra Sarıkaya, and Gamze Kaya for their support in constructing, developing, and analysing the items used for this study. Also, I want to thank to all students and teachers of ISTEK Schools participated in this study and to the management –especially to Deniz Demirtaş- for their great vision to let this study happen.

I am grateful to the teachers attended the standard setting sessions especially to Gülşah Çekiç, Filiz Uz, and Nağme Karamustafaoğlu, not only for their participation in the study, but also for their beneficial recommendations and their sincere friendship.

I want to thank my friends and colleagues Özlem Mecit and Selen Sencar Tokgöz, with whom I have shared a lot of experience during the whole period of my doctoral study.

I am also grateful to my parents Tülay and Cengiz Özkaya whose love and confidence have always been with me; and to my dear sister Ayça Özkaya, who has been with me again to help, support, and encourage.

Finally, I want to express my love and gratitude once more to my dear husband Güray for not giving up believing and encouraging me. Without his support, confidence, and patience, this study would have never been finished.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

As I graduated from university to begin my professional career as a mathematician, it soon became clear to me that my chosen field was not something that was just taught, but something that could also be applied to everyday life. It helped me to think in a way that helped me to solve problems that I had to face, and also to argue for, or against, points of view held by myself, or others. It became clear to me that mathematics was not learnt for mathematics sake but that it could be applied to everyday life.

Through the ages, mathematics has always been considered one of the elitist disciplines. It has defined the position of individuals in society. The knowledge of mathematics conferred status. It has also been considered to preserve of males. Women were considered inferior to men when it came to mathematics. Its philosophically importance in ancient Greece is best summed up by the following inscription that was said to have been carved over the entrance to Plato's Academy:

"Let no one destitute of geometry enter my doors"

In this age of information and technology those lacking skills or knowledge of mathematics would also find entrance Plato's Academy barred. Today, they would also find it difficult to find job opportunities in many areas, enter university, or even pursue their high school careers in institutions that demand a high academic performance. Here in Turkey, there is a great demand, and competition, for admittance to Anatolian High Schools and Science High Schools. Both of these types of institutions have a more demanding mathematics and science curricula. For the 2009-2010 academic year, the number of applicants for these institutions was 764,623. (Egitek, 2009)

Despite this high demand for entrance to these schools, the average mathematics performance in entrance exams is disturbingly low. The mean score of mathematics, for 8th grade students in the 2009 SBS examination , which is a countrywide summative assessment consisting of Turkish, mathematics, science, social science and English was 2.35 based on 20 items (Egitek, 2009). It should be pointed out, that this low performance in mathematics is not common to this examination, or year. The mean mathematics score for the 2007 OKS examination (former form of SBS), based on 25 items was 3.35 (Egitek, 2007). In the 2006 OSS examination, the countrywide university entrance examination, the mean mathematics, for the first section, was 8.5 based on 30 items (OSYM, 2006).

These poor performances in mathematics, expressed in percentage terms of correct responses to questions, clearly demonstrate the problems of the teaching and learning of mathematics. However, these statistics reveal nothing about the weaknesses of the students, in terms of the knowledge and skills that they require. In other words, there has been no attempt, to date, to identify the students' performance qualitatively in terms of the descriptors spotting levels attained or not attained.

Besides these unsuccessful results in the examinations conducted in Turkey, results of the international assessment studies revealed the Turkish students' low performances (EARGED, 2005). In Programme for International Student Assessment (PISA) 2003, for example, 27.7 % percent of the Turkish students were under the first level among the six hierarchical levels described (EARGED, 2005). The first level of performance in mathematics in Programme for International Student assessment (PISA) was defined as follows:

"At Level 1, students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli." (OECD, 2004; p:47).

When the performance level description of this first level is examined, the skills required for reaching this level are defined as explicit, clear, and routine. This level requires very basic knowledge and skills for the fundamental mathematics achievement. However, more than one fourth of the students could not reach this basic standard. There had been a number of studies investigating the possible reasons of this result (Yıldırım, 2006; Çet, 2006; İş Güzel, 2006). Another international assessment study called Trends in Mathematics and Science Study (TIMSS) revealed, unfortunately, the same degree of low performance for 8th grade students in mathematics. 59% of the Turkish students were cumulated on the lowest level of performance called "Low International Benchmark" identified as follows: "Students have some knowledge of whole numbers and decimals, operations, and basic graphs."

These international studies are important not only because of their indication of the Turkish students' comparative success in these exams based on the rank order among other countries, but also their clear descriptions of the level of performances. As mentioned above, the results of the national assessment studies are announced based on the scores, ranks or both; however, the corresponding levels of performances are never revealed. In recent years, this situation has started to emerge problems after some schools started to introduce with some international programmes and to implement them integrated with the national curriculum. For example, many private schools and few state schools are implementing the programmes of International Baccalaureate Organization (IBO). IBO is an international, independent, and non-profit organization which presents three different educational programmes for the students all around the world (www.ibo.org). The assessment principles of these programmes stand on indentifying the performances of students in terms their weaknesses and strengths and, as a result, to develop their understanding and performance in all areas.

How this contradicts with the national curriculum's requirements is obvious, because of the high dedication to exams and lack of defining the performances of students qualitatively as explained above. This contradiction leads to a handicapped implementation which involves a double assessment of students: one for the national programme which mostly based on the percentages or ratios of correct responses and one for the international programme based on describing the performances and identifying the students' level on these performances. However, it is, of course, not an efficient way of school assessment system.

The improvements in the national curriculum, on the other hand, were based on the needs students and requirements of changing world (MEB, 2005). In the introduction part of the new Primary School Programme for mathematics, it is mentioned that programme was prepared with respect to the national and international research studies, programmes of other countries, and the experiences gained in our country (MEB, 2005). In the assessment system, as well as the content and methods of teaching and learning, occurred a lot of changes. The sources of information for student assessment varied by including the performance tasks and projects; however, there is still the lack of overall description of student level of performance in terms of knowledge and skills.

1.1 The purpose of the study

Under the light of the above discussions, the main purpose of this study was to identify the skills and cognitive processes for mathematics learning to overcome the above mentioned problems to some extent. To this purpose, a mathematics achievement test constructed for the aim of detecting the degree of reaching the outcomes in the 4th grade curriculum was used. The levels of performance in mathematics was attempted to be identified by putting cut points for different levels with two different methods of standard setting.

The overall design of the study can be summarized as follows: The 4th grade students from 8 different private primary schools were administered an achievement test in the context of mathematics curriculum of Ministry of Education. The results of the test were analysed by using Item Response Theory (IRT) models. Then, two methods of standard setting were used to identify the levels of performance of students who were grouped in terms of the mathematical skills. One of these methods was based on the IRT techniques; on the other hand, the second one was mostly derived from the judgments of experts. For validity evidences of these groupings, then, path analyses indicating the prerequisite relationships among the skills were also analysed.

1.2 Research questions

To these purposes of the study, the following research questions were asked and tried to be answered in the context of this study:

1. What cognitive characteristics exemplify different performance levels of 4th grade mathematics students?
    a. What are the characteristics of skills that were demonstrated by students at different score level in mathematics?
    b. Is there any prerequisite relationship among the skills that were achieved by the students at different performance levels in mathematics?
2. Is there congruence between the item mapping results of two standard setting methods: one IRT-based method (Bookmark Method) and one judgmental method?
3. What factors effect judges decisions in categorising the items with respect to performance levels in mathematics?
4. Is there any other format consideration in describing the skills achieved at different performance level of students?

1.3 Definition of important terms

The following terms were frequently used in the rest of the study; therefore, they were explained shortly here and in detail in the related places in the following chapters.

Standard Setting: According to Cizek, standard setting is "the process of establishing one or more cut scores on examinations. The cut scores divide the distribution of examinees' test performances into two or more categories." (2007, p.5).

Bookmark Method (IRT Based Method): This method based on the judges' placements the "bookmarks" in the *ordered item booklet* where the items were ordered from the easiest to hardest (Cizek, 2007).

Ordered Item Booklet: The items in the Ordered Item Booklet were placed with respect to their difficulties. The item difficulty parameters were detected by using the IRT model.

IRT Model: In this study, one-parameter model or in other words, Rasch model was used to detect the difficulty and ability parameters (Hambleton, Swaminathan & Rogers, 1991).

Judgmental Method: This method of standard setting based on the expert judges' decisions for determining the categories. The judges decide the discrimination among the categories with respect to the outcomes they measure (Berberoğlu, Demirtaşlı, İş Güzel, & Konak, 2008).

Performance Levels: The scale scores were grouped into four different categories named as Below Basic, Basic, Proficient, and Advanced.

1.4 Significance of the study

As explained above, although the great improvements in the national curriculum in mathematics programme and the general assessment system, there is still a large need to identify the skills corresponding to each group of scores and to define the national standards with respect to the high stake national exams like SBS and ÖSS. There are comparative studies based on the international assessment programmes like PISA and TIMSS (Yıldırım, 2006; Çet, 2006; İş Güzel, 2006), however, the number of studies on national assessments is very limited (Berberoğlu, Demirtaşlı, İş Güzel, & Konak, 2008).

This study can be taken as a preliminary attempt to identify the standards for the performance in mathematics. In the context of mathematics assessment that had been conducted for 4th grade students from 8 different private primary schools, the study targeted to describe the performances of those students in terms the knowledge and skills required to be accounted as the member of performance levels such as Below Basic, Basic, Proficient, and Advanced. As explained before, the achievement of students should be elaborated more than just the percentage corrects scores.

Also, the study aimed to define the relationships among the mathematical skills such as understanding, computation or problem solving, and among the sub-dimensions of them. This attempt to define the relationships would reveal the conceptual hierarchy among the mathematical processes or the prerequisite skills needed to gain higher order ones. Several recommendations for teachers in terms of teaching strategies and techniques could be derived from the results of the study.

Lastly, the study, on the other hand, focused on the teachers' thoughts and approaches to the capacity of the students in terms of their ability or probability to answer a group of items. Since the study was mostly based on the teachers' judgments on the students' capability to answer the questions, it could be observed how the teachers had conceptualized the students' thinking processes. Factors affecting teachers' decisions in terms of the differences in the information and feedback given to the teachers were investigated and related implications were discussed.

CHAPTER 2

REVIEW OF RELATED LITERATURE

In this chapter, the literature about related studies will be reviewed. The chapter includes the review of studies focusing on the mathematical knowledge and skills and ways of identifying them, the studies on setting standards or cut points for performance levels, methods for standard setting and comparison of these methods, and the factors affecting the participants' decisions during standard setting studies.

2.1 Identifying the Mathematical Knowledge and Skills

In this section, the approach of several studies and programmes to the mathematical competency, understanding, and skills were examined. Since the current study focused basically on identifying meaning of the scores taken from the certain test, it was important to review the curriculum that the current study referred to and the other programmes. Therefore, primary mathematics programme of Turkish Ministry of Education, mathematics subjects of Primary Years and Middle Years Programmes of International Baccalaureate Organization, mathematics parts of PISA and TIMSS programmes were reviewed and compared in terms of content areas, competencies, and skills.

2.1.1 International Baccalaureate Primary Years Programme

International Baccalaureate Organization (IBO) is an international, independent, and non-profit organization which presents three different educational programmes for students aged 3-19 in 2715 schools all around the world (www.ibo.org). The mission statement of IBO best explains its way of approach to education:

> "The International Baccalaureate Organization aims to develop inquiring, knowledgeable and caring young people who help to create a better and more peaceful world through intercultural understanding and respect. To this end the IBO works with schools, governments and international organizations to develop challenging programmes of international education and rigorous assessment. These programmes encourage students across the world to become active, compassionate and lifelong learners who understand that other people, with their differences, can also be right." (IBO, 2007a; p:v).

Primary Years Programme (PYP) is one of those three programmes of IBO which covers the ages of 3-12 and follows an inquiry-based approach (IBO, 2007a). It depends on the construction of knowledge by the students, focuses the concepts and skills, aims to use varied assessment procedures, and promotes international-mindedness (IBO, 2002).

PYP Mathematics, on the other hand, describes mathematics as a language which gives the students a way of constructing meaning (IBO, 2003). The framework of the programme in terms of the mathematical content is like as follows:

Data Handling: Statistics and probability

Measurement

Shape and Space

Pattern and Function

Number

Also, PYP presents the skills required for learning mathematics or "the stages how best the students will learn" as follows (IBO, 2003; p.3.2):

Constructing meaning

Transferring meaning into signs and symbols

Understanding and applying

2.1.2 International Baccalaureate Middle Years Programme

As a part of the programme, mathematics courses should have goals to ensure the requirements of three fundamental concepts and five areas of instruction (IBO, 2007b). Each topic in the mathematics programme should have a relation or rationale to adopt the fundamental concepts and/or areas of interaction. That is, central idea of MYP is not to teach the content but to teach the concepts and skills.

Although MYP provides the possibility of a flexible curriculum, a framework for the content of mathematics is given. This framework is five topics to cover in every grade level of MYP. The concepts and skills of this framework comprise the following areas:

Number: Numerals, decimals, ordinality, cardinality, divisibility, pattern, number sets

Algebra: Variables, relations, functions, expressions, equations, coordinate systems, inequalities, sequences

Geometry and Trigonometry: Shapes, mensuration, similarity, enlargement, angles, vectors, Pythagoras' Theorem

Statistics and Probability: Discrete and continuous data, graphical analysis, sampling, probability

Discrete Mathematics: Sets, Venn diagrams, logic, networks, trees

These topics may have different weights in each year curriculum with respect to the grade level, the requirements of national programmes, and the objectives of the school. Also, the schools may decide on the subtopics of this framework.

MYP, also, states targets for learning mathematics in terms of the skills specific to mathematics (IBO, 2007). These objectives are the basis for the final assessment criteria of MYP. Every student should reach to a predefined level in terms of these criteria to finish MYP successfully. The following are the four basic objectives of MYP Mathematics (IBO, 2000: pp.16-17):

A Knowledge and Understanding

To know and understand concepts, and demonstrate skills from five branches

To be able to use a variety of mathematical forms and to move confidently between them

B Investigating patterns

To select and use appropriate mathematical knowledge when investigating problems

To select and apply mathematical skills and techniques when investigating problems

To recognise patterns and structures and describe them as relationships or rules when investigating problems

To draw conclusions consistent with findings

To justify mathematical relationships when investigating problems

C Communication in Mathematics

To communicate mathematical facts, ideas, methods, results, and conclusions using appropriate language and symbols, and a variety of media and technologies

D Reflection in Mathematics

To reflect on their methods and processes

To consider possible alternative approaches

To evaluate the significance and reliability of findings

2.1.3 Programme for International Student Assessment

Another important study which presented a framework for defining the mathematics performance in terms of the concepts and skills that mathematics required was the Programme for International Student Assessment (PISA). The study was a product of Organisation for Economic Co-operation and Development (OECD) and cooperatively developed by the participating countries.

Mathematical Competencies

Thinking and reasoning

Argumentation

Communication

Modelling

Problem posing and solving

Representation

Using symbolic, formal, and technical language and operations

Use of aids and tools

Content-"overarching ideas"

Quantity

Space and shape

Change and Relationships

Uncertainty

2.1.4 Program of National Ministry of Education

The Ministry mathematics program does state the target outcomes in detail for every content area and sub-area without identifying the global skills required for every content area. In other words, it can be identified as a more “content-based” programme rather than a “skill-based” one. A criticism and a recommended alternative for this situation were made by the Academic Steering Committee of CİTO Turkey (2008). The Committee stated that the National Programme had mentioned the skills that should have been developed by the mathematics education through grades 1-5 consistently along the general approach of the system, however, when the learning outcomes had been presented; they had been stuck into the content areas and had had to be repeated (İş Güzel, 2008).

2.1.5 Trends in International Mathematics & Science Study

International Association for the Evaluation of Educational Achievement (IEA) has been developing international assessment studies in various countries around the world to compare the students’ performances in diverse areas for years (Martin, Mullis & Foy, 2008). One of the famous studies of IEA is Trends in International Mathematics and Science Study (TIMSS), which was conducted in 59 countries in 2007. The study consisted of the following content area and skills:

Content

Number

Geometric Shapes and Measures

Data Display

Skills

Knowing

Applying

Reasoning

2.1.6 Comparison of All Programs

The content and skill developments of five different national and international programmes for learning and teaching mathematics were summarized until here. These programmes presented the similar content and expected the students to gain similar skills for being counted as good learners of mathematics, although they were developed for different purposes and recommended diverse methods of learning and teaching mathematics. The following two tables (Table 2.1 and 2.2) present a summary of these programmes in terms of the mathematical content that they include and the skills and competencies that they require.

Table 2.1 Comparison of Different International Mathematics Programs in terms of Content

	PISA	**PYP**	**MYP**	**TIMSS**	**MEB**
CONTENT	Quantity	Number	Number	Number	Numbers
	Space& Shape	Shape& Space	Geometry& Trigonometry	Geometric Shapes& Measures	Geometry
	Change& Relationships	Pattern& Functions	Algebra		(Algebra)
	Uncertainty	Data Handling	Statistics& Probability	Data Display	Probability & Statistics, Data*
		Measurement	Discrete Math		Measurement

*For Grades 1-5 Data, for Grades 6-8 Probability & Statistics are used
Algebra starts from Grade 6

Table 2.2 Comparison of Different International Mathematics Programs in terms of Thinking Skills

	PISA	PYP	MYP	TIMSS	MEB**
SKILLS & COMPETENCIES	Thinking& reasoning	Constructing Meaning	Knowledge& Understanding	Knowing	Problem Solving
	Argumentation	Understanding & Applying	Application& Reasoning	Applying	Communication
	Communication	Transferring Meaning into Signs	Communication	Reasoning	Reasoning
	Modelling		Reflection& Evaluation		Relationships
	Problem Posing & Solving				
	Representation				
	Using Language*& Operations				
	Use of Aids& Tools				

*Language includes symbolic, formal, and technical language

** In the National Ministry of Education program, these skills were not presented explicitly, hierarchically, or related with the content areas (as mentioned previously).

2.2 Standard Setting

The studies focused on the researches aiming to set the cut scores defining the "minimally competent examinee" in the context of an educational setting. This meant defining the passing and failing examinees with respect to those cut scores and the descriptions of their performance levels. The further studies extended the issue from defining two performance levels (i.e pass and fail) to more performance levels; for example, basic, proficient, and advanced (Cizek & Bunch, 2007) or beginning, intermediate, advanced, and exiting (Skorupski & Hambleton, (2005).

The first studies on standard setting in educational contents and purposes depend on the judges' decisions. The one conducted by Angoff in 1971 was not only the most important and mostly referenced study in the related literature but also was the method that the majority of the other methods were emerged from (Cizek & Bunch, 2007; Skorupski & Hambleton, 2005).

One example for the IRT based standard setting studies was application of a one parameter IRT model for the Dutch National Assessment Program in Education (Van der Schoot, 2002). The results of an eight grade mathematics performance survey were used for setting achievement levels of the students' performance in different topics of mathematics curriculum. These achievement levels would then be used for the evaluation of the effectiveness of the educational system. Therefore, as Van der Schoot mentioned, the mostly criticized Angoff method was not preferred because of its reliance on the judges' inconsistent estimates. The researcher claimed that the IRT based model in the study prevented those inconsistent judgments with its presentation of the ability scale and the relative difficulty of items to the judges.

The basis of the method in that study was the "P50-P80 segments". Van der Schoot defined two ability points on the scale: P50 is the point where the probability of answering an item correctly is 50% and P80 is the point where the probability of answering an item correctly is 80%. The line that connected these two points was called as P50-P80 segment and was shown directly on the ability scale for each item. The expert judges of teachers, school counselors, and teacher educators were asked to identify the achievement levels for minimum, satisfactory, and advanced standards by using the tables showing the P50-P80 segments on the ability scale. As indicated before, these tables helped the judges to easily identify the relative difficulty of the items and the ability distribution of the examinees on the same scale. The judges used these tables in the last round of the study after identifying the achievement levels with only the help of items and item contents in the previous rounds.

The researcher discussed an interesting point of the importance and benefit of the IRT based standard setting method used. With this method, the participants were able to both consider the intended targets of the curriculum to be attained and also the percentage of passing students after the definition of the cut score for accepted achievement level.

One of the studies comparing the results of two different methods for item mapping in a large scale assessment for primary students' mathematics performance was conducted by Berberoğlu, Demirtaşlı, İş Güzel, and Konak (2008). The study aimed to categorize the items by the proficiency levels using two different standard setting methods and to investigate the congruence between the results obtained by implementing these two methods. Since one of the methods was a judgmental method with items writers and teachers as the panellists, the study also aimed to mention the importance of the understanding of performance levels by those judges. The instrument used in the study was the computerized mathematics test for third graders, which was a part of the Turkish Pupil Monitoring System which was an assessment procedure developed and conducted by CİTO-Turkey.

In that study, one of the methods used was an IRT based item mapping method, which arrayed the items with respect to different response values of 50%, 67%, and 80%. A two parameter IRT model was used with item difficulty and item discrimination parameters. The method was explained in detail, later in Berberoğlu (2009). The procedure can be summarized as:

- Defining the correct response values with the probability of 50% and 80% for each item,
- Experts' investigation of the jump points on the distribution of those probability values and the corresponding items' content differences
- Defining the cut points for four different competency levels as Basic, Competent, Advanced, and Distinguished and identifying the performance level descriptions for each level.

The second method in the study was based on experts' judgments about the classification of the items to different competency levels same as the first method. There were 12 judges participated in the study who were experienced in item writing, mathematics curriculum, corresponding mathematical cognitive skills, and item analysis techniques.

The judges were asked to identify an item as "1" that could be answered correctly with at least 50% probability by the group of students with "Basic" level of competency. They were asked to identify an item as "2" that could be answered correctly with at least 50% probability by the group of students with "Competent" level of competency. The other two competency levels "Advanced" and "Distinguished" were also classified as "3" and "4", respectively.

A significant and high correlation was found between the judgmental and IRT based arraying of items with 0.499 Kendall's Tau and 0.574 Spearman' rho. Only 6 of the 48 items were matched differently by the two methods. These six items, which required computational skills, were identified in lower levels by the judges.

Another study including a comparison between two different standard setting methods was the one, which proposed an alternative method for Angoff method of standard setting (Impara & Plake, 1997). In the study, two groups of teachers assemble to identify the cut scores for passing students on one grade 2 and one grade 5 mathematics tests. Two groups used two different standard setting methods: traditional Angoff method based on identifying proportion correct for each item and the proposed yes/no method based on the participants' judges whether an item could be correctly answered by a "borderline" student. The "yes" answers were transformed into values of 1 and "no" answers were transformed into values of 0. The total of those values were taken as the cut score estimates for each judge.

The final cut score was calculated for both the Angoff method and the proposed yes/no method by averaging the estimates of each participant and corresponding cut scores were found at the end of first round. The judges were given the group's cut score and the percentage of students who would not attain that score. They were given time to discuss their results and the feedback and then they made their final estimates with the same method in the Round 1.

Impara and Plake indicated that the judges using the yes/no method did make fewer changes in their estimates from Round 1 to Round 2 when compared with the judges using the Angoff standard setting method: "It is also notable that the group using the traditional Angoff method shifted by more than six score points between Round 1 and Round 2 (after seeing the actual performance and impact data), whereas the group using the yes/no method shifted by less than one score point between Round 1 and Round 2." (p:357). Also, the range of judges' estimates in traditional Angoff method was much wider than the range in yes/no method. In other words, the variance of estimates of Angoff method in round 2 was 110, whereas, the variance of estimates in yes/no method was 18. These two results showed that the yes/no method had an advantage over traditional Angoff method in terms of the variability in both within judges' estimates in the same round and between judges estimates in two consecutive rounds.

2.2.1 Factors affecting participants' decisions

There are a limited number of studies concerning the factors effecting the decisions of judges who are taking part in the standard setting studies. Ferdous and Plake (2005) investigated the effect of the feedback given to the judges and attempted to identify whether the norm-referenced or criterion-referenced feedback more affected their decisions. The study was conducted with a standard setting study for a Grade 5 mathematics assessment in a Midwestern state in USA. A modified Angoff standard setting method which was explained above (Impara & Plake, 1997) had been used in the study. In the study, 22 panellists participated whose average teaching experience was 18 years.

Ferdous and Plake preferred to separate the participants into three groups with respect to their score estimates for a minimally proficient student in Grade 5 mathematics. Out of a 105 possible points, the participants who estimated the cut score for a minimally proficient student as 82 were defined as "high rating group". On the other hand, "moderate" and "low" rating groups estimated the cut score as 70 and 62, respectively.

The factors affected the participants’ decisions were classified as follows:

Factor 1: Role of performance level descriptors defining the skills and knowledge of a barely proficient student and this factor was taken as a criterion referenced perspective.

Factor 2: Role of students in their class and school, which gives the participants a view for identifying the items as easy or difficult for specific students in their class or school was taken as a norm referenced perspective.

Factor 3: Role of states legislation. The participants could be thinking to set lower standards for minimally proficient student to let more students be classified as proficient because of the fact that all states were evaluated with the percentage of proficient students in every subject.

These factors were tried to be identified by Ferdous and Plake with the following general questions (p:267):

- How did the performance level descriptors for “proficient” affect your decisions about how to classify the test questions?
- How did the performance levels of students in your class and school affect your decisions about how to classify the test questions?
- How did the consequences for students and school for students meeting or not meeting the proficient standards affect your classification of the test question?

The researchers wanted to identify the differences among those groups in terms of the effects norm- and criterion-referenced feedback given to participants. Originally, as the researchers mentioned, criterion referenced information or thoughts should have been more influential on the decisions for item ratings because the minimum knowledge or skills that a “barely” proficient student should have been the evidence for how he or she would perform. However, the research indicated that in the low and high rating groups, the impact data which showed them the percentage of students who would not attain the passing score was more influential.

Dawber and Lewis (2002) elaborated the participants' understanding of the Bookmark method for standard setting, item selection strategies, and the factors effecting their judgements. Both the results of a survey and the protocols of a think aloud procedure were used for identifying above issues in two different standard setting studies for high school exams of mathematics (2 exams) and science (1 exam). There were totally 69 participants who were equally divided into 3 different groups of 23 teachers. Demographic information of race, gender, and years of experience were collected from the participants.

When item selection strategies were taken into consideration, the survey presented the participants three options: (1) identifying an interval of items and selecting an item within the interval, (2) identifying a single item and focusing on that item and the preceded items, (3) identifying a single item and focusing on the skills assessed by that item. The results of the survey showed that the participants mostly used the first and second strategies. The participants who had used the first strategy mentioned that the size of the interval of items decreased as they proceeded in the rounds which indicated less indecision.

When factors affecting the participants' decisions for Bookmark places were investigated, in Round 1, when the participants worked independently without the effect of other participants' opinions, the most important factors were revealed as the participants' "experience in working with students" (82%) and "the difficulty of the items" (91%). "Knowledge of state content standards" (77%) and "Understanding of performance level descriptors" (68%) were the following factors effecting their decisions. On the other hand, in Round 2, after the participants discussed their placements of Bookmarks in their groups, the most affecting factor became "Opinions expressed by small group members" (95%). Although the effect of "experience in working with students" (55%) decreased from Round 1 to Round 2; effect of "the difficulty of the items" (90%) protected its importance.

The think aloud protocols revealed consistent results with the options presented to the participants in the survey study. Codes were taken from the protocols and the themes found in those protocols were the content of the test (content coverage, specific content such as geometry, numbers, etc, number of steps required for solving the problems, item difficulty), personal experience working with students, and small group discussion.

Skorupski and Hambleton (2005) aimed to investigate the thoughts of panellists during a standard setting study by giving them a survey with structured and unstructured items conducted at different moments of the study. The method of standard setting was the yes/no method suggested in Impara and Plake (1997) with slight changes. During the standard setting study for a Grade 5 and Grade 6 ESL assessment, a 61-item questionnaire was given to the panellists and they were asked to answer the items before starting any test section, at the end of training phase, at the end of first round, following the discussions of the first round, and at the completion of the final round. Following are examples of 5-likert scale type of items with levels of strongly agree, agree somewhat, undecided, disagree somewhat, and strongly disagree (p:255):

1. I am very confident in my understanding of the standard setting task described to me.

2. At this time, I completely understand the differences among the four Performance Levels.

Besides the structured items, there were several constructed response items which required participants to freely write their opinions. Following are examples of those unstructured items in the questionnaire (p:254):

1. Why do you think you were asked to participate in this two day meeting?

3. Do you have any idea about the method you will be using to set standards on the test? If yes, briefly describe your present understanding.

The study revealed important conclusions: (1) At the beginning of the study, the panelists had had very different ideas about the performance level descriptors, which indicated a need for orientation and initial training. (2) The panelists did not feel totally comfortable with their decisions due to their statements of lack of confidence in their ratings. The researchers suggested follow up studies after the participants' final ratings. (3) The type of the item whether it was multiple choice or constructed response affected the participants' confidence about the procedure. It was mentioned that: "…performance tasks with polytomous scoring create special challenges for panellists" (p:233). (4) The panellists mentioned that they had felt rushed during the study. The researchers suggested to use less time consuming tasks for the standard setting tasks or to give the participants more time than a two-day meeting.

Apart from the studies fundamentally aiming to investigate the factors affecting the panellists' decisions, few studies stated comments from the results indicating possible reasons for item mappings. Berberoğlu, Demirtaşlı, İş Güzel, and Konak (2008) presented graphically that there was a relation between the distribution of the items to the performance levels and item type. The judges placed the most of the open ended and multiple choice items to "competent" performance level, whereas hot spot items were mostly placed in "Basic" level. Also, when the cognitive skills required for the items were examined, it could be seen that problem solving skill was thought as a characteristic for "Advanced" and "Distinguished" levels, however, computational skill and computational understanding were indicators of "Competent" level.

CHAPTER 3

METHODOLOGY

In this chapter, the sample, instrument and the procedures used in the study are described.

3.1 Research Design

The purpose of the study was to define the skills characterizing the different performance levels of 4th grade students in mathematics using two different methods of standard setting: (1) an IRT-based method called the Bookmark Method (Cizek, 2007), and (2) a judgmental method (Berberoğlu, 2009). The cognitive characteristics exemplifying the performance levels were defined and the congruence and discrepancies between two used methods were investigated.

The first step in the study was to prepare the measuring instrument that would be used for the purposes of the study. The items were prepared by the researcher and one mathematics teacher working at the Measurement & Evaluation Department of a private school.

The test was conducted in four sessions. All of the four parts of the test were one of the different sections of a combined examination of four courses: Turkish, Mathematics, Science & Technology, and Social Sciences. For the purposes of the study, the results of mathematics sections of these four combined test were taken and used.

The ten participant mathematics teachers were selected for being the panellists in standard setting study. The teachers were divided into two groups and assigned to one of the methods: six teachers to IRT-based method and seven teachers to judgmental methods (three teachers took part in both studies). Between the implementation of two studies, four months were left to let the common teachers not affected by the previous study.

3.2 Population and Sample

There are two groups of sample in this study. Since the study aims to identify the performance level descriptors of the 4th grade students, the target population is all 4th grade students in Turkey. The accessible population is the 4th grade students in İstanbul which makes a total of 222937. The sample chosen was a convenient sample for this study. It was difficult to have random sample from the population because the study requires both the conduction of the test and the analysis of the results by the participant teachers. Therefore, eight different private primary schools from seven different districts of İstanbul were chosen as the sample with 269 students. All the 4th grade students at the sample schools were chosen as the participants; however, due to missing participation to any of the consecutive sessions of the conduction of the tests or very low results, three of the students were then eliminated from the study. The number of students, classes, and schools with respect to the districts are presented in the Table 3.1.

Table 3.1 Number of students, classes, and schools with respect to districts

District	Number of Schools	Number of Classes	Number of Students
Kadıköy	2	5	91
Beşiktaş	1	2	24
Üsküdar	1	2	30
Bakırköy	1	2	35
Gaziosmanpaşa	1	2	35
Sarıyer	1	2	25
Kartal	1	2	32
Total	8	17	272*

*Three of the students were then eliminated

And also, the study aims to compare the results of two standard setting methods used for identifying the performance levels for 4th grade mathematics students. Ten primary mathematics teachers were selected as the participants. The teachers were assigned to two different methods and three of them attended the sessions for both methods. Two of the teachers have been working with the classroom teachers on teaching mathematics for two years. Two of the teachers have been working as assessment experts in the Measurement & Evaluation Department of a private school. They were experienced in preparing item specifications, constructing items, and analyzing the results. All participants were female. The average of participants' age was 31.5 and the average of their years of experience was 9 as shown in the Table.

Table 3.2 Information about the teachers participated in standard setting study

Number of Participants	Mean Age	Mean Years of experience
10	31.5	9

3.3 Instrument

The main instrument in this study was Grade 4 Common Mathematics Test (CMT) constructed and conducted by the Measurement and Evaluation Department of the sample institution. The test was conducted in the schools at the same time in four consecutive sessions.

The CMT was a mathematics test constructed to monitor the students' achievement on the intended outcomes of the fourth grade mathematics curriculum of National Ministry of Education. The test consisted of 62 multiple choice items with four alternatives. These items were prepared with respect to the learning outcomes which were classified under four main content areas and twenty two sub-areas of Primary Mathematics Grade 4 Curriculum (MEB, 2005).

Table 3.3 Distribution of items to content areas and sub-areas

CONTENT AREA	SUB AREAS	ITEMS
NUMBERS	Natural numbers	16,19,20,21,22,23,27,48,53,
	Addition with natural numbers	25,28,35,38,41,46,62
	Subtraction with natural numbers	14,24,25,38,39,41
	Multiplication with natural numbers	26,39,46
	Division with natural numbers	
	Fractions	32,34,42,44,52
	Addition with fractions	52
	Subtraction with fractions	43
	Decimals	40,49,51
GEOMETRY	Angle and Its Measure	1,6,8,10,12,13,57
	Triangles, Squares, and Rectangles	3, 7
	Solids	
	Symmetry	2, 4, 5,15,18,60,
	Patterns and tesselations	9,37,61
MEASUREMENT	Measuring length	55
	Perimeter	58,59
	Area	
	Measuring time	52
	Weighing	29,56,
	Measuring liquid	30,47,54
DATA	Column Graphs	11, 17,33,36,
	Probability	45,50

Table shows the corresponding content areas and sub-areas of each item in the test. All the sub-areas except division with natural numbers, solids, and area were covered by the items in the test. In the mathematics program, for every grade level, the weight of the each area and each sub-area were defined with respect to the number of learning outcomes that should be covered. These weights in percentages were given as follows (MEB, 2005):

Table 3.4 Ratio of each sub-area in the grade 4 mathematics program of Ministry of Education

CONTENT AREA	SUB AREAS	PERCENTAGE (%)	
NUMBERS	Natural numbers	6	
	Addition with natural numbers	6	
	Subtraction with natural numbers	6	
	Multiplication with natural numbers	9	
	Division with natural numbers	9	
	Fractions	7	
	Addition with fractions	2	
	Subtraction with fractions	3	
	Decimals	7	
	Total		**55**
GEOMETRY	Angle and Its Measure	6	
	Triangles, Squares, and Rectangles	7	
	Solids	2	
	Symmetry	2	
	Patterns and tesselations	2	
	Total		**19**
MEASUREMENT	Measuring length	4	
	Perimeter	4	
	Area	4	
	Measuring time	3	
	Weighing	3	
	Measuring liquid	4	
	Total		**22**
DATA	Column Graphs	2	
	Probability	2	
	Total		**4**
	GRAND TOTAL	**100**	

Berberoğlu (2009) mentioned importance of identifying which outcome had been intended to measure by each item before starting to construct the test. Making the item specification table for a test is indispensible, and the construction of this table should be based on the purpose and context of the test. The items in the CMT were developed with respect to the learning outcomes defined in the mathematics programme of MEB; therefore, those weights shown in the Table were taken into consideration and tried to be kept similar in the total of the test. The Table presents the number and percentage of items in each sub-area.

Table 3.5 Number and percentage of items related to each sub-area

CONTENT AREA	SUB AREAS	NUMBER	PERCENTAGE (%)
NUMBERS	Natural numbers	9	14,5
	Addition with natural numbers	7	11,29
	Subtraction with natural numbers	6	9,68
	Multiplication with natural numbers	3	4,84
	Division with natural numbers	0	0
	Fractions	5	8,06
	Addition with fractions	1	1,61
	Subtraction with fractions	1	1,61
	Decimals	3	4,84
	Total		**56,43**
GEOMETRY	Angle and Its Measure	7	11,29
	Triangles, Squares, and Rectangles	2	3,23
	Solids	0	0
	Symmetry	6	9,68
	Patterns and tesselations	3	4,84
	Total		**29,04**
MEASURMENT	Measuring length	1	1,61
	Perimeter	2	3,23
	Area	0	0
	Measuring time	1	1,61
	Weighing	2	3,23
	Measuring liquid	3	4,84
	Total		**14,52**
DATA	Column Graphs	4	6,45
	Probability	2	3,23
	Total	**4**	**9,68**
	GRAND TOTAL	**109,67**	

The table shows that there are both differences in the percentage of items in the test covering each sub-area and area and the weights of these sub-areas and areas with respect to total number. There are several reasons for this difference: (1) there were items like 25, 38, and 41 that could be related with more than one sub-area and the corresponding learning outcome. As a result of this multiple correspondence, both the total percentage exceeded the 100 % and the balance among the weights of the number of items related with content areas and sub-areas differed from the original percentages of those areas in the mathematics program. (2) Since the "content-based classification" of the items and developing a test with respect to this classification led to weaknesses in terms of the skills required for identifying the performance levels (). Therefore, the items had been developed by taking the skills required independent from the content dimension which led to discrepancies in the weight of learning outcomes.

3.4 Analysis of Data

In this study, the descriptive analyses of the results of the Common Mathematics Tests (CMT) including mean, standard deviation, minimum and maximum scores, skewness and kurtosis of the distribution were conducted by SPSS 11.01 (SPSS Inc, 2001).

The reliability of the test scores was detected by using the Cronbach alpha (or coefficient alpha). The reliability analysis was conducted by again SPSS 11.1.

One of the standard setting methods stood on the Item Response Theory (IRT). To detect the items with one-parameter model, the software BILOG-MG (SSI, 2003) was used.

The results of the two standard setting methods revealed two different item mappings, in other words, two different classification of items into the performance levels. The congruence between these two item mappings was analysed with coefficients of Kendall's Tau and Spearman' s rho (Green, Salkind & Akey, 2000).

After the item mappings were analysed, the performance descriptors for each level of performance, such as Below Basic, Basic, Proficient, and Advanced were identified. The common psychometric characteristics of the items that were classified together in the same performance level were examined and the statements for descriptors were decided (OECD, 2003; Martin, Mullis & Foy, 2008).

The last analysis and examination was the evidence for the "construct validity" for the description of performances for each level. The descriptions were identified by the common characteristics of what the items in those categories had intended to measure. However, an empirical check was needed to be able to show the validity of those descriptions. The path analysis model of Structural Equation Modeling was used for this purpose (Jöreskog & Sörbom, 2001). The skills that were grouped by the items into hierarchical levels such as Basic, Proficient, etc. entered the path analysis to demonstrate that the skills grouped in the lower levels of performance were prerequisites for the skills grouped in the higher levels of performance. The software LISREL 8.54 was used (SSI Inc, 2003). The variables used were presented below:

IDENTNUM: Identifying and modelling numbers

RECOGSHP: Recognizing angles and shapes

SYMMETRY: Finding and using symmetry context

PATTERNS: Finding and defining patterns

ONESTEP: Conducting one step operations

MULTI: Conducting multi step operations

ROUTINE: Solving routine problems

NONROUTN: Solving non routine (complex) problems

In the following sections, the methods for standard setting were explained and sampled in detail.

3.4.1 Identifying Item Difficulty Indices

The difficulty indications of the items were estimated by using One Parameter Logistic IRT Model or with its more popular name Rasch Model (Hambleton, Swaminathan & Rogers, 1991).

This model stands on the logistic function

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n \qquad (1)$$

set forth by Hambleton, Swaminathan and Rogers (1991). In this equation, $P_i(\theta)$ is the probability of answering item *i* correctly by a chosen examinee, b_i is the difficulty parameter of that item *i*, θ is the ability of that chosen examinee, e is the transcendental number with value 2.718, and n is the number of items in the test. As the basic function defining the Rasch model, equation (1) gives the opportunity to conclude the following results:

1. Equation (1) is not defining a single function, because, by using it, we can calculate the probability of a correct response for every item i by the same examinee with the ability θ. Therefore, equation (1) is the group or family of functions (Verhelst, 2004).
2. When equation (1) is examined, it can be seen that one examinee's probability of answering an item correctly depends on only a single item characteristic, i.e. the item difficulty (Hambleton, Swaminathan & Rogers, 1991). This is the reason why this model is also named as one parameter logistic model.
3. This equation proposes a model where the probability of answering item *i* correctly is 50% (or 0.5 or 1/2) only when the ability *θ* of the chosen examinee equals the difficulty parameter b_i of item *i*. Additionally, one can make the mathematical conclusion that the greater ability *θ* is required for an examinee to have a probability of 50% to response the item correctly, if the item has a greater value of b_i parameter. Vice versa, the smaller the value of b_i, the smaller the ability is required for an examinee to have a probability of 50% to

response the item correctly. Hence, the function is an increasing function.

4. The probabilistic idea of one examinee’s chance to answer one item correctly can be shaped by equation (1) and the above conclusions in a theoretical graph in the Figure 3.1. The graph is traditionally called “Item Characteristic Curve” (Hambleton, Swaminathan & Rogers, 1991; Verhelst, 2004).

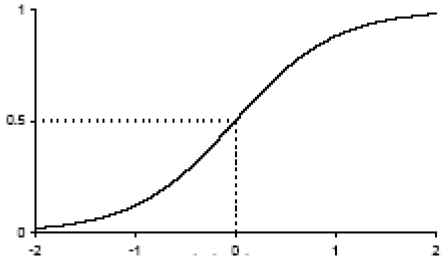


Figure 3.1 Item Characteristic Curve

3.4.2 Finding Ability Scores Corresponding to Item Difficulty Parameters

The item difficulty indices were calculated with BILOG-MG and were listed in a MS-Excel Sheet. This sheet would then be used to convert the difficulty parameter of an item into the ability score which was required for answering that item correctly with a defined probability. As explained before in the Chapter 2, this probability value was chosen, under the light of previous studies, as 67 %.

The abilities required to correctly answer each item with a probability of 67% could be calculated with the help of equation (1):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1,2,3,\ldots, n \qquad (1)$$

As indicated before, $P_i(\theta)$ is the probability of an examinee with an ability θ to answer the item i correctly. Since this probability was taken as 67%, the equation became:

$$\frac{67}{100} = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad (2)$$

Now, it can be continued to solve for θ:

$$67 + 67e^{(\theta - b_i)} = 100 \cdot e^{(\theta - b_i)} \qquad (3)$$

$$\frac{67}{33} = e^{(\theta - b_i)} \qquad (4)$$

$$2.0303 = e^{(\theta - b_i)} \qquad (5)$$

$$\ln 2.0303 = \ln e^{(\theta - b_i)} \qquad (6)$$

$$\theta\text{-}b_i = .708 \qquad (7)$$

$$\theta = b_i + .708 \qquad (8)$$

The final equation (8) states that any examinee who is supposed to have the 67% chance to answer the item i correctly will need to have the ability score which is .708 more than the difficulty of the item i. Since the items in the test were calibrated with Rasch model of IRT, the ability levels of the examinees and the item difficulties were brought to the same scale, therefore, there was no harm to find the θ values with the linear equation (8). The derivation of the equation (8) let to finding the abilities required for answering each 62 items in the test.

As mentioned above, an MS-Excel sheet was used to list the difficulty parameters of 62 items. MS-Excel gives the opportunity to write and solve mathematical equations for the given variables; therefore, the equation (8) was solved for b_i to find out the θ values with the function property of MS-Excel.

3.4.3 Developing the Ordered Item Booklet

In two different standard setting sessions, the participants were given the items in the booklets called "Ordered Item Booklet". The common characteristic of the booklets used in two methods was the presentation of the items in an ascending order of difficulty. However, other information on the pages of the booklets was differentiated with respect to the purposes of their usage. These points are explained in the following sections.

3.4.3.1 OIB in Bookmark Method

The items were presented to the participant teachers with the method similar to the one that was suggested by Cizek and Bunch (2007) in the Bookmark Method. The method was named as the Ordered Item Booklet (OIB) in which the items were ordered with respect to difficulty from the easiest one to the most difficult one. A similar procedure was followed in the current study and the ordered items were given to the participants as one item on one page. This single page contained the information of the item difficulty, the order of the item on the booklet and the item itself.

As mentioned above, this collection of the items was called OIB, because the items were ordered with respect to their difficulty. All the items in the test used for the study was detected with Item Response Theory (IRT) analyses to define their item difficulties and these difficulties were used to construct the OIB. The original place of the item in the test was then replaced with its new place with respect to its difficulty.

As it was indicated before, the item difficulty parameters were needed to develop the OIB in the Bookmark Method for standard setting, therefore, the items in the test were calibrated with respect to the one parameter IRT model and item difficulty parameters (b_i) were detected by using the software BILOG-MG Version 3.0 (Scientific Software International, 2003).

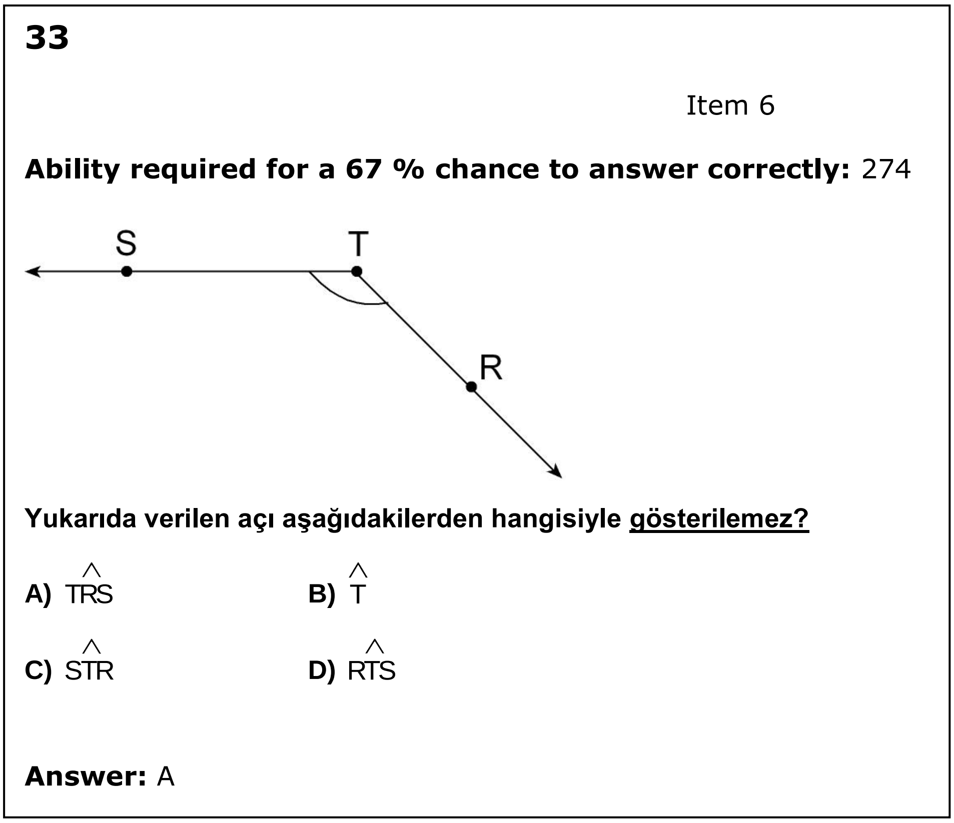
The OIB could be prepared with the determination of the order of each item, item difficulty indices, and the ability levels required for each item to be answered correctly with 67% probability. The ability levels would be used as the main information by the teachers to determine the places of bookmarks put for the standard setting purposes. Therefore, the ability levels were transformed to a new scale that could be perceived more easily by the judges. The new scale was the distribution of ability scores of the all students which had a mean value of 250 and a standard deviation of 50. The new values for the ability levels were calculated by equation (9):

$$\text{Scaled Ability} = \frac{\text{Ability}\theta - \text{Mean}\theta}{\text{StandardDeviation}} \cdot 50 + 250 \qquad (9)$$

Each θ value was transformed to a new value by using the equation (9) and these values were used as the indicator of the item difficulty. A sample page from the OIB used in the Bookmark Method was shown in Figure 3.2.

3.4.3.2 OIB in Judgmental Method

The participants of judgmental method were presented the OIB like the group of panelists of IRT-based method, but the OIB's of this group differed from the other ones in terms of the information in them. Every page of this OIB includes the order of the item, its original place in the test, correct answer, the content area and the sub-area that the item was belong to, and the learning outcome that was intended to be measured by the item. A sample page can be found in Figure 3.3.



33

Item 6

Ability required for a 67 % chance to answer correctly: 274

Yukarıda verilen açı aşağıdakilerden hangisiyle gösterilemez?

A) $\widehat{TRS}$ **B)** $\widehat{T}$

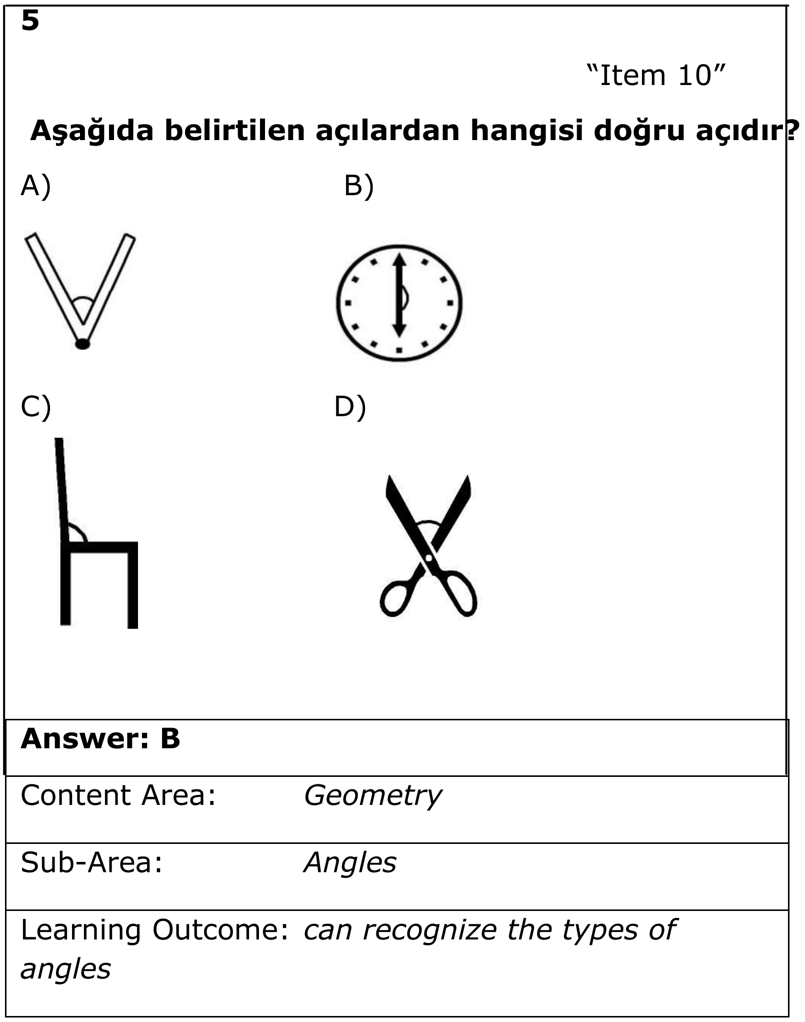
C) $\widehat{STR}$ **D)** $\widehat{RTS}$

Answer: A

Figure 3.2 Sample Page from Ordered Item Booklet (OIB)

3.4.4 Finding the Corresponding Raw Scores

One part of the information given to the participants using IRT-based method for standard setting was showing them the corresponding raw scores that would be allocated to cut score points they had set for performance levels. In other words, the participants examine the items starting from the easiest one to the most difficult one; put their bookmarks with respect to their decisions about the cut points for performance levels, according to the difficulty indices of these "border" items, the required ability scores are found; and finally that ability score and the corresponding raw score are matched.



5

"Item 10"

Aşağıda belirtilen açılardan hangisi doğru açıdır?

A) B)

C) D)

Answer: B

Content Area:	*Geometry*
Sub-Area:	*Angles*
Learning Outcome:	*can recognize the types of angles*

Figure 3.3 Sample page form OIB used for judgmental method

The match between the ability scores and the raw scores can be obtained from the Phase-3 output of BILOG-MG. For every examinee, the output gives the raw score taken by the examinee and also the calculated ability score. A part from the Phase-3 output is shown in Figure 3.4. The BILOG output supplies the user the raw score of each subject and his/her allocated ability. These important data were used to transform the ability required for the item that any bookmark had been placed to the closest raw score. For example, suppose a participant placed the bookmark for Basic performance level on page 9. This meant that any person, who could be counted as having the Basic performance level, should have at least had the ability of -.025 (Table 3.4). The BILOG output mentioned above could then be used to transform this ability θ of -.025 to the closest raw score which was 31. In other words, a student who could answer the 31 items of 62 correctly would be grouped as Basic performance level holders.

Therefore, for finding the raw score corresponding to an item, the required ability value for answering that item with the probability of 67 % is matched with the nearest ability score in that output. The raw score taken from that examinee will then naturally be the raw score corresponding to that item.

```
GROUP   SUBJECT IDENTIFICATION                              MARGINAL
 WEIGHT   TEST     TRIED  RIGHT  PERCENT    ABILITY     S.E.    PROB
 ----------------------------------------------------------------------
   1  10028966462                     |                |
   1.00   MATHTEST    62    20    32.26 |   -0.4445    0.0094 |  0.000000
   1  10070014272                     |                |
   1.00   MATHTEST    62    33    53.23 |    0.3991    0.1956 |  0.000000
   1  10136933264                     |                |
   1.00   MATHTEST    62    55    88.71 |    1.3404    0.0790 |  0.000000
   1  10178524468                     |                |
   1.00   MATHTEST    62    33    53.23 |    0.3991    0.1956 |  0.000000
   1  10516666584                     |                |
   1.00   MATHTEST    62    33    53.23 |    0.3991    0.1956 |  0.000000
```

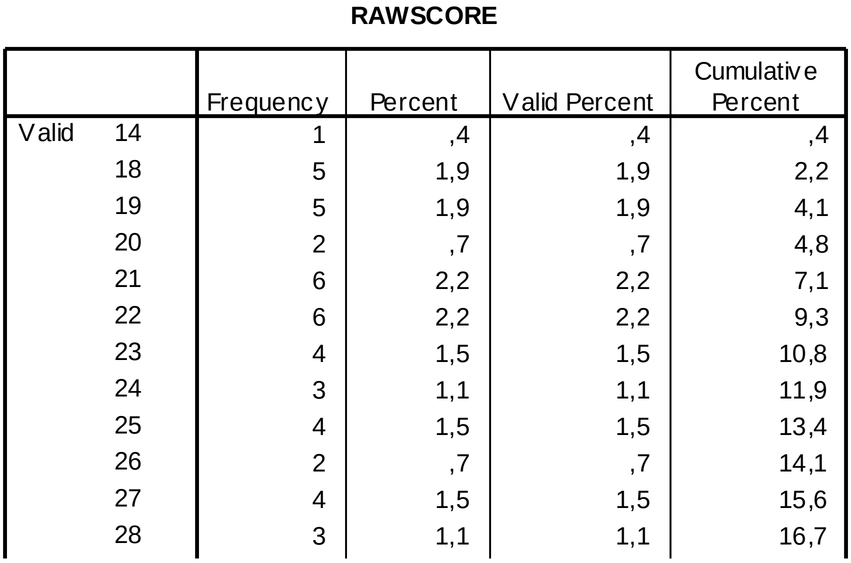
Figure 3.4 Extract from a sample BILOG-MG output

For example, the item 40 in the original test is on the 20th order of OIB with a required ability score of .377. The nearest value in the output, part of which can be seen in the Figure, is .399. This value is the ability theta for a student who scored 33 in the test. Therefore, the corresponding raw score for the item 40 is 33. For all the items, the corresponding raw scores were found with this procedure. For easily finding, the output was ordered with respect the ability value, and therefore, ability theta value required for each item could be found easily and consistently, by comparing this theta value for the nearest value in the list.

3.4.5 Finding Percentages

Cizek and Bunch (2007) recommends that before the last round of the Bookmark Method, the participants can be presented more impact data which demonstrates the percentage of the students at or above the raw score allocated for each item.

For finding these percentages of students at or above each raw score, SPSS 11.1 software was used. The descriptive analyses including frequency, percent, and cumulative percent for every score value were analysed and calculated. The Figure demonstrates a part of the output from the SPSS Frequency analyses. In the table, raw score, frequency of that raw score in the sample, and the cumulative percents can be seen from the output.



RAWSCORE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	14	1	,4	,4	,4
	18	5	1,9	1,9	2,2
	19	5	1,9	1,9	4,1
	20	2	,7	,7	4,8
	21	6	2,2	2,2	7,1
	22	6	2,2	2,2	9,3
	23	4	1,5	1,5	10,8
	24	3	1,1	1,1	11,9
	25	4	1,5	1,5	13,4
	26	2	,7	,7	14,1
	27	4	1,5	1,5	15,6
	28	3	1,1	1,1	16,7

Figure 3.5 Extract from a sample SPSS output

To be able use this data as the feedback information for the participants, the cumulative percent for each raw score value should be transformed to the percentage of the students at or above that raw score. The cumulative percent for a score shows the percentage of the students at that score plus the percentage of the students whose score are lower than that score. Therefore, the percentage of the examinees above that score can be found by subtracting the cumulative percent given in the table from 100%. Since the percentage of the students at that score is also expected, then the percent of the students who took that score should be added to result of the above described subtraction. To avoid adding a previously subtracted value, the cumulative percentage matched with the previous raw score can be subtracted from 100%. For example, the cumulative percent for the score 28 is given as 16.7 % in the table. To find the percentage of students at or above that raw score, the previous cumulative percentage value of 15.6 is read from the output and it is subtracted from 100%. The remainder gives the percentage of the students who scored 28 and more in the test. A part of the table showing that conversion is shown in the Figure.

order of difficulty	itemno	RP %67	raw score	Percentage at or above
1	item1	-0,705	14	100
2	item27	-0,557	14	100
3	item17	-0,432	28	84.4
4	item2	-0,404	29	83.3
5	item10	-0,299	30	81.4
6	item33	-0,275	30	81.4
7	item4	-0,142	31	78.8
8	item11	-0,101	31	78.8
9	item3	-0,025	31	78.8

Figure 3.6 Extract from the table showing the abilities required for each item, the corresponding raw scores, and percentages of students at or above each raw score

3.5 Standard Setting Methods

3.5.1 Bookmark Method

The first method used in this study for setting the mathematics performance levels of fourth grade students is an IRT based procedure called Bookmark Method (Cizek & Bunch, 2007). The method basically depends on the participants' marking the cut points in the ordered collection of items which is formed with respect to the parameters taken from IRT analyses.

In the bookmark method, the participants were asked to deduce on every item whether the item was likely to be answered correctly by the student who was supposed to have the minimum qualifications of the given performance level. In other words, the participants indicated the places to put bookmarks named Basic, Proficient, and Advanced on the sequence of items ordered according to their difficulty. These three bookmarks demonstrated the borders for the four performance levels of Below Basic, Basic, Proficient, and Advanced. When a participant put the Basic bookmark in front of an item, this meant that any person who was defined to have Basic performance level should have at least answered this item correctly. And on the contrary, the person would be counted as Below Basic performance level if he or she could not have answered that item correctly. The same idea was valid for placing Proficient and Advanced bookmarks.

3.5.1.1 Round One of Standard Setting Procedure with Bookmark Method

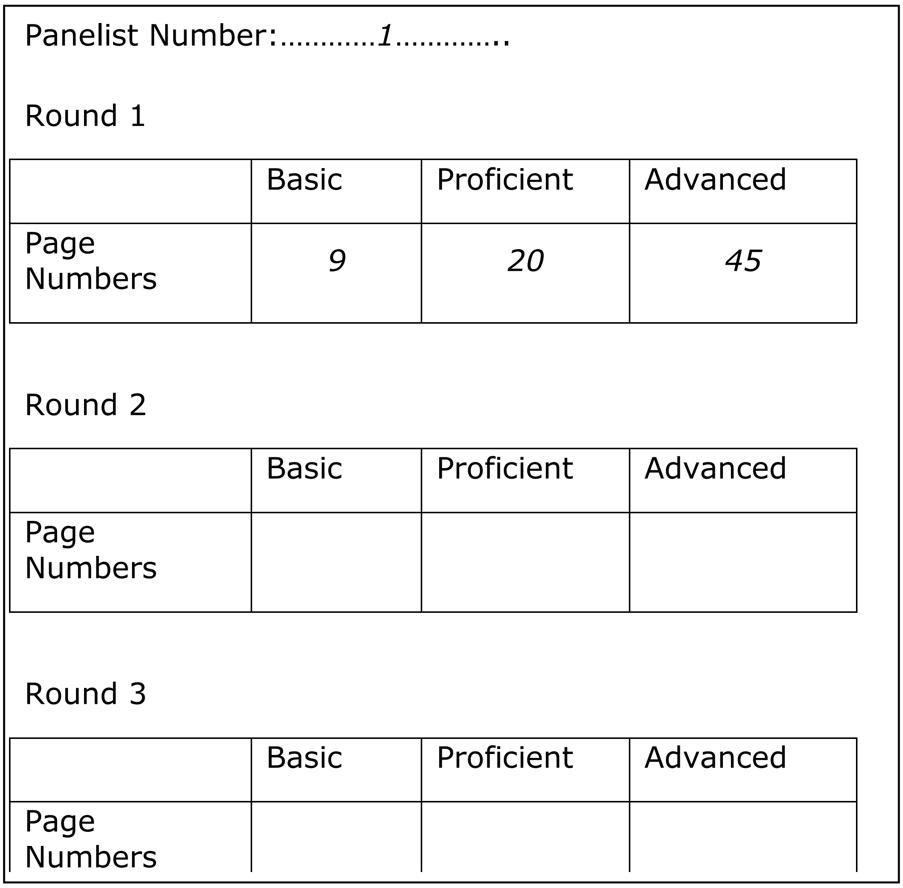
Before starting this first round, the panellists were given a detailed introduction about the procedure. They were introduced by the Ordered Item Booklets (OIB) and the steps of the implementation and the expectations from them were explained with the examples. The full text of the introduction given to the panellists can be found in Appendix.

The OIB contained the items in the increasing order of difficulty with the required ability for answering each item with 67% chance. A sample page taken from OIB used in the implementation was presented in the Figure 3.2. This had been the "item 6" in the originally ordered test, however, the order of the item became 33 after the items were ordered with respect to their difficulty indices. The page contained the information about both the original and the manipulated orders of the item. Another data given to the panelists through this page was the "ability required for a 67% chance to answer correctly". The value "274"of this ability **θ** was taken from the Table 3.4 indicating the values on the transformed scale. The page also included the item itself and the correct answer.

The task for the participants, as explained before, was to place three bookmarks indicating the borders for the performance levels. Since the booklet they were given presented the items in the increasing difficulty order, they started with placing the bookmark for Basic level and continued with Proficient and Advanced. Therefore, the participants were asked to solve the items in OIB one by one and to think of the answers for the following questions for each item(www.sagepub.com/cizek/bookmarktraining):

- What makes this item more difficult than the previous items?
- What skill or knowledge is required to answer this item correctly?
- Think of group of students assumed to have the Basic performance level. Would at least 67% of them solve this item correctly?

After those brief explanations, the participants solved the items, checked their results with the answers of the items and decided where to put the bookmarks for Basic, Proficient, and Advanced performance levels. Their decisions were collected by a similar form like the one given in the Figure 3.7. This form in the Figure 3.3 was used by each participant for all three rounds to let them observe their decisions and the change in those decisions through the rounds. In the sample in Figure 3.7, one of the panellists who was named as the panellist one, decided that any student who could be defined as having the Basic performance level should have at least answered the ninth item in the OIB. By the same way, for being counted as Proficient and Advanced, twentieth item and forty fifth items, respectively, should have been answered correctly.



Panelist Number:…………*1*…………..

Round 1

	Basic	Proficient	Advanced
Page Numbers	*9*	*20*	*45*

Round 2

	Basic	Proficient	Advanced
Page Numbers			

Round 3

	Basic	Proficient	Advanced
Page Numbers			

Figure 3.7 Sample Form Filled by All Participants to Indicate Their Bookmark Decisions at the end of each Round

After all the items were examined and the bookmarks were placed by the participants, the researcher collected the pages containing the Figure 3.3 from all participants. The data from all these papers were entered into a spreadsheet and summary statistics were also calculated and presented. The core aim of the bookmark method was to define the cut scores for Basic, Proficient, and Advanced performance levels. The placement of the bookmarks, however, could only present these cut scores in terms of the order of pages. To define the actual cut scores, those "order of pages" should have been transformed to raw scores indicating the performance levels set by the bookmarks. This transformation could be done with the phase 3 output given by the BILOG (Cizek and Bunch, 2007) as explained before.

The researcher collected the participants' papers including their decisions on placing bookmarks and she constructed the feedback table as explained in the previous paragraph. The table consisted of each participant's decision for placing Basic, Proficient, and Advanced bookmarks. In Cizek and Bunch (2007), it was suggested to present the participants the frequency table for bookmark decisions for each performance level as the feedback from the first round. Since the number of participants was 6, which could be taken as a reasonable amount to present the whole data, the frequency table was not preferred as the feedback information from round 1.

3.5.1.2 Round Two of Standard Setting Procedure with Bookmark Method

After the completion of the first round, the participants came together as a whole group and discussed their work. The researcher directed the discussion around the answers of the questions of how they had described the performance levels such as Basic, Proficient, and Advanced and how their bookmark decisions had reflected those descriptions. The participants mentioned their rationales behind selecting the bookmark points, discussed the differences among their difficulty locations of the items, and the variety of cut scores. These discussions led the participants to think about their decisions and listen to the other panelists' points of views.

At the end of these discussions and the researcher's preparation of the feedback information from the first round of standard setting procedure, the panelists were again given the ordered item booklets to start the second round. They were going to use their OIBs to place the bookmarks with respect to their decisions for performance levels. In addition, they were given the feedback table presenting all panelists' bookmark decisions for basic, proficient, and advanced performance levels and, on the other hand, corresponding raw scores for every bookmark place.

The information given in the Table 3.5 should have led the participants to observe the appropriateness of their decisions about the difficulty of the individual items to the performances in terms of raw scores. For example, a participant who put the Basic bookmark to page 9 would then realize that the ability required for answering this item correctly with 67% chance equaled to answering 31 of 62 items correctly. The impact data presented to the participants in the second round of standard setting procedure helped them to have a more realistic idea about the ability distribution of the examinees.

The participants continued to examine the OIB to make their judgments for the bookmark placements and filled up the second round of the table shown in Figure 3.3. After all the participants finished filling up their forms, they were collected by the researcher and she prepared the feedback tables containing the information of each participant's judgments for each cut score, corresponding θ values, and the raw scores.

3.5.1.3 Round Three of Standard Setting Procedure with Bookmark Method

At the beginning of the last round, the panelists were given two different tables giving them information. The first one was presenting the all participants' judgments about the bookmarks; corresponding theta cut scores and raw cut scores. In addition to this, they were given the information of theta value required to solve each item correctly with 67% chance, the corresponding raw score values for that response value, the percentage of examinees at or above each recommended raw cut score.

This impact data presented the important relationship between the theta cut score and the raw cut score (Cizek and Bunch, 2007). By introducing the participants this information, they did then have the idea of how their placements of bookmarks had affected raw cut scores for defining performance levels. Besides, the judges were then aware of the ability distribution of the group of examinees by the data of percentage at or above the corresponding raw score.

At the beginning of the Round 3, the participants took all these information, discussed the impact data and worked on the items once more as a whole group. Then, they individually made their last changes on the placement of the bookmarks and submitted their latest judgments to the researcher. The researcher collected the judgments of the participants and constructed the table for the final bookmark decisions and calculated the mean raw scores for each performance level.

3.5.2 Judgmental Method

The second group of participants were assigned to set the cut points for the performance levels without given any information about the examinees' relative performance on the test like the feedback tables presented to the other group at the end of the Rounds 1 and 2. However, they were given more information about the knowledge and skills that every item required and more time to work on the charts of outcomes of the grade 4 mathematics program. This method was more focused on the thinking procedures of examinees when they had been answering the items.

3.5.2.1 Round One

The participants of judgmental method were presented the OIB like the group of panelists of IRT-based method, but the OIB's of this group differed from the other ones in terms of the information in them. Every page of this OIB includes the order of the item, its original place in the test, correct answer, the content area and the sub-area that the item was belong to, and the learning outcome that was intended to be measured by the item. A sample page can be found in Figure.

Before starting to work on the OIB's and to place the points for the levels of performance, the participants first studied the chart of the learning outcomes organized with respect to content areas and sub-areas of the grade 4 mathematics program developed by Turkish Ministry of Education (MEB, 2005). The translated version of this chart can be found in APPENDIX. They studied the learning outcomes and discussed them as a whole group. By the instructions of the researcher, they discussed and made estimations about which outcomes should have been reached by the group of students having different level of performance levels. They brainstormed on classifying the learning outcomes in a different way, which had been organized by the Ministry under the content areas. They were asked to try to classify them with respect to the skills required for reaching the learning outcomes without thinking of their contents. In Skorupski and Hambleton (2005), these preliminary estimations were analyzed and assessed with respect to these

After these preliminary discussions, the participants were instructed to classify the items into performance levels as follows:

When examining the items in order of difficulty, the participants ask the following question by themselves:

Can an examinee who has the Below Basic level of performance answer this item correctly with a probability of 67%?

If the answer for this question is “yes”, the participant classifies the item as Below Basic. If the participant thinks that the item is not easy enough to be answered correctly by the Below Basic level of performance, then the next question is:

Can an examinee who has the Basic level of performance answer this item correctly with a probability of 67%?

Again, if the answer is “yes”, the process ends and the item is classified as Basic; but the process continues to identify the Proficient and Advanced performance levels.

3.5.2.2 Round Two

The participants went over the items one by one with the whole group, compared their placement of the items into the performance levels, and discussed their different placements of items. The participants were also instructed to come to an agreement by discussing the items, their level of difficulty, and the knowledge and skills required for solving them. Again, they were not forced to come to a full agreement, but were asked to elaborate the item specifications.

The participants were instructed to use the following coding for demonstrating their placements for the items: 1 for Below Basic, 2 for Basic, 3 for Proficient, and 4 for Advanced. After they worked on the items and classified the items to performance levels by using the above coding, then the median value of judges' rankings for each item could be calculated. The table showing this median value was presented the participants and they were asked to come to a final agreement in the last round.

3.5.2.3 Round Three

The table explained above was shared with the participants to make their last decisions and to clarify the final cut points for the performance levels. They had to make changes on only the items which had been classified as a representative for a performance level which was lower than the performance level of the previous item. Their final classifications were collected by the researcher and these categorizations were taken as the identification of the performance levels.

3.6 Identification of Performance Level Descriptors

After the termination of standard setting sessions with both of the methods, the results were analyzed by the researcher for comparing the classifications developed by the two methods.

CHAPTER 4

RESULTS

4.1 Descriptive Summary

In this section, the scale results of the test used in the study will be presented. The Common Mathematics Test (CMT) was a multiple choice achievement test and conducted with 269 grade four students from the 8 different private primary schools. The descriptive statistical results of the CMT can be found in the Table 4.1.

Table 4.1 Descriptive statistics for CMT

STATISTICS	COMMON MATHEMATICS TEST
Number of Items	62
Number of Examinees	269
Mean	38.84
Standard Deviation	10.65
Minimum	14
Maximum	60
Skewness	-.12
Kurtosis	-.79
Alpha	.907
Mean Percent Correct	.63
Mean Biserial	.477

The results indicated a high reliability with Cronbach Alpha value of .907. Although the mean was a moderately high value, the wide range from 14 to 60 indicated the differences among the examinees. The item discriminations showed reasonable values. In the preliminary analyses, two items showed very low percent correct values (.12 and .10) and those items did not fit the IRT model, therefore, they were not included in the descriptive analyses and standard setting studies.

Before starting the IRT calibration of the items, the assumption of unidimensionality of the test was checked. The scree plot diagram in SPSS factor analysis was used to show that there existed only one dimension in the test. With the jump from the first eigenvalue of 9.841 to the second eigenvalue of 2.247, the results indicated one dimension of the test. The scree plot is presented in the Figure.

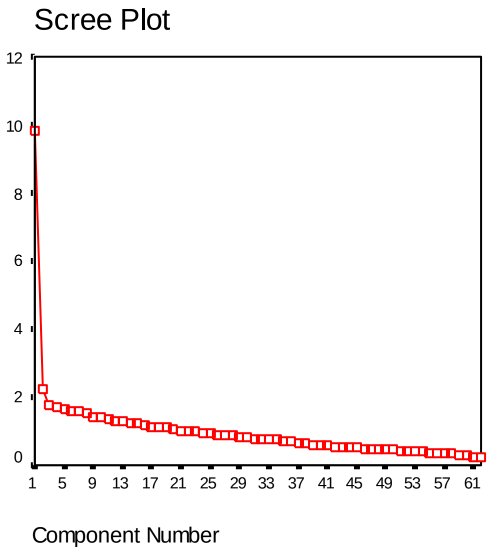


Figure 4.1 Scree plot for CMT

4.2 Results of IRT Based Item Mapping

As mentioned before, two different methods for standard setting based on a Grade 4 mathematics test were used in this study, to compare the item mappings. The first method was an IRT based item mapping which detected the items with the One Parameter Model calibrating the items with respect to the item difficulty indices and allowing the differences in discriminating power between items (Van der Schoot, 2002). The items were calibrating by the software BILOG-MG Version 3.0 (Scientific Software International, 2003) using one-parameter model. The code for the process was presented in the APPENDIX. The mean values of estimated item difficulty parameters and standard errors were given in Table.

Table 4.2 Item parameters summary

	Mean Item Difficulty	Mean Standard Error
IRT- one parameter model	-1.263	0

As can be easily seen from Table 3.1, the first item in the original test was, by chance, the easiest item and therefore, it was put in the first place in the OIB. However, item 27 was the second easiest item in the test and it was seen in the second place in OIB. The most difficult item in the test was the item 59 and, therefore, took its place as the sixty-second item. The difficulty parameters of the items by using One Parameter Logistic IRT Model (Hambleton, Swaminathan & Rogers, 1991) and by using the software BILOG-MG Version 3.0 (Scientific Software International, 2003).

Table 4.3 Orders of the items in the original test, the corresponding orders according to their difficulties, and difficulty parameters

order of difficulty	**item no**	**difficulty parameter**	**order of difficulty**	**item no**	**difficulty parameter**
1	item1	-1.413	32	item41	0.030
2	item27	-1.265	33	item6	0.042
3	item17	-1.140	34	item42	0.086
4	item2	-1.112	35	item12	0.130
5	item10	-1.007	36	item14	0.151
6	item33	-0.983	37	item39	0.151
7	item4	-0.850	38	item46	0.152
8	item11	-0.809	39	item62	0.152

Table 4.3 cont'd

9	item3	-0.733	40	item53	0.163
10	item20	-0.645	41	item50	0.216
11	item43	-0.579	42	item34	0.238
12	item35	-0.563	43	item45	0.259
13	item5	-0.532	44	item24	0.301
14	item25	-0.516	45	item51	0.322
15	item32	-0.486	46	item60	0.396
16	item57	-0.486	47	item8	0.428
17	item18	-0.471	48	item28	0.449
18	item15	-0.414	49	item22	0.470
19	item56	-0.386	50	item13	0.491
20	item37	-0.345	51	item52	0.501
21	item40	-0.331	52	item54	0.672
22	item26	-0.254	53	item31	0.727
23	item21	-0.228	54	item49	0.727
24	item19	-0.191	55	item44	0.828
25	item36	-0.143	56	item58	0.862
26	item7	-0.095	57	item30	0.885
27	item55	-0.095	58	item61	1.118
28	item16	-0.072	59	item9	1.185
29	item48	-0.037	60	item47	1.284
30	item29	-0.003	61	item38	1.313
31	item23	0.019	62	item59	1.438

Item difficulty parameters were converted into ability level of an examinee required for likely answering the item. The abilities required to correctly answer each item with a probability of 67% could be calculated with the help of equation (1):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1,2,3,\ldots, n \qquad (1)$$

As indicated before, $P_i(\theta)$ is the probability of an examinee with an ability θ to answer the item i correctly. When this probability was taken as 67 % and the necessary conversions were conducted in equation (1), the formula to calculate the ability θ required to answer each item correctly became as follows:

$$\theta = b_i + .708 \qquad (2)$$

By substituting the b_i with the item difficulty parameters found by calibrating the items with BILOG-MG, the θ values for each item were calculated. The Table presented the items, their difficulty parameters, and the corresponding required ability θ values. The table was organized in an ascending order with respect to the item difficulty parameters. For standard setting purposes, the item difficulty ordered presentation of the items were needed, therefore, the items were named in terms of both their location in the original test and their difficulty order.

Table 4.4 Ability Levels Required Answering each Item Correctly by Response Value of 67%

order of difficulty	item no	difficulty	ability θ for RP 67%	order of difficulty	item no	difficulty	ability θ for RP 67%
1	item1	-1.413	-0.705	32	item41	0.030	0.738
2	item27	-1.265	-0.557	33	item6	0.042	0.750
3	item17	-1.140	-0.432	34	item42	0.086	0.794
4	item2	-1.112	-0.404	35	item12	0.130	0.838
5	item10	-1.007	-0.299	36	item14	0.151	0.859
6	item33	-0.983	-0.275	37	item39	0.151	0.859
7	item4	-0.850	-0.142	38	item46	0.152	0.860
8	item11	-0.809	-0.101	39	item62	0.152	0.860
9	item3	-0.733	-0.025	40	item53	0.163	0.871
10	item20	-0.645	0.063	41	item50	0.216	0.924
11	item43	-0.579	0.129	42	item34	0.238	0.946
12	item35	-0.563	0.145	43	item45	0.259	0.967
13	item5	-0.532	0.176	44	item24	0.301	1.009
14	item25	-0.516	0.192	45	item51	0.322	1.030
15	item32	-0.486	0.222	46	item60	0.396	1.104
16	item57	-0.486	0.222	47	item8	0.428	1.136
17	item18	-0.471	0.237	48	item28	0.449	1.157
18	item15	-0.414	0.294	49	item22	0.470	1.178
19	item56	-0.386	0.322	50	item13	0.491	1.199
20	item37	-0.345	0.363	51	item52	0.501	1.209
21	item40	-0.331	0.377	52	item54	0.672	1.380
22	item26	-0.254	0.454	53	item31	0.727	1.435
23	item21	-0.228	0.480	54	item49	0.727	1.435
24	item19	-0.191	0.517	55	item44	0.828	1.536
25	item36	-0.143	0.565	56	item58	0.862	1.570
26	item7	-0.095	0.613	57	item30	0.885	1.593
27	item55	-0.095	0.613	58	item61	1.118	1.826
28	item16	-0.072	0.636	59	item9	1.185	1.893
29	item48	-0.037	0.671	60	item47	1.284	1.992
30	item29	-0.003	0.705	61	item38	1.313	2.021
31	item23	0.019	0.727	62	item59	1.438	2.146

The easiest item in the test had the item difficulty parameter of -1.413 (which had been originally located in the first order in the test), and the hardest item had the item difficulty parameter of 1.438 (which had been originally located in fifty ninth order in the test). When the ability θ values are examined, it can be observed that corresponding θ for the easiest item (Item 1) is also the smallest θ required for answering an item correctly (-0.705), whereas, corresponding θ for the most difficult item (Item 59) is the highest θ required for answering an item correctly (2.146).

So, the panelists of measurement experts and mathematics teachers were given the Ordered Item Booklets (OIB) as explained in the Chapter 3 (Cizek and Bunch, 2007). Every page in the booklet includes the order of the item, its original place in the test, the ability required for a 67% chance to answer the item correctly (in a rescaled form), the item itself, and the correct answer (Figure 3.2). The participants first examined the items, solved them, and compared them with the answers. Then they made their first round participation of the bookmarks identifying the cut points for the performance levels.

4.2.1 Results of First Round

After the bookmark participations were collected from the participants and summarized and organized by the researcher, the following Table was developed. This Table would then be given to the participants to use as feedback information presenting the decisions of each participant and their means. In the Table, for each performance level cut point (Basic, Proficient, and Advanced), there exists the individual page numbers, the theta ability value required for correctly answering the item on that page, and the corresponding raw score.

Table 4.5 Feedback Information Collected and Summarized after Round 1

	Basic			Proficient			Advanced		
Panelist No	Page numbers	Theta Cut Score	Raw Score	Page number	Theta Cut Score	Raw Score	Page numbers	Theta Cut Score	Raw Score
1	9	-0.025	31	20	0.363	32	45	1.030	48
2	6	-0.275	30	19	0.322	32	47	1.136	48
3	6	-0.275	30	15	0.222	32	46	1.104	48
4	11	0.129	31	23	0.480	46	49	1.178	49
5	5	-0.299	30	19	0.322	32	37	0.859	48
6	7	-0.142	31	15	0.222	32	46	1.104	48
Summary Statistics									
Mean Theta Cut Score		-0.148			0.322			1.069	
Mean Raw Cut Score		31			34			48	

At the bottom of the Table, one can find the mean of the theta values allocated with each participant’s decision. Also, the raw scores are presented. For example, the mean raw score for Basic performance level means that an examinee should have at least 31 out of 62 to be counted as reaching the Basic performance level. The scores below 31 will be taken as having the performance of Below Basic. Similarly, a student should get at least 34 and 48 to be counted as Proficient and Advanced performance levels, respectively.

This information namely the mean cut scores points for the performance levels, was effective on the participants’ decisions. When they were having a discussion before they started to make their second round of bookmark placements, they frequently discussed the meaning of those raw scores in terms of the students’ probability of achieving. For example, for the cut score point of 31 for Basic level of performance was found very high. The participants stated that the score of 31 meant 50% of the total score (62) and it was too high for being counted as Basic.

The participants also discussed the other participants’ individual item selections for each performance level. They turned to the items again and discussed the knowledge or skills required for each item and also compared their selections. After all of these discussions, they made their second round of item placements.

4.2.2 Results of Second Round

The participants made their decisions about the cut points for performance level for the second time, and these results are in the Table.

Table 4.6 Feedback Information Collected and Summarized after Round 2

	Basic			Proficient			Advanced		
Panelist No	Page numbers	Theta Cut Score	Raw Score	Page number	Theta Cut Score	Raw Score	Page numbers	Theta Cut Score	Raw Score
1	4	-0.404	29	17	0.237	32	40	0.871	48
2	4	-0.404	29	17	0.237	32	44	1.009	48
3	4	-0.404	29	17	0.237	32	46	1.104	48
4	9	-0.025	31	19	0.322	32	52	1.380	56
5	4	-0.404	29	16	0.322	32	44	0.009	48
6	7	-0.142	31	15	0.222	32	46	1.104	48
Summary Statistics									
Mean Theta Cut Score		-0.297			0.263			0.913	
Mean Raw Cut Score		30			32			49	

When the Table 3.6 is examined carefully, it can be observed that the participants' judgments about the places of the bookmarks were effected by the feedback given at the end of the round 1. During their discussions beginning the second round, most of the participants mentioned that their first judgments had been an overestimation for the lower group ie the Basic performance group. They also stated that they had taken most of the first questions in the OIB as very easy; however, the Table 3.5 showed them the group of examinees had been a very low ability group.

After the changes, it can be easily observed that the cut points for Basic performance level were carried to lower page numbers, which indicated that the Basic performance level should have been represented by easier items. Four of the six participants put their bookmarks for Basic level in front of the page number 4. This meant that most of the participants had decided that an examinee could have been counted as reaching the Basic level of performance if only he/she had answered the first three items correctly. On the other hand, in the first round, there was one participant who had been thinking that this number of items to be answered correctly should have been minimum 10. So, it is obvious that they lowered their expectations. However, this decrease in the number of correct items required did not make a great change on the raw score. The minimum raw score for Basic level of performance did become 30, which was only one point less than the first cut score point.

This is an indication of the fact that the replacement of the bookmark from the item 9, for example, to the item 4 did not mean a great change in the difficulty indices and, and related with this, in the ability values required for answering those items. This group of items could be taken as similar level of items with respect to their difficulties. This fact was mentioned in the discussion of the participants at the end of round two, when they were presented this Table.

IRT theory takes its strength from its ability to bring the parameters of items, that is, the item difficulty in this study, and the ability values required for answering those items correctly (Swaminathan & Hambleton). Therefore, the Table presents a lot of information about the difficulty parameters and the abilities required for answering each item correctly with the 67% probability. However, here, there is no information about the real performances of the examinees. To let the experts identify the items that characterize the minimum achievement level that an examinee should achieve to be a member of the group called Basic, Proficient or Advanced, it is needed to present them the impact data. In other words, the participants should be given the information that demonstrate the effect of their bookmark placement on the corresponding raw score and the percentage of the examinees at or above this score (Cizek, 2007). Therefore, at the beginning of the round 3, the Table was presented to the participants. The table was an expanded version of the Table in which the raw score allocated for each theta (θ) value and the percentage of the examinees at or above this raw score were also included.

Table 4.7 Raw scores and percentages at or above these scores

order of diff.	item no	RP %67	raw score	% examinees at or above	order of diff.	item no	RP %67	raw score	% examinees at or above
1	item1	-0.705	14	99.6	**32**	item41	0.738	48	22.7
2	item27	-0.557	14	99.6	**33**	item6	0.75	48	22.7
3	item17	-0.432	28	83.3	**34**	item42	0.794	48	22.7
4	item2	-0.404	29	81.4	**35**	item12	0.838	48	22.7
5	item10	-0.299	30	78.8	**36**	item14	0.859	48	22.7
6	item33	-0.275	30	78.8	**37**	item39	0.859	48	22.7
7	item4	-0.142	31	76.2	**38**	item46	0.860	48	22.7
8	item11	-0.101	31	76.2	**39**	item62	0.860	48	22.7
9	item3	-0.025	31	76.2	**40**	item53	0.871	48	22.7
10	item20	0.063	31	76.2	**41**	item50	0.924	48	22.7
11	item43	0.129	31	76.2	**42**	item34	0.946	48	22.7
12	item35	0.145	31	76.2	**43**	item45	0.967	48	22.7
13	item5	0.176	31	76.2	**44**	item24	1.009	48	22.7
14	item25	0.192	31	76.2	**45**	item51	1.030	48	22.7
15	item32	0.222	32	72.1	**46**	item60	1.104	48	22.7
16	item57	0.222	32	72.1	**47**	item8	1.136	48	22.7
17	item18	0.237	32	72.1	**48**	item28	1.157	48	22.7
18	item15	0.294	32	72.1	**49**	item22	1.178	49	17.5
19	item56	0.322	32	72.1	**50**	item13	1.199	49	17.5
20	item37	0.363	32	72.1	**51**	item52	1.209	49	17.5
21	item40	0.377	33	66.2	**52**	item54	1.38	56	4.1
22	item26	0.454	45	29	**53**	item31	1.435	57	2.2
23	item21	0.480	46	26.8	**54**	item49	1.435	57	2.2
24	item19	0.517	47	24.2	**55**	item44	1.536	57	2.2
25	item36	0.565	47	24.2	**56**	item58	1.570	57	2.2
26	item7	0.613	47	24.2	**57**	item30	1.593	58	1.1
27	item55	0.613	47	24.2	**58**	item61	1.826	58	1.1
28	item16	0.636	47	24.2	**59**	item9	1.893	58	1.1
29	item48	0.671	47	24.2	**60**	item47	1.992	59	0.4
30	item29	0.705	48	22.7	**61**	item38	2.021	59	0.4
31	item23	0.727	48	22.7	**62**	item59	2.146	60	0.4

4.2.3 Results of Third Round

The participants detected the Table to identify the cut scores for the performance levels in Grade 4 mathematics. The criterion that would direct their decision about the cut points for Basic, Proficient, and Advanced levels of performance was the points in the distribution where a "jump" occurred in the raw score allocated to the theta required.

When the participants examined the Table, they made their last decisions for the places of the bookmarks for performance levels. The Table presents their last decisions.

Table 4.8 Final Decisions of Participants for Cut Scores for Each Performance Level

	Basic			Proficient			Advanced		
Panelist No	Page numbers	Theta Cut Score	Raw Score	Page numb er	Theta Cut Score	Raw Score	Page numbers	Theta Cut Score	Raw Score
1	3	-0.432	28	22	0.454	45	52	1.380	56
2	3	-0.432	28	22	0.454	45	52	1.380	56
3	3	-0.432	28	22	0.454	45	52	1.380	56
4	4	-0.404	29	22	0.454	45	52	1.380	56
5	4	-0.404	29	22	0.454	45	52	1.380	56
6	3	-0.432	28	23	0.480	46	52	1.380	56
Summary Statistics									
Mean Theta Cut Score		-0.423			0.459			1.380	
Mean Raw Cut Score		28			45			56	

4.3 Results of Judgmental Method

4.3.1 Results of First Round

The participants studied each item and filled up the forms showing every item and its estimated level of performance. They did not discuss with the other participants and made their decisions individually. The Table shows the participants' first placements.

It can be seen that the participants' identifications were not continuously hierarchical. Although the items were organized in an ascending order of difficulty, there were items identified as, for example, Proficient, but the next one identified as Basic. Since the participants studied on the items only with the content information and corresponding learning outcome, they did not conclude with relative performance of examinees, they were not affected by this information.

However, the main issue for the task given to the participants was to identify the cut points for the performance levels; therefore, during the discussions, the participants were also directed by the researcher to keep this manner of being in a hierarchical order. But, if they really believed that an item was belong to a performance level which was lower than the performance level of the previous item, they were not forced to change it.

4.3.2 Results of Second Round

The participants were instructed to use the following coding for demonstrating their placements for the items: 1 for Below Basic, 2 for Basic, 3 for Proficient, and 4 for Advanced. The following Table shows the results of this second round. In the table, there are the placements of each participant and the median of these placements.

Table 4.9 Judges' placements of the items in the second round and median of the placements

Order of Item	P1	P2	P3	P4	P5	P6	P7	Median
1	1	1	1	1	2	2	2	1
2	1	1	1	1	2	2	2	1
3	1	1	1	1	1	2	2	1
4	1	1	1	2	2	1	2	1
5	1	1	1	1	2	2	2	1
6	1	2	1	2	2	3	2	2
7	1	2	2	2	2	2	2	2
8	1	2	2	3	3	3	2	2
9	1	3	2	2	2	2	2	2
10	1	2	2	3	3	3	2	2
11	1	2	2	3	3	3	2	2
12	1	1	2	3	3	3	4	3
13	1	2	2	3	3	3	2	2
14	1	2	2	2	2	2	2	2
15	1	2	2	3	2	3	2	2
16	1	3	2	2	3	2	2	2
17	1	2	2	3	3	3	3	3
18	2	2	2	3	2	3	2	2
19	2	2	2	3	3	3	2	2
20	2	3	3	3	4	4	2	3
21	2	3	3	3	4	3	2	3
22	2	3	3	3	3	3	4	3
23	2	3	3	3	3	3	3	3
24	2	3	3	3	3	3	3	3
25	2	3	3	3	3	3	2	3
26	2	3	3	3	3	2	2	3
27	2	3	3	3	3	3	2	3
28	2	3	3	3	3	3	2	3
29	2	3	3	3	3	3	2	3
30	2	3	3	4	4	4	3	3
31	2	3	3	4	4	4	3	3
32	2	3	3	3	3	3	3	3
33	3	3	3	3	3	2	3	3
34	3	3	3	3	3	3	2	3
35	3	4	3	3	3	3	3	3
36	3	4	3	3	3	3	2	3
37	3	4	3	3	3	3	3	3
38	3	4	3	3	3	2	3	3
39	3	4	4	3	3	3	2	3
40	3	4	3	3	3	3	2	3
41	3	4	3	3	3	3	4	3
42	3	4	3	3	3	2	3	3
43	3	4	3	3	3	3	4	3
44	3	4	3	3	3	3	3	3
45	3	4	3	4	4	3	4	4
46	3	4	3	4	4	3	2	3
47	3	4	3	4	4	3	3	3
48	3	4	4	4	4	4	2	4
49	3	4	4	3	3	3	3	3
50	3	4	4	3	3	3	3	3
51	3	4	4	4	4	4	3	4
52	3	4	3	4	4	4	4	4
53	3	4	4	3	3	3	2	3
54	3	4	4	4	4	4	4	4

Table 4.9 Cont'd

55	3	4	4	4	4	4	4	4
56	4	4	4	4	4	4	3	4
57	4	4	4	4	4	4	3	4
58	4	4	4	4	4	4	2	4
59	4	4	4	4	4	3	4	4
60	4	4	4	4	4	4	4	4
61	4	4	4	4	4	4	4	4
62	4	4	4	4	4	4	4	4

4.3.3 Results of Third Round

This Table was shared with the participants to make their last decisions and to clarify the final cut points for the performance levels. They had to make changes on only the items which had been classified as a representative for a performance level which was lower than the performance level of the previous item. For example, item 13 was classified as an item that could be answered by the examinees of Basic level of performance, whereas the item 12 was put into the Proficient level of Performance. The participants turned to the items once more and discussed the skills required for the items 12 and 13, and also the series of items until item 20 where there were discrepancies between the computed levels of those items. The expert participants tried to change either the levels of items 12 and 17 from Proficient to Basic or the levels of items 13, 14, 15, 16, 18, 19 from Basic to Proficient by thinking the learning outcomes of the items and the skills required. After their discussions, they decided to change the levels of items 12 and 17 from Proficient to Basic. They had to study the items 45, 48, and 53 in the same manner. The classification of the items with respect to final agreement of the participants on the median value of their previous placements and the transformed form of the codes given formerly to the items for showing the levels of performance are given in the Table 4.10.

Table 4.10 Judges' Agreement on Performance Levels

Order of Item	Median	Agreement Result	Performance Level
1	1	1	BB
2	1	1	BB
3	1	1	BB
4	1	1	BB
5	1	1	BB
6	2	2	B
7	2	2	B
8	2	2	B
9	2	2	B
10	2	2	B
11	2	2	B
12	**3**	**2**	**B**
13	2	2	B
14	2	2	B
15	2	2	B
16	2	2	B
17	**3**	**2**	**B**
18	2	2	B
19	2	2	B
20	3	3	P
21	3	3	P
22	3	3	P
23	3	3	P
24	3	3	P
25	3	3	P
26	3	3	P
27	3	3	P
28	3	3	P
29	3	3	P
30	3	3	P
31	3	3	P
32	3	3	P
33	3	3	P
34	3	3	P
35	3	3	P
36	3	3	P
37	3	3	P
38	3	3	P
39	3	3	P
40	3	3	P
41	3	3	P
42	3	3	P
43	3	3	P
44	3	3	P
45	**4**	**3**	**P**
46	3	3	P
47	3	3	P
48	**4**	**3**	**P**
49	3	3	P
50	3	3	P
51	4	4	A
52	4	4	A
53	**3**	**4**	**A**
54	4	4	A
55	4	4	A
56	4	4	A

Table 4.10 Cont'd

57	4	4	A
58	4	4	A
59	4	4	A
60	4	4	A
61	4	4	A
62	4	4	A

4.4 The Congruence of the Item Mappings between IRT-Based and Judgmental Methods

As mentioned in the previous chapters, the purpose of the study was to define the cut points on the ability scale for different performances and to identify the knowledge and skills that the students should reach for being classified in these levels of performance. Since two different methods were used, the congruence between these two methods is also important. The scaling of items by the experts and by the IRT-based procedure should be in correlation to state the performance level descriptors.

The Kendall coefficient of concordance between the median values of the judges' scaling and by the IRT-based mapping was significant with τ (62)=.696, $p<.001$. The relationship was also identified by calculating the Spearman's rho coefficient. The Spearman's rho correlation coefficient between the median values of the judges' scaling and by the IRT-based mapping was also significant with ρ (62)=.752, $p<.001$. The Table shows the coefficients and significance values.

Table 4.11 Correlation Results

			MEDIAN	**IRT**
Kendall's tau_b	MEDIAN	Correlation Coefficient	1.000	.696**
		Sig.(2-tailed)		.000
		N	62	62
	IRT	Correlation Coefficient	.696**	1.000
		Sig.(2-tailed)	.000	
		N	62	62
Spearman's rho	MEDIAN	Correlation Coefficient	1.000	.752**
		Sig.(2-tailed)		.000
		N	62	62
	IRT	Correlation Coefficient	.752**	1.000
		Sig.(2-tailed)	.000	
		N	62	62

**Correlation is significant at the 0.01 level.

4.4.1 The Mismatch Items

There 6 items out of 62 about which the judges did not have an agreement about their placement to the performance levels. These items were thought to be identifying the different performance levels with IRT-based and Judgmental methods. These mismatch items and their classification by the participants using two different methods were shown in the Table.

Table 4.12 Comparison of Performance Levels

Item Order	Performance Level IRT-based Method	Performance Level Judgemental Method
3	Basic	Below Basic
4	Basic	Below Basic
5	Basic	Below Basic
20	Basic	Proficient
21	Basic	Proficient
51	Proficient	Advanced

The items 3, 4, and 5 were classified as Basic performance level by the participants who used the IRT-based method. The reason was that on the ability scale where the point representing the required ability to answer item 3 correctly with 67% probability, there occurred a jump (Table). Moreover, while the percentage of students at or above the corresponding score to the ability required for answering item 3 correctly was 83.3%, the percentage of students at or above the corresponding score to the ability required for answering the previous item (namely, item 2) was 99.6%. This was an evidence for the participants to place their border between Below Basic and Basic performance levels.

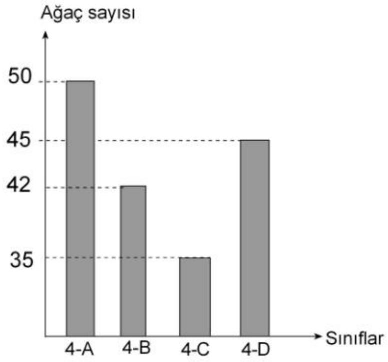
However, the participants, who were deciding the cut points with a more qualitative way where they used only the descriptors of the items and did not have any information about the relative performance of the examinees on those items, classified the items as Below Basic. When these items are carefully examined, they have common characteristics with the items in the Below Basic performance level and they slightly differ from the items measuring the similar outcomes that were placed in the Basic level.

For example, the item 3 can be seen in the Figure 4.2 with its content area, sub-area, and the learning outcome intended to measure. The item requires the examinee to read the data shown in the column graph correctly and to select the correct representation of those given data in the table. It was identified as the minimum skill needed for this content area.

Though the similar issues and the same content idea, the item 6 requires slightly different skills for answering. For this item, the examinee should transform the given data in the tally sheet into a column graph. Besides, the complex order of the names in the horizontal axis makes the question more difficult than the item 3 in the Figure 4.2. The judgmental procedure for mapping the items into performance levels was able to identify this small detail with the help of the experienced teacher evaluations and their group discussions. The item 6 is shown in the Figure.

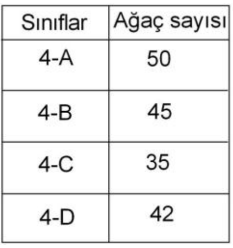
3

Item 17



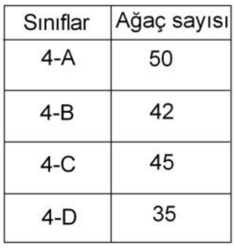
Yukarıda bir okuldaki dördüncü sınıf öğrencilerinin orman haftasında diktikleri ağaç sayısı görülmektedir. Bu grafik aşağıdaki tablolardan hangisiyle de gösterilebilir?

A)



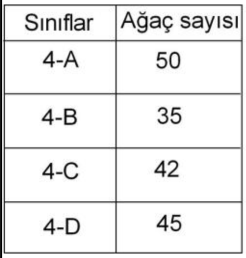
Sınıflar	Ağaç sayısı
4-A	50
4-B	45
4-C	35
4-D	42

B)



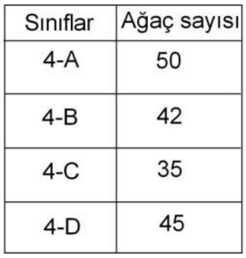
Sınıflar	Ağaç sayısı
4-A	50
4-B	42
4-C	45
4-D	35

C)



Sınıflar	Ağaç sayısı
4-A	50
4-B	35
4-C	42
4-D	45

D)



Sınıflar	Ağaç sayısı
4-A	50
4-B	42
4-C	35
4-D	45

Answer: D

Content Area: Data

Sub-Area: Column Graphs

Learning Outcome: can comment on the column graphs

Figure 4.2 Item 3

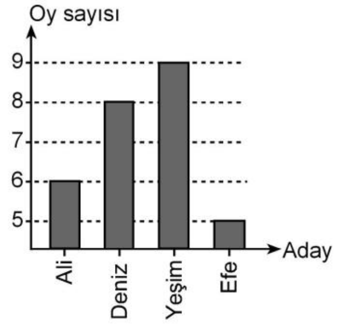
33)

ADAY	OY SAYISI
Ali	𝍸 /
Efe	𝍸
Yeşim	𝍸 ///
Deniz	𝍸 ////

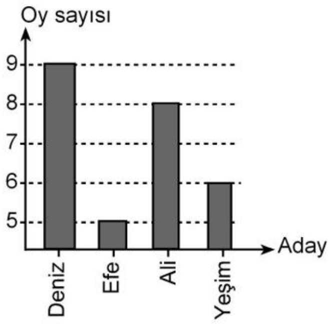
Yukarıda 4-A sınıfında yapılan başkanlık seçiminin sonuçlarını gösteren tablo görülmektedir.

Bu tablo aşağıdaki sütun grafiklerinden hangisinde doğru gösterilmiştir?

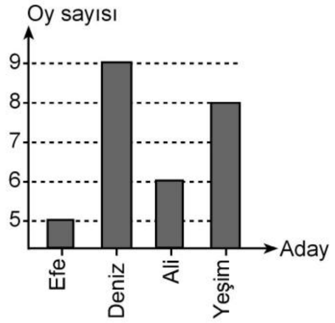
A)



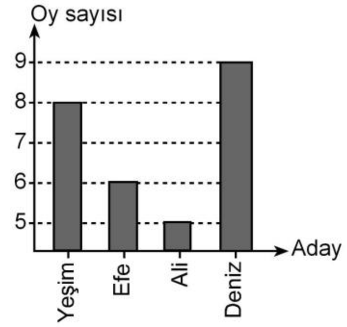
B)



C)



D)



Answer:C

Content Area: Data

Sub-Area: Column Graphs

Learning Outcome: can construct the column graphs

Figure 4.3 Item 6

4.5 Identification of the Performance Level Descriptors

As it was mentioned in the previous sections, the main purpose of the current study was to classify the scores gained from the mathematics exam into performance levels and identify the common characteristics of the mathematical abilities of the students belonging to each level. This could be done by clearly stating the skills required for solving the items grouped to different performance levels such as Below Basic, Basic, etc.

All the items in the test were developed for measuring an outcome stated in the national mathematics programme of Ministry of Education (MEB, 2005). The outcome statements could have possibly been written as the descriptors for those performance levels, however, as it was declared before, the national programme did have a very content dependent outcome sentences, which obviously did not let discriminate core and general skills or thinking process required for reaching expected performance levels. Therefore, the group of outcomes were summarized in general statements focusing on the skill rather than the content. For example, the following outcomes were taken from the 4th grade mathematics programme (MEB, 2005):

- Students can conduct addition with natural numbers.
- Students can conduct subtraction with natural numbers.
- Students can conduct multiplication with natural numbers.
- Students can conduct division with natural numbers.
- Students can conduct addition with fractions.
- Students can conduct subtraction with fractions.

The common characteristics of these outcomes were their focus on the computation skills, therefore, the items measuring these outcomes were then characterised as the following general statement:

"Students can model and conduct one-step operations (addition, subtraction, multiplication, division) with natural numbers and fractions."

Therefore, all the items and their outcomes were examined one by one to identify the general knowledge and skills they measured. The common characteristics of skills were discriminated from the content definitions; therefore, the performance level descriptors could have been identified in a more skills-oriented way (İş Güzel, 2008; Berberoğlu, 2009). Also, Kelly (1999) described the performance levels by the detailed descriptions of items; however, she identified the level descriptions in the summary of skills. The following tables present the detailed item structure of the performance levels and the descriptions for performance levels for Below Basic, Basic, Proficient, and Advanced levels of performance.

Table 4.13 Performance Level Descriptors for Below Basic Level

Performance Levels	Below Basic
Ordered Item Scale	Items 1-5
Performance Level Descriptions	They can order a group of natural numbers and can identify the minimum or maximum of them. Students can recognize the types of angles (acute, straight, etc.). Students can read the data from a column graph and organize it in a table. They also can identify the symmetry lines of regular geometric shapes.

When the descriptions are examined in Table 4.14, it can be seen that the skills required for this level are very limited. From the perspective of the content areas, it is expected from the student to have a very general knowledge and understanding of geometry, numbers, and data gathering. The characteristics items for this level are the first 5 questions in the ordered item booklet, which means the 5 easiest items in the test. Sample items can be found in APPENDIX A.

Table 4.14 Performance Level Descriptors for Basic Level

Performance Levels	Basic
Ordered Item Scale	Items 6-19
Performance Level Descriptions	Students can recognize the value of the given digit in a six digit number. Students can model and conduct the one-step operations (addition, subtraction, multiplication, division) with natural numbers and fractions. They can transform the data tables to column graphs. They can classify the triangles with respect to their angles. They can recognize the mirror images of geometric figures. They also can identify the symmetry lines of irregular and complex figures. They can conduct one step operations related to weight units in daily situations.

Starting from the Basic level, the descriptors began to have the hierarchical structure. It can be observed from the Table 4.15 that the mathematical processes defined in the Basic level required the skills defined in the Below Basic level and besides put new cognitive abilities on them.

Table 4.16 demonstrates characteristics of the proficient level students who started to gain higher order skills like transferring from one format to another and problem solving. When the level is assessed in terms of the weight of the content areas, it can be said that these students have began to understand and apply the number concept more proficiently.

Table 4.15 Performance Level Descriptors for Proficient Level

Performance Levels	Proficient
Ordered Item Scale	Items 20-50
Performance Level Descriptions	Students can conduct the multi-step operations (addition, subtraction, multiplication, division) defined in daily life situations. They can match the numbers (natural numbers, fractions, or decimals) with their models and transform the numbers to the points on the number line. They can match at most six digit numbers with their expanded forms. They can comment the statements including probability expressions. Students can transfer the verbal expressions or visual representations into mathematical expressions. Students can recognize the types of angles in plane geometric figures and in real life situations. They can transform the units of length and weight and solve routine problems using these units.

The highest level of performance which was labelled as Advanced required the cognitive processes of estimation, problem solving, and identifying and producing relationships. This finding is consistent with all the national and international studies aimed to identify the benchmarks or descriptors for performance levels (Martin, Mullis, & Foy, 2008; OECD, 2003; Kelly, 1999; İş Güzel, 2008; NAEP, 2001). In all these studies, the highest level of performance included reasoning, estimation, multi step problem solving, and drawing conclusions from complex situations. Table 4.17 and the sample items in APPENDIX A present examples measuring the higher order skills.

Table 4.16 Performance Level Descriptors for Advanced Level

Performance Levels	Advanced
Ordered Item Scale	Items 51-62
Performance Level Descriptions	Students can identify the relationships in the number or shape patterns and can find the missing or following component that is consistent with this relationship. They can solve non-routine and multi-step problems which require understanding and application of knowledge of several content areas like numbers, measurement, geometry, etc. They can also use the problem solving strategies and the estimation strategies based on measurement or operation.

To summarize, the results from the standard setting sessions both by the judgemental method and by the IRT based method revealed consistent and matching item mappings. The cut points were marked similarly with respect to these item mappings. And in addition to these results, the descriptors for performance levels were designated in a hierarchical manner which required fundamental thinking skills for the lower levels, however, required complex reasoning and problem solving skills for higher levels.

4.6 Path Analysis for Identifying the Prerequisite Cognitive Processes

As mentioned above, the formation of the performance level descriptors could be taken as an evidence for the validity of the study in terms of the construct validity. The items which had been constructed to measure the learning outcomes of mathematics programme were then mapped consistently by the judgmental and statistical methods grouping skills in a hierarchical structure. On the other hand, it was still needed to show empirical evidence for the validity of this model in terms of the skills' relationships. This was satisfied by the path analysis conducted with LISREL software (SSI Inc., 2003).

The causal submodel of LISREL models includes only the observed variables and the analysis aims to check the fit of the proposed model to the used data (Jöreskog & Sörbom, 2001). The relationships or the bivariate regressions between the directly observed or measured variables are detected and the statistically significant ones are included in the model. Then, the model is checked with respect to its fit with the data.

In the current study, the mathematical skills were grouped generally as "mathematical understanding", "mathematical computation", and "problem solving" consistent with İş Güzel (2008). These cognitive processes could easily be the framework generalising the learning outcomes when their content dimensions were eliminated. However, to be able to make a more detailed path analysis to find empirical evidence for the hierarchical structure of the required skills for each performance level, these three overall dimensions were divided into subdimensions of cognitive processes based on the descriptions of processes in the performance levels. The subdimensions of the cognitive processes in mathematics used in the study were given below. The abbreviations in parentheses showed the names of the variables used in the LISREL analysis.

"Mathematical Understanding"

Identifying and modelling numbers (IDENTNUM)

Recognizing angles and shapes (RECOGSHP)

Finding and using symmetry context (SYMMETRY)

Finding and defining patterns (PATTERNS)

"Mathematical Computation"

Conducting one step operations (ONESTEP)

Conducting multi step operations (MULTI)

"Problem Solving"

Solving routine problems (ROUTINE)

Solving non routine (complex) problems (NONROUTN)

These sub-dimensions were developed by investigating the descriptors of performance levels and the items correspondingly to find out the general skill definitions. For example, the dimension "identifying and modelling numbers" were related to the following statements of performance descriptors:

1. They can order a group of natural numbers and can identify the minimum or maximum of them. (Below Basic)
2. Students can recognize the value of the given digit in a six digit number. (Basic)

The related items with these sub-dimensions from the test were taken to create the score for the variable IDENTNUM. These items were items 27, 20, 37, 40, 21, 19, 23, 42, and 22 in the original test and the IDENTNUM variable were then calculated for each examinee by adding the scores for each of these items. The rest of the variables were calculated in the same way.

The sub-dimensions did span one or more consecutive performance levels protecting the hierarchical structure. Moreover, the relationships among the variables proposed in the model to test with LISREL based on this structure. The sub-dimensions and their corresponding performance levels can be summarized as follows:

Identifying and modelling numbers (IDENTNUM): Below Basic, Basic and Proficient

Recognizing angles and shapes (RECOGSHP): Below Basic and Basic

Finding and using symmetry context (SYMMETRY): Below Basic and Basic

Finding and defining patterns (PATTERNS): Advanced

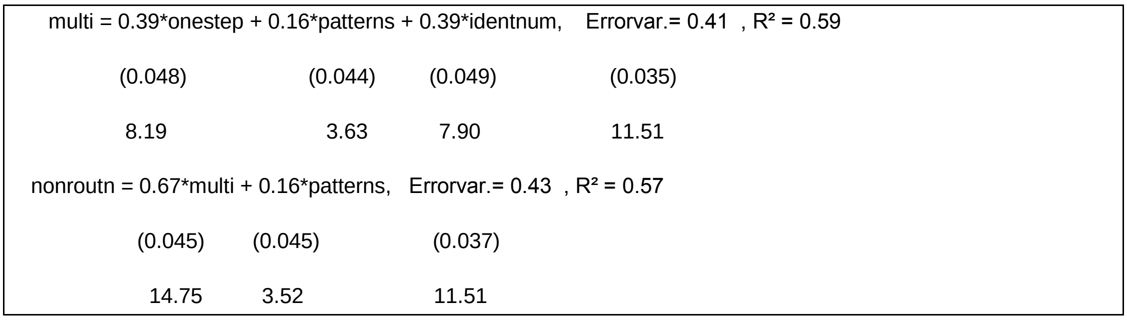
Conducting one step operations (ONESTEP): Basic

Conducting multi step operations (MULTI): Proficient

Solving routine problems (ROUTINE): Proficient

Solving non routine (complex) problems (NONROUTN): Advanced

Therefore, the significant relationships between the sub-dimensions were put in the path analysis and the following equations were found and presented in Figure (refer to APPENDIX C for the LISREL Syntax and APPENDIX D for the output):



```
  multi = 0.39*onestep + 0.16*patterns + 0.39*identnum,   Errorvar.= 0.41  , R² = 0.59

        (0.048)           (0.044)      (0.049)            (0.035)

         8.19              3.63         7.90               11.51

nonroutn = 0.67*multi + 0.16*patterns,   Errorvar.= 0.43  , R² = 0.57

          (0.045)     (0.045)            (0.037)

           14.75       3.52               11.51
```

Figure 4.4 Structural Equations for the Proposed Model

The model revealed significant relationships and high R² values for the equations. The structural equations demonstrated that the higher order thinking skills like non-routine problem solving or conducting multi step operations could be mostly explained by the achievement in the related but lower level thinking skills such as identifying numbers and conducting one step operations. The model fit well with the data and the selected goodness-of -fit indices were presented in the Table 4.17.

Table 4.17 Selected Goodness-of-Fit Indices

Goodness-of-Fit Indices	Goodness-of-Fit Index Values
RMSEA	.056
AGFI	.96
GFI	.99
Standardized RMR	.016
Chi-square (p-value)	3.72 (P = 0.16)

CHAPTER 5

DISCUSSION, CONCLUSIONS, RECOMMENDATIONS

5.1 Performance Level Descriptors

The main purpose of this study was to fill up one of the most important gaps in the area of measurement in education in Turkey. Recent studies showed that the results of the measurement activities in education were not basically used for feedback to the students, reporting to the parents, or identifying the quality of the curriculum (Berberoğlu, 2007). Despite the lack of all these issues in education lead to important disadvantages, its missing usage in giving feedback to students and parents should be the one that must be overcome immediately. Quantitative measurement results should be defined qualitatively to identify the performances of students who gained these scores. When the quantitative results are the only indicators of success or achievement in education, the measurement of student performance, unfortunately, focuses on the ranking of students by relatively ordering the scores. Students' level of reaching the higher order skills of thinking loses its importance. One of the announced reasons of the change in the high school entrance and examination system by the Ministry of Education was to transform the purpose of that examination system into a process for using its results for reviewing and renewing the curriculum (MEB, 2005).

The current study, therefore, aimed to describe the meanings of the scores gained in a test. Mathematics descriptors were identified for the performances of 4th grade students in a group of private schools in İstanbul.

The performance descriptors indicated several important results in terms of the characteristics of the students' knowledge and skills.

Firstly, the levels of performance demonstrated a hierarchical character. Lower levels of certain skills could be observed in lower competency levels, whereas more developed and complex skills could be observed in advanced levels. For example, students with a Basic level of competency could conduct one step operations including weight units, while "Proficient" students were able to solve routine problems with these units. Moreover, the students who had reached Advanced levels could both solve non-routine problems and use estimation strategies while solving those problems.

Secondly, each performance level included nearly all of the content areas consistently with the hierarchical manner explained above. One can follow the development of certain skills through the levels of performance and also these are independent of the content area. For example, problem solving was identified as higher order thinking skill, which was developed gradually through the levels from Basic to Advanced. It started as an ability to conduct operations in real life situations, developed as routine problem solving and finalized as using estimation strategies. Besides, when the items requiring problem solving skills were analyzed, it could be seen that these items were related to several content areas and sub-areas such as natural numbers, fractions, geometry, measurement, etc. This can be taken as the expert judges' decisions about these performance levels were not totally based on the content area of the items, but the skills required for that item.

Thirdly, the performance levels identified in this study showed a parallelism with the ones in several other researches. Berberoglu (2009) presented the competency levels of 3rd grade students in the sub-area of numbers. Although the grade levels investigated in the current study and the one in Berberoglu (2009) were different, conceptual similarities could be observed. For example, the following skill was identified as one of the skills that a Proficient level student should carry: *Students can transfer the verbal expressions or visual representations into mathematical expressions.* A similar skill was stated in Berberoğlu (2009) as the skill of 4th level students, which was corresponded the Advanced level in the current study: *Students* *can match the given operation with the appropriate problem expression (p: 19, translated).* It can be inferred that the similar skills were identified as higher order levels of thinking in both studies, assuming the small difference in the competency levels were due to that one study was conducted with 3rd graders whereas the other one with 4th graders.

Another similar classification of skills in both studies was related with the data tables and graphs. Berberoglu stated that reading data from the figure graphs was a skill for the 2nd level of students. In the current study as well, the skill of reading the data from a column graph and organizing it in a table required a student to have at least the Basic level of performance. This exact match of these skills in both studies was also important in terms of the evidence that the real performance of students were independent of the content of the items. Although the content changed from figure graphs in 3rd grade to column graphs in 4th grade, the skill of reading data from these graphs kept its place as a Basic level character.

The performance level descriptors' validity was also detected with LISREL path analysis and the model revealed good fit for the data. Moreover, from the path analytic model proposed and fit in the analysis, several conclusions could be derived for better mathematics teaching and learning practices.

The study, firstly, showed that mathematical concepts such as numbers had been the fundamentals for developing any further or higher thinking skills. Teachers should focus on the concepts as a basic and the applications of these concepts should be built on them.

Secondly, the general and fundamental skills which were important for developing and improving the children's cognitive processes should be identified and focused independent of the content. For example, conducting multiplication with natural numbers and fractions were not different skills or cognitive processes for the children. The differences emerged from the technical details or came from the algorithmic procedures, but the concept of multiplication was kept same. Teaching these as separate two tasks would both lead to heavy work load for children's cognitive processes and also prevent them to learn holistically. On the contrary, the content should be taken as the tool to use for making the abstract concepts more concrete.

Lastly, problem solving skills were significantly separated from the other constructs such as understanding and computation, but at the same time those skills were prerequisites for problem solving. The path analytic model obviously showed that complex problem solving skills were discriminated from the multi step operations and routine and simple problems. Problem solving also accompanied by the complex strategy use and estimation techniques and should be defined well by the teachers to prevent the misidentifications (İş Güzel, 2009).

5.2 Standard Setting Methods

When the literature on identifying the level descriptors for students' performances based on the results of certain tests, there was the concern that the expert judgments could be subjective or the procedures could be cumbersome for these panelists . Also, several researchers sought the ways of checking the validity of these judgments' item mappings . While the methods (Angoff, 1971) which were used most frequently and for a long time were more expert judgment based, new, modified, and more objective methods were developed to overcome these reliability and objectivity problems.

One of the purposes of this study was to investigate the participants' judgments about setting the cut points for levels of 4th grade students' mathematics performance and congruence of this arraying of items with another one made by an IRT-based method. Panelists were experienced teachers or measurement experts in the first method and it was standing on the panelists' knowledge and study on the mathematics programme outcomes of every item and the skills required for answering these items. During the standard setting process, teachers classified each item with respect to the following question/s:

Can a student who has the Below Basic / Basic / Proficient / Advanced level of performance answer this item correctly with a probability of 67%?

On the other hand, the second method was based on the identification of the "jump" points on the scale of ability scores required for answering each item correctly. The scale of ability scores was developed by using the item difficulty indices gathered from BILOG-MG, the software for detecting IRT parameters. The jump points indicated the great changes in the ability values which showed that there was a difference in the skills required for answering two items on the two sides of this jump point. Therefore, when this information was given to the teachers, it definitely affected their decisions on the places of the cutting points and they made their final decisions under the light of this impact information.

However, the teachers using the judgmental method stated that they decided the places of these cutting points mostly with respect to the skills that were attempted to be measured by each item, their estimations for the difficulty of the item, and their experiences and thoughts about the students' possible performances on these items. It was also observed during the discussion sessions between the consecutive rounds that teachers were affected by the placements of other teachers and their thoughts.

Although the sources and types of information which was effective on the decisions of two groups were different, the congruence between the mappings by these two methods indicated a high correlation of ρ (62) =.752 of Spearman coefficient. The importance of this result was that we could take the participants' judgments as evidence for IRT based method.

Despite the high correlation between the settings of two methods, there were 6 mismatched items. When they were examined, it could be inferred that three of these six items (items 3, 4, and 5) were classified as easier items by the judges than the IRT-based method. The remaining three of mismatched items (items 20, 21, and 51), on the contrary, were classified as more difficult items by the judges than the IRT-based method. When discussed with the judges, they commented on their thoughts of the performances of the students and the discrepancy of these thoughts from the students' real performances. They mentioned that the students whom they had guessed as low ability group had been weaker, on the contrary, the students, who were received as high ability group, scored better than they had guessed. They also stated that it was difficult to estimate the performances of the students who had performed on the extreme points, in other words, who scored very low or very high.

Two of these six mismatched items (items 4 and 5), which were categorized as Below Basic by judgmental method but Basic by IRT-based method, were about content area of geometry and one of them (item 3) was about content area of data. All three items were attempting to measure understanding of fundamental mathematical language or definitions. Two items (item 20, 21), which were classified as Proficient by judgmental method but Basic by IRT-based method, were belonging to the content area of numbers and they were again measuring the understanding of mathematical concepts. The last of these mismatched items (item 51), which was classified as Advanced by judgmental method but Proficient by IRT-based method was belonging to the content area of geometry. The item required the application of ability of computing to the daily life situations.

5.3 Factors Affecting the Panelists' Decisions

Although two methods of standard setting used in this study were basically different in terms of the feedback given to the panelists and hereby of the procedures, both of the methods consisted of panelists' decisions. Therefore, the factors affected the panelists could be observed in both studies. The most important factor affected the panelists using the IRT-based method was the impact data. The impact data was presented to the panelists as the percentage of students who scored at or above the raw scores that were allocated with the points where they had put the bookmarks. Although these panelists were focusing on the items, their contents, their perceived difficulties, and to some extent the skills required for answering the item until the last round of the session where they were presented the impact data, then, they totally structured their decisions upon these data. This result was consistent with some findings of the study conducted by Ferdous and Plake (2005), which indicated that, for most of the panelists, the norm-referenced feedback had been more influential for their decisions.

The literature showed that, when judgmental procedures were used and norm-referenced feedback or impact data were not given to the participants, panelists were mostly affected by the small group discussions for changing or finalizing their performance level discussions (Dawber & Lewis, 2002). Similarly, in this study, when the panelists were discussing their decisions between the rounds in both methods, they were affected by the opinions of other participants. Especially, the teachers who had more experience with the 4th grade students were also more effective on the other teachers. This fact, on the other hand, was also consistent with the reviewed studies (Ferdous & Plake, 2005; Dawber & Lewis, 2002). Both studies showed that most of the judges referred to their experiences with the students or they thought one of their students as the one characterizing the skills required for a specific level of competence, such as Basic or Proficient.

Two of the studies (Skorupski & Hambleton, 2005; Ferdous & Plake, 2005) indicated that teachers attending the standard setting sessions conducted for taking state-wide decisions about the cut score for reaching standards, were affected by the possible results of their decisions. In other words, teachers might be thinking to set lower standards for letting more of their students reach higher levels. Therefore, it can be said that political issues may affect setting educational standards. However, since the current study was conducted for only research purposes and this was clearly stated to the participant teachers, this factor was not influential for this study.

5.4 Future Directions

The current study identified the performance levels for 4th grade mathematics and their descriptions in terms of the skills required in the context of the test used. The students participated in the study were the students from 8 different private primary schools owned by the same foundation. Therefore, further studies can follow the performances of the students belong to the same levels of performance in this current study to check their improvements. Cizek (2007) recommended the method called "Vertically Moderated Standard Setting (VMSS)", which could give the opportunity to compare students' levels of performance across grades.

> "Addressing the challenges, it would seem, would involve developing and implementing standard-setting methods that set performance levels in concert, that is, across all affected grade levels (and perhaps subject areas) with some method smoothing out differences between grades. One approach to the challenge is found in what has come to be known as vertically-moderated standard setting (VMSS)." (p.253).

Furthermore, the study can be expanded to more general assessment procedures like SBS. The primary school students are taking the assessments at the end of the grades 6, 7, and 8 –called SBS-. The main purpose of these assessment programmes is to evaluate primary curricula for the contexts of Turkish, mathematics, science, social studies, and second language. However, the results are not analysed for identifying the descriptions of students' performance levels, and this leads to a perception of these exams as "competition exams" by the students, parents, and even by the schools. This study can be a simple model for expanding the implementation to the context of high-stake exams for institution- or nation-wide conclusions.

One of the most important limitations of the current study was its dependency to the multiple choice items. In the literature, most of the studies on both the comparative studies and standard setting included mixed types of items like constructed and multiple choice items ( ). Since the test was administered and assessed commonly among 8 primary schools, only multiple choice items could be used. However, the study should be expanded to a form including other types of items, and even to a form assessing with "performance tasks".

In the current study, the mathematics curriculum was taken as a whole, in other words, the results were not analysed with respect to different content areas of numbers, geometry, data, and measurement. The performance level descriptions included the skills required for all content areas as a combination. However, in the related literature, there were studies which discriminated the skills for different content areas and presented descriptions for performance levels separately (Berberoğlu, 2009;

Berberoğlu, Demirtaşlı, İş Güzel & Konak, 2008; IBO, 2007b). Another recommendation for further implication of this current study can be identifying the specific skills required for each content area. To this purpose, the number of items related to each content area in the measurement tool should be approximately equalized and the frameworks for each content area should, therefore, be extended. This will also help to describe the relationships between the knowledge and skills for each content area more specifically.

REFERENCES

Berberoğlu, G. (2006). Cito Öğrenci İzleme Sistemi. Paper presented at the meeting “Cito Türkiye Vizyon 2007”. Retrieved on May 26, 2007 from *http: // www . cito.com.tr/cito turkiye/etkinlikler/~/ media/ cito tr/ bestanden/ cito gb istanbul % 208 % 2012 % 2006%20jk.ashx*

Berberoğlu, G. (2009). Madde Haritalama Yöntemi ve Cito Türkiye Öğrenci İzleme Sistemi (ÖİS) Uygulamalarında Yeterlilik Düzeylerinin Belirlenmesi. *Cito Eğitim: Kuram ve Uygulama*. Ankara: Cito.

Berberoğlu,G., Demirtaşlı, N., İş Güzel, Ç. and Konak, Ö.A. (2008). Item Mapping in the Turkish Pupil Monitoring System. Paper presented at the 6th Conference of the International Test Commission, Liverpool.

Çet, S. (2006). A Multivariate Analysis in Detecting Differentially Functioning Items Through the Use of Programme for International Student Assessment (PISA) 2003 Mathematics Literacy Items. Unpublished Doctoral Thesis, Middle East Technical University, Ankara, Turkey.

Cizek, G.J. & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oqaks, California: Sage Publications

Dawber, T. & Lewis, D. M. (2002). The Cognitive Experience of Bookmark Standard Setting Participants. Paper presented at the annual meeting of American Educational Research Association. Retrieved on August 2, 2009 from

http://www.Education.ualberta.ca/educ/psych/crame/files/dawlew.pdf

EARGED (2005). PISA 2003 Projesi Ulusal Ön Rapor. Retrieved on February 8, 2006 from http://earged.meb.gov.tr/pisa/dil/tr/pisa2003.html

Egitek (2009). 2009 Ortaöğretime Geçiş Sistemi Sayısal Bilgileri. Retrieved on August 4, 2009 from http://egitek.meb.gov.tr/sinavlar/Istatistikler/2009/sbs/I_Yerles_Taban_Tavan/I_YerlesTaban_Tavan_SayisalBilgiler.pdf

Egitek (2009). 2009 Yılı Seviye Belirleme Sınavları Test ortalama ve Standart Sapmaları. Retrieved on August 4, 2009 from http://egitek.meb.gov.tr/sinavlar/Istatistikler/2009/sbs/I_Yerles_Taban_Tavan/I_YerlesTaban_Tavan_SayisalBilgiler.pdf

Ferdous, A.A & Plake, B.S (2005). Understanding the Factors that Influence Decisions of Panelists in a Standard-Setting Study. *Applied Measurement in Education*, 18,3,257-267.

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

IBO (2002). PYP: A Basis for Practice. Geneva: International Baccalaureate Organization.

IBO (2003). *Primary Years Programme Mathematics: Scope and Sequence*. Geneva: International Baccalaureate Organization.

IBO (2007a). *PYP Coordinator's Handbook 2007-2008.* Cardiff : International Baccalaureate Organization.

IBO (2007b). *Middle Years Programme Mathematics Guide*. Cardiff : International Baccalaureate Organization.

Impara, J.C. & Plake, B.S. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34,4,353-366.

İs Güzel, Ç. (2006). A Cross-Cultural Comparison of the Impact of Human and Physical Resource Allocations on Students' Mathematical Literacy Skills in the Programme for International Student Assessment (PISA) 2003 Mathematics Literacy Items. Unpublished Doctoral Thesis, Middle East Technical University, Ankara, Turkey.

İş Güzel, Ç. (2008). Cito Türkiye Öğrenci İzleme Sistemi'nde (ÖİS) Matematik Alanının Yapısı. *Cito Eğitim: Kuram ve Uygulama*. Ankara: Cito.

İş Güzel, Ç. (2009). Cito Türkiye Öğrenci İzleme Sistemi'nde (ÖİS) Problem Çözme Becerilerinin Ölçülmesi, Etkinlik ve Sorularla Örneklendirilmesi ve Değerlendirilmesi. *Cito Eğitim: Kuram ve Uygulama*. Ankara: Cito.

Jöreskog, K. & Sörbom, D. (2001). *LISREL 8: User's Reference Guide.* Scientific Software International, Inc.,Lincolnwood.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) Achievement Scales Using Scale Anchoring.* Unpublished Doctoral Thesis, Boston College, Boston, USA.

Martin, M. O., Mullis, I. V.S. & Foy, P. (2008). TIMSS 2007 International Mathematics Report. TIMSS & PIRLS International Study Centre, Boston College.

MEB (2005). İlköğretim Matematik Dersi 1-5. Sınıflar Öğretim Programı. Retrieved in September 13, 2005 from http://ttkb.meb.gov.tr

OECD (2003). PISA 2003 Assessment Framework. OECD. Retrieved on April 4, 2009 from, http://www.pisa.oecd.org/ dataoecd/ 46 /14/ 33694881.pdf

OSYM (2006). 2006 ÖSYS Başvuru veSınavlara İlişkin Sayısal Bilgiler. Retrieved on August 4, 2009 from
http://www.osym.gov.tr/BelgeGoster.aspx?F6E10F8892433CFF1A954
7B61DAFFE2A778DA91DA71E47ED

Scientific Software International (2003). BILOG-MG Version 3.0

Skorupski, W.P. & Hambleton, R.K. (2005). What are Panelists Thinking When They Participate in Standard-Setting Studies?. *Applied Measurement in Education*, 18,3, 233-256.

Van der Schoot,F.C.J.A (2002). The Application of an IRT-based Method for standard Setting in a Three-stage Procedure*.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Verhelst, N.D. (2004). Item Response Theory. Section G of Preliminary Pilot version of the Manual for "Relating Language examination the Common European Framework of Reference for Languages: learning, teaching, assessment. Retrieved on January 24, 2008 from *http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionG.pdf*

Yıldırım, H.H. (2006). *The Differential Item Functioning (DIF) Analysis of Mathematics Items in the International Assessment Programs*. Unpublished Doctoral Thesis, Middle East Technical University, Ankara, Turkey.

APPENDIX A

Items Exemplifying the Performance Level Descriptors

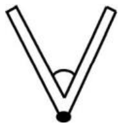
Below Basic Level

5

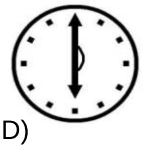
"Item 10"

Aşağıda belirtilen açılardan hangisi doğru açıdır?

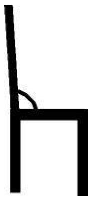
A)



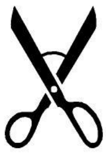
B)



C)



D)



Answer: B	
Content Area:	*Geometry*
Sub-Area:	*Angles*
Learning Outcome:	*can recognize the types of angles*
Skill required:	*understanding*
Percentages of Correct Responses	
Below Basic	70.2 %
Basic	93.3 %
Proficient	97.8 %
Advanced	93.8 %

Figure A1 Item 5

Below Basic Level

2

“Item 27”

Ay	**Ziyaretçi sayısı**
Nisan	6085
Mayıs	6079
Haziran	6080
Temmuz	6090

Yukarıdaki tablo bir müzeyi dört ay boyunca ziyaret eden kişilerin sayısını göstermektedir.

Tablodaki bilgilere göre müze ziyareti hangi ayda en az kişi tarafından yapılmıştır?

A) Nisan B) Temmuz

C) Haziran D) Mayıs

Answer: D

Content Area:	*Numbers*
Sub-Area:	*Natural Numbers*
Learning Outcome:	*can order at most six digit numbers*
Skill required:	*understanding*

Percentages of Correct Responses

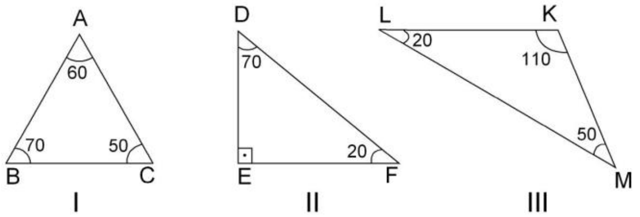
Below Basic	78.9 %
Basic	94.8 %
Proficient	97.8 %
Advanced	100 %

Figure A2 Item 2

Basic Level

9

“Item 3”



Yukarıdaki üçgenler, sırasıyla aşağıdakilerin hangisinde doğru verilmiştir?

A) Dar, geniş ve dik açılı üçgen

B) Geniş, dik ve dar açılı üçgen

C) Dar, dik ve geniş açılı üçgen

D) Dik, dar ve geniş açılı üçgen

Answer: C

Content Area:	*Geometry*
Sub-Area:	*Triangles, Squares, Rectangles*
Learning Outcome:	*can classify triangles w.r.t their angles*
Skill required:	*understanding*

Percentages of Correct Responses	
Below Basic	54.4 %
Basic	89.6 %
Proficient	97.8 %
Advanced	96.9 %

Figure A3 Item 9

Basic Level

12

"Item 35"

4 ■ 6 1 3
+ 1 7 8 ▲ 9
★3 4 7 2

Yukarıdaki toplama işleminde ■, ▲ ve ★ yerine aşağıdakilerden hangisi yazılmalıdır?

A) ■ : 4
▲ : 5
★: 6

B) ■ : 5
▲ : 6
★: 7

C) ■ : 4
▲ : 6
★: 6

D) ■ : 5
▲ : 5
★: 6

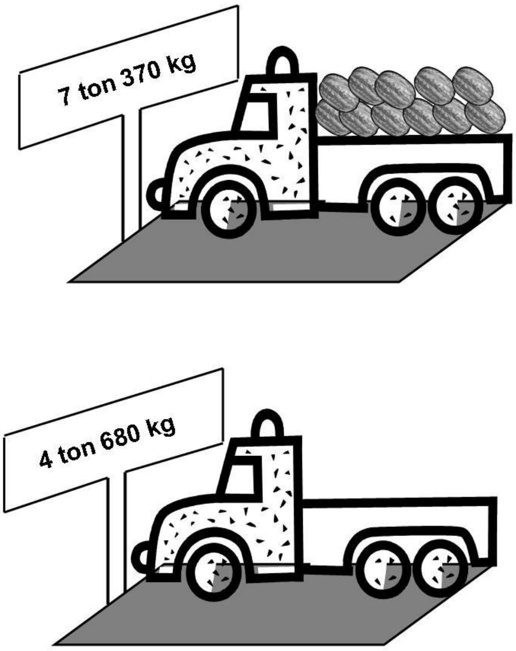
Answer: D	
Content Area:	*Numbers*
Sub-Area:	*Triangles, Squares, Rectangles*
Learning Outcome:	*can make addition with at most four digit natural numbers*
Skill required:	*operation*
Percentages of Correct Responses	
Below Basic	54.4 %
Basic	84.3 %
Proficient	91.3 %
Advanced	96.9 %

Figure A4 Item 12

Basic Level

19

“Item 56”



Adana’dan İstanbul’a karpuz taşıyan bir kamyonun karpuzlarla birlikte ağırlığı 7 ton 370 kg ’dır.
Bu kamyonun yüksüz ağırlığı 4 ton 680 kg olduğuna göre, bu kamyonun taşıdığı karpuzların ağırlığı ne kadardır?

A) 2 ton 150 kg

B) 2 ton 30 kg

C) 2 ton 690 kg

D) 2 ton 600 kg

Answer: D	
Content Area:	*Numbers*
Sub-Area:	*Weighing*
Learning Outcome:	*can solve problems including weight units*
Skill required:	*problem solving*
Percentages of Correct Responses	
Below Basic	49.1 %
Basic	75.4 %
Proficient	95.7 %
Advanced	100 %

Figure A5 Item 19

Proficient Level

38

“item 46”

Ece marketten tanesi 1300 YKr’tan 3 kutu süt, 850 YKr’tan 1 paket çikolata ve tanesi 40 YKr’tan 6 adet sakız satın aldı.

Ece’nin ödediği para kaç YKr’tur?

A) 4990 B) 4870

C) 4750 D) 2190

Answer: A	
Content Area:	*Numbers*
Sub-Area:	*Multiplication*
Learning Outcome:	*can construct and solve problems including multiplication*
Skill required:	*computation*

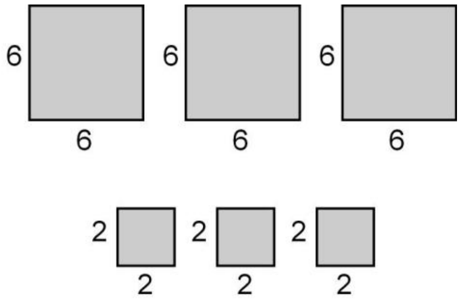
Percentages of Correct Responses	
Below Basic	33.3 %
Basic	53.7 %
Proficient	84.8 %
Advanced	93.8 %

Figure A6 Item 38

Advanced Level

62

“Item 59”



Elif, yukarıda kenar uzunlukları verilen 6 adet karesel kartonu kesmeden ve bükmeden hiç boşluk olmayacak şekilde bir araya getirip dikdörtgen oluşturuyor.

Bu dikdörtgenin çevre uzunluğu kaç birimdir?

A) 48 B) 52

C) 72 D) 96

Answer: B	
Content Area:	*Measurement*
Sub-Area:	*Area*
Learning Outcome:	*can compute the area measures of square and rectangular regions by using square units*
Skill required:	*problem solving*
Percentages of Correct Responses	
Below Basic	3.5 %
Basic	11.9 %
Proficient	28.3 %
Advanced	62.5 %

Figure A7 Item 62

APPENDIX B

BILOG-MG Code for One-Parameter Model for 62-item Test

```
>GLOBAL DFName = 'C:\Documents and
Settings\selcen\Belgelerim\Mat2345.prn',
        NPArm = 1;
>LENGTH NITems = (62);
>INPUT NTOtal = 62,
       NALt = 2,
       NIDchar = 11;
>ITEMS INAmes = (ITEM01(1)ITEM62);
>TEST1 TNAme = 'MATHTEST',
       INUmber = (1(1)62);
(11A1, 62A1)
>CALIB CRIt = 0.0500,
       ACCel = 1.0000,
       CHIsquare = (20, 6),
       RASch;
>SCORE
```

Figure B1 Extract from BILOG-MG Code

APPENDIX C

SIMPLIS syntax for identifying path relations for observed variables with LISREL 8.54

```
Thinking Skills
Observed Variables
onestep multi routine nonroutn patterns identnum symmetry
recogshp
Covariance Matrix from file 'C:\Documents and Settings\pc\My
Documents\skills\Think-2.cov'
Sample Size = 269
Relationships
nonroutn = patterns multi
multi = patterns onestep identnum
symmetry
recogshp
Path Diagram
End of Problem
```

Figure C1 Extract from SIMPLIS Syntax

APPENDIX D

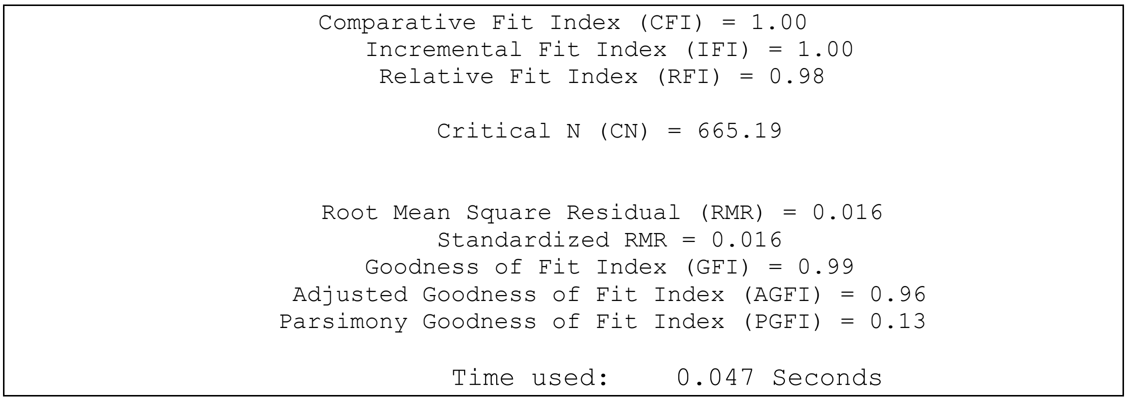
Selected output from LISREL analysis for defining the relationships among skills

```


               LISREL 8.54 BY   Karl G. Jöreskog & Dag Sörbom


 The  following  lines  were  read  from  file  C:\Documents  and  Settings\pc\My
Documents\skills\think-2.spj:

 Thinking Skills
 Observed Variables
 onestep multi routine nonroutn patterns identnum symmetry recogshp
 Covariance    Matrix    from    file    'C:\Documents    and    Settings\pc\My
Documents\skills\Think-2.cov'
 Sample Size = 269
 Relationships
 nonroutn = patterns multi
 multi = patterns onestep identnum
 symmetry
 recogshp
 Path Diagram
 End of Problem

 Sample Size =   269

 Thinking Skills

         Covariance Matrix

               multi   nonroutn    onestep   patterns   identnum
            --------   --------   --------   --------   --------
    multi       1.00
 nonroutn       0.74       1.00
  onestep       0.67       0.56       1.00
 patterns       0.47       0.47       0.38       1.00
 identnum       0.67       0.54       0.56       0.41       1.00

 Thinking Skills
 Number of Iterations =  4
 LISREL Estimates (Maximum Likelihood)

         Structural Equations

    multi  =  0.39*onestep  +  0.16*patterns  +  0.39*identnum,  Errorvar.=  0.41
, R² = 0.59
           (0.048)        (0.044)         (0.049)                     (0.035)
            8.19           3.63            7.90                        11.51

 nonroutn = 0.67*multi + 0.16*patterns, Errorvar.= 0.43  , R² = 0.57
           (0.045)      (0.045)                   (0.037)
            14.75        3.52                      11.51
```

Figure D1 Extract from LISREL output

```
 Reduced Form Equations

    multi = 0.39*onestep + 0.16*patterns + 0.39*identnum, Errorvar.= 0.41,
R² = 0.59
           (0.048)        (0.044)         (0.049)
            8.19           3.63            7.90

 nonroutn = 0.26*onestep + 0.27*patterns + 0.26*identnum, Errorvar.= 0.61,
R² = 0.39
           (0.037)        (0.052)         (0.037)
            7.16           5.16            6.96
          Correlation Matrix of Independent Variables
             onestep   patterns   identnum
            --------   --------   --------
  onestep       1.00
              (0.09)
               11.51

 patterns       0.38       1.00
              (0.07)     (0.09)
                5.74      11.51

 identnum       0.56       0.41       1.00
              (0.07)     (0.07)     (0.09)
                7.92       6.19      11.51

         Covariance Matrix of Latent Variables
               multi   nonroutn    onestep   patterns   identnum
            --------   --------   --------   --------   --------
    multi       1.00
 nonroutn       0.74       1.00
  onestep       0.67       0.51       1.00
 patterns       0.47       0.47       0.38       1.00
 identnum       0.67       0.51       0.56       0.41       1.00
                         Goodness of Fit Statistics
                             Degrees of Freedom = 2
               Minimum Fit Function Chi-Square = 3.72 (P = 0.16)
       Normal Theory Weighted Least Squares Chi-Square = 3.69 (P = 0.16)
                Estimated Non-centrality Parameter (NCP) = 1.69
             90 Percent Confidence Interval for NCP = (0.0 ; 11.33)

                       Minimum Fit Function Value = 0.014
              Population Discrepancy Function Value (F0) = 0.0064
             90 Percent Confidence Interval for F0 = (0.0 ; 0.043)
            Root Mean Square Error of Approximation (RMSEA) = 0.056
            90 Percent Confidence Interval for RMSEA = (0.0 ; 0.15)
              P-Value for Test of Close Fit (RMSEA < 0.05) = 0.34

                 Expected Cross-Validation Index (ECVI) = 0.11
            90 Percent Confidence Interval for ECVI = (0.11 ; 0.15)
                        ECVI for Saturated Model = 0.11
                       ECVI for Independence Model = 3.19

     Chi-Square for Independence Model with 10 Degrees of Freedom = 834.13
                           Independence AIC = 844.13
                               Model AIC = 29.69
                             Saturated AIC = 30.00
                           Independence CAIC = 867.10
                               Model CAIC = 89.42
                             Saturated CAIC = 98.92

                         Normed Fit Index (NFI) = 1.00
                       Non-Normed Fit Index (NNFI) = 0.99
                    Parsimony Normed Fit Index (PNFI) = 0.20
```

Figure D1 Cont’d



```
                     Comparative Fit Index (CFI) = 1.00
                        Incremental Fit Index (IFI) = 1.00
                         Relative Fit Index (RFI) = 0.98

                             Critical N (CN) = 665.19


                     Root Mean Square Residual (RMR) = 0.016
                             Standardized RMR = 0.016
                        Goodness of Fit Index (GFI) = 0.99
                   Adjusted Goodness of Fit Index (AGFI) = 0.96
                  Parsimony Goodness of Fit Index (PGFI) = 0.13

                          Time used:    0.047 Seconds
```

Figure D1 Cont’d

APPENDIX E

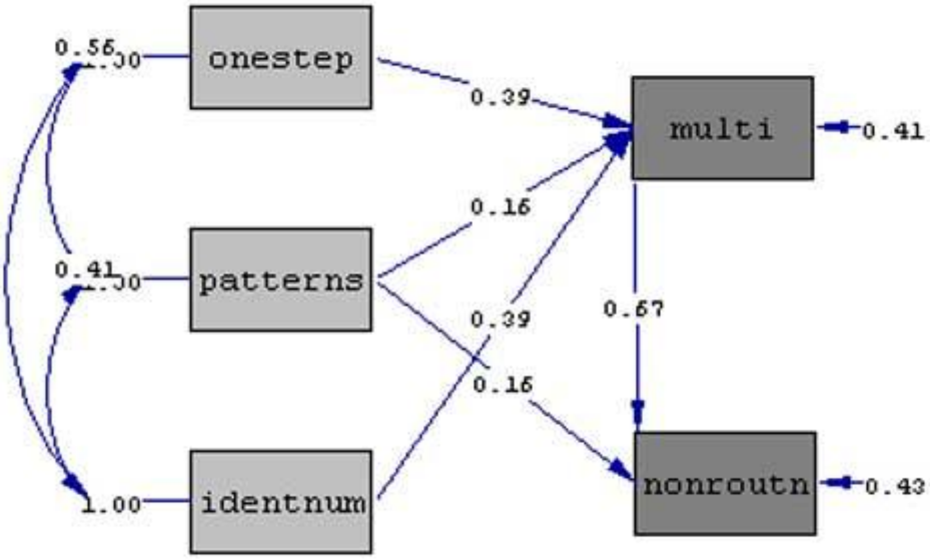


Figure E1 Diagram for Path Analysis

APPENDIX F

Learning Outcomes in 4th Grade Mathematics in the Ministry of National Education Programme

Table F1 Mathematics programme of Ministry of Education for primary school grade 4

ÖĞRENME ALANLARI			
SAYıLAR	GEOMETRI	ÖLÇME	VERI
ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı
Doğal Sayılar	Açı ve Açı Ölçüsü	Uzunlukları Ölçme	Sütun Grafiği
1. 4, 5 ve 6 basamaklı doğal sayıları okur ve yazar. 2. 4, 5 ve 6 basamaklı doğal sayıların bölüklerini ve basamaklarını, basamaklarındaki rakamların basamak değerlerini belirtir. 3. 4, 5 ve 6 basamaklı doğal sayıları çözümler. 4. Doğal sayıları en yakın onluğa veya yüzlüğe yuvarlar. 5. Bir örüntüyü sayılarla ilişkilendirir ve eksik olan bölümü tamamlar. 6. En çok altı basamaklı doğal sayıları sıralar.	1. Açının kenarlarını ve köşesini belirtir. 2. Açıyı isimlendirir ve sembolle gösterir. 3. Açıları, standart olmayan birimlerle ölçerek standart açı ölçme biriminin gerekliliğini açıklar. 4. Açıları standart açı ölçme araçlarıyla ölçerek açıları; dar, dik, geniş ve doğru açı olarak belirler. 5. Ölçüsü verilen bir açıyı çizer. 6. Açıların ölçülerini tahmin eder ve tahminini açıyı ölçerek kontrol eder.	1. Atatürk'ün önderliğinde ölçme birimlerine getirilen yeniliklerin gerekliliğini nedenleriyle açıklar. 2. Standart uzunluk ölçme birimlerinden kilometre ve milimetrenin kullanım alanlarını belirtir. 3 Milimetre-santimetre, santimetre-metre ve metre-kilometre arasındaki ilişkileri açıklar. 4. Belirli uzunlukları farklı uzunluk ölçme birimleriyle ifade eder. 5. Bir uzunluğu en uygun uzunluk ölçme birimiyle tahmin eder ve tahminini ölçme yaparak kontrol eder. 6. Uzunluk ölçme birimlerinin kullanıldığı problemleri çözer ve kurar.	1. Sütun grafiğini oluşturur. 2. Sütun grafiğini yorumlar.
			Olasılık
			1. Olasılık belirten kelimeleri uygun cümlelerde kullanır.

Table F1 Cont'd

ÖĞRENME ALANLARI			
SAYıLAR	GEOMETRI	ÖLÇME	VERI
ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞR. ALANL.
Doğal Sayılarla Toplama İşlemi	**Üçgen, Kare ve Dikdörtgen**	**Çevre**	
1. En çok dört basamaklı doğal sayılarla toplama işlemini yapar. 2. Toplamı en çok dört basamaklı olan iki doğal sayının toplamını tahmin eder ve tahminini işlem sonucu ile karşılaştırır. 3. Toplamları en çok dört basamaklı olacak şekilde en çok dört basamaklı doğal sayıları, 100'ün katlarıyla zihinden toplar. 4. Doğal sayılarla toplama işlemini gerektiren problemleri çözer ve kurar.	1. Üçgen, kare ve dikdörtgeni isimlendirir. 2. Üçgen, kare ve dikdörtgenin kenarlarını isimlendirir. 3. Kare ve dikdörtgenin, kenar ve açı özelliklerini belirler. 4. Köşegeni belirler. 5. Üçgenleri kenar uzunluklarına göre sınıflandırır. 6. Üçgenleri açı ölçülerine göre sınıflandırır. 7. Üçgenin iç açılarının ölçülerinin toplamını belirler. 8. Açıölçer, gönye veya cetvel kullanarak dik üçgen, kare ve dikdörtgeni çizer.	1. Düzlemsel şekillerin çevre uzunluklarını belirler. 2. Kare ve dikdörtgenin çevre uzunlukları ile kenar uzunlukları arasındaki ilişkiyi belirler. 3. Aynı çevre uzunluğuna sahip farklı geometrik şekiller oluşturur. 4. Düzlemsel şekillerin çevre uzunluklarını hesaplamayla ilgili problemleri çözer ve kurar.	
Doğal Sayılarla Çıkarma İşlemi	**Geometrik Cisimler**	**Alan**	
1. En çok dört basamaklı doğal sayılarla çıkarma işlemini yapar. 2. En çok üç basamaklı iki doğal sayının farkını tahmin eder, tahminini işlem sonucu ile karşılaştırır. 3. Üç basamaklı doğal sayılardan 100'ün katı olan doğal sayıları zihinden çıkarır. 4. Doğal sayılarla çıkarma işlemini gerektiren problemleri çözer ve kurar.	1. İzometrik kâğıttaki çizimleri eş küplerle oluşturur.	1. Bir alanı, standart olmayan alan ölçme birimleriyle tahmin eder ve birimleri sayarak tahminini kontrol eder. 2. Düzlemsel bölgelerin alanlarının, bu alanı kaplayan birim karelerin sayısı olduğunu belirler. 3. Karesel ve dikdörtgensel bölgelerin alanlarını birim kareleri kullanarak hesaplar.	
	Simetri	**Zamanı Ölçme**	
	1. Düzlemsel şekillerdeki simetri doğrularını belirler ve çizer. **Örüntü ve Süslemeler** 1. Uygun karesel, dikdörtgensel ve üçgensel bölgeleri kullanarak ve boşluk kalmayacak şekilde döşeyerek süsleme yapar.	1. Dakika ile saniye arasındaki ilişkiyi açıklar. 2. Saat-dakika, dakika-saniye arasındaki dönüşümleri yapar. 3. Yıl-ay-hafta-gün arasındaki ilişkileri açıklar. 4. Zamanı ölçme birimlerinin kullanıldığı problemleri çözer ve kurar.	

Table F1 Cont'd

ÖĞRENME ALANLARI			
SAYıLAR	GEOMETRI	ÖLÇME	VERI
ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞR. ALANL.
Doğal Sayılarla Çarpma İşlemi		**Tartma**	
1. Çarpımı en çok beş basamaklı doğal sayı olacak şekilde iki doğal sayıyla çarpma işlemini yapar. 2. Üç doğal sayı ile yapılan çarpma işleminde sayıların birbirleriyle çarpılma sırasının değişmesinin, sonucu değiştirmediğini gösterir. 3. En çok üç basamaklı doğal sayıları 10, 100 ve 1000'in en çok dokuz katı olan doğal sayılarla kısa yoldan çarpar. 4. En çok üç basamaklı doğal sayıları 10, 100 ve 1000 ile zihinden çarpar. 5. En çok iki basamaklı doğal sayıları 5, 25 ve 50 ile kısa yoldan çarpar. 6. En çok iki basamaklı iki doğal sayının çarpımını tahmin eder ve tahminini işlem sonucu ile karşılaştırır. 7. Doğal sayılarla çarpma işlemini gerektiren problemleri çözer ve kurar.		1. Tonun kullanıldığı yerleri belirtir. 2. Ton-kilogram, kilogram-gram ve gram-miligram arasındaki ilişkileri belirtir. 3. Ton, kilogram, gram ve miligramla ilgili problemleri çözer ve kurar.	
		Sıvıları Ölçme	
		1. Litre ve mililitre arasındaki ilişkiyi belirtir. 2. Litre ve mililitre arasında dönüşümler yapar. 3. Bir kaptaki sıvının miktarını, litre ve mililitre birimleriyle tahmin eder ve ölçme yaparak tahminini kontrol eder. 4. Litre ve mililitre ile ilgili problemleri çözer ve kurar.	

Table F1 Cont'd

ÖĞRENME ALANLARI			
SAYıLAR	GEOMETRI	ÖLÇME	VERI
ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞR. ALANL.
Doğal Sayılarla Bölme İşlemi			
1. Bölme işleminde bölümün basamak sayısını işlem yapmadan belirler.			
2. Üç basamaklı doğal sayıları en çok iki basamaklı doğal sayılara böler.			
3. Son üç basamağı sıfır olan en çok beş basamaklı doğal sayıları 10, 100 ve 1000'e kısa yoldan böler.			
4. Bir bölme işleminin sonucunu tahmin eder ve tahminini işlem sonucu ile karşılaştırır.			
5. İki adımlı işlemleri yapar.			
6. Doğal sayılarla bölme işlemini gerektiren problemleri çözer ve kurar.			
Kesirler			
1. Payı ve paydası en çok iki basamaklı doğal sayı olan kesirleri, kesrin birimlerinden elde ederek isimlendirir.			
2. Payı ve paydası en çok iki basamaklı olan kesirleri sayı doğrusunda gösterir.			
3. Kesirleri karşılaştırır.			
4. Eşit paydalı en çok dört kesri, büyükten küçüğe veya küçükten büyüğe doğru sıralar.			
5. Payları eşit, paydaları birbirinden farklı en çok dört kesri, büyükten küçüğe veya küçükten büyüğe doğru sıralar.			
6. Bir çokluğun belirtilen bir basit kesir kadarını belirler.			

Table F1 Cont'd

ÖĞRENME ALANLARI			
SAYıLAR	GEOMETRI	ÖLÇME	VERI
ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞRENME ALANLARı	ALT ÖĞR. ALANL.
Kesirlerle Toplama İşlemi			
1.Paydaları eşit kesirlerle toplama işlemi yapar.			
Kesirlerle Çıkarma İşlemi			
1. Paydaları eşit kesirlerle çıkarma işlemi yapar. 2. Kesirlerle toplama ve çıkarma işlemlerini gerektiren problemleri çözer ve kurar.			
Ondalık Kesirler			
1. Bir bütün 10 ve 100 eş parçaya bölündüğünde, ortaya çıkan kesrin birimlerinin ondalık kesir olduğunu belirtir. 2. Ondalık kesirleri virgül kullanarak yazar. 3. Ondalık kesirlerin tam kısmını, kesir kısmını ve basamak adlarını belirtir. 4. İki ondalık kesri karşılaştırarak aralarındaki ilişkiyi büyük, küçük veya eşit sembolüyle gösterir.			

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Özkaya Seçil, Selcen
Nationality: Turkish (TC)
Date and Place of Birth: 20 February 1976 , Eskişehir
Marital Status: Married
Phone: +90 216 3263415
email: selcensecil@hotmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Secondary Science & Mathematics Education	2002
BS	METU Mathematics	1998
High School	Eskişehir Anadolu High School	1994

WORK EXPERIENCE

Year	Place	Enrollment
2008- Present	İSTEK Schools	Deputy General Manager
2007-2008	İSTEK Schools	Head of Measurement & Evaluation Dep.
2004-2007	Eyüboğlu Schools	Mathematics Teacher & MYP Coordinator
1999-2004	METU Secondary Science & Mathematics Education	Research Assistant

FOREIGN LANGUAGES

Advanced English, Intermediate French

PUBLICATIONS

1. Özkaya Seçil, S. (2002). "Investigation of tenth grade students' problem solving strategies in geometry". Unpublished Master Thesis. METU: Ankara

2. Özkaya Seçil, S. ve Bulut, S (2002) "Öğrencilerin Geometri Problemleri Çözerken Kullandıkları Farklı Yöntemler " V. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi, ODTÜ, Ankara