MIXED METHODS RESEARCH AT THE INTERSECTION OF MATHEMATICS TEACHERS’ ADOPTION OF CURRICULUM CHANGE AND ASSESSMENT OF STUDENTS’ READINESS LEVEL THROUGH EDUCATIONAL NEUROSCIENCE METHODS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BENGİ BİRGİLİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF EDUCATIONAL SCIENCES, CURRICULUM AND INSTRUCTION

OCTOBER 2022

Approval of the thesis:

MIXED METHODS RESEARCH AT THE INTERSECTION OF MATHEMATICS TEACHERS' ADOPTION OF CURRICULUM CHANGE AND ASSESSMENT OF STUDENTS' READINESS LEVEL THROUGH EDUCATIONAL NEUROSCIENCE METHODS

submitted by **BENGİ BİRGİLİ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy of Educational Sciences, Curriculum and Instruction, the Graduate School of Social Sciences of Middle East Technical University** by,

Prof. Dr. Sadettin KİRAZCI
Dean
Graduate School of Social Sciences ____________

Prof. Dr. Hanife AKAR
Head of Department
Department of Educational Sciences ____________

Prof. Dr. Hanife AKAR
Supervisor
Department of Educational Sciences ____________

Assist. Prof. Dr. Tuna ÇAKAR
Co-Supervisor
MEF University
Department of Computer Engineering ____________

Examining Committee Members:

Prof. Dr. R. Nükhet ÇIKRIKÇI (Head of the Examining Committee)
Istanbul Aydın University
Department of Primary Education ____________

Prof. Dr. Hanife AKAR (Supervisor)
Middle East Technical University
Department of Educational Sciences ____________

Assoc. Prof. Dr. Gülçin TAN-ŞİŞMAN
Hacettepe University
Department of Educational Sciences ____________

Assist. Prof. Dr. Şerife SEVİNÇ
Middle East Technical University
Department of Mathematics and Science Education ____________

Assist. Prof. Dr. Elif ÖZTÜRK
Middle East Technical University
Department of Educational Sciences ____________

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: Bengi BİRGİLİ

Signature:

ABSTRACT

MIXED METHODS RESEARCH AT THE INTERSECTION OF MATHEMATICS TEACHERS' ADOPTION OF CURRICULUM CHANGE AND ASSESSMENT OF STUDENTS' READINESS LEVEL THROUGH EDUCATIONAL NEUROSCIENCE METHODS

BİRGİLİ, Bengi
Ph.D., The Department of Educational Sciences, Curriculum and Instruction
Supervisor: Prof. Dr. Hanife AKAR
Co-supervisor: Assist. Prof. Dr. Tuna ÇAKAR

October 2022, 391 pages

The main purpose of this study was to investigate the quality of middle school mathematics teachers' instructional and assessment processes after curriculum policy change through an ecological approach and the quality of the students' responses and reactions towards different item types via employing a multimodal mixed methods concurrent dominant status design. Three major phases were processed to achieve this aim. Firstly, Phase 1 includes the practices that reveal the quality of teachers' authentic teacher-made items (N = 380) via document analysis. Secondly, Phase 2 relied on a quantitative survey where a sample of mathematics teachers (n = 350) affiliated in public and private schools in an upper middle class district of Istanbul, Turkey were administered the TMMESP-Questionnaire, in which items were theoretically constructed, to identify teachers' teaching method and assessment preferences. Thirdly, Phase 3 relied on multimodal phase in which fifth grade students (n = 32) were administered open-ended and multiple-choice mathematics examination items to evaluate their reactions and responses to the items with respect to active use of metacognitive subskills and affective processes via eye-tracking and biometric tools. Finally, a local deep data model was theoretically sketched based on these multimodal data. The results indicate that the teachers tended to prepare mathematics items mostly relying on traditional objective testing. Most of them lacked the skills needed to enact collaborative instruction and assessment processes. More advantageously, when the students were exposed to higher-order items, they could more reliably reflect their metacognitive skills and affective process. The quality of the instruction, assessment and items made by the teachers are not yet sufficiently prepared for the students' readiness of future competencies. Among implications, teachers should help their students to acquire metacognitive skills while responding to open-ended items.

Keywords: Mathematics curriculum change, In-class assessment, Metacognition, Mixed methods research, Educational Neuroscience

ÖZ

MATEMATİK ÖĞRETMENLERİNİN ÖĞRETİM PROGRAMI DEĞİŞİKLİĞİNİ BENİMSEMESİ VE ÖĞRENCİLERİN HAZIR BULUNUŞLUK DÜZEYLERİNİN DEĞERLENDİRİLMESİNİN KESİŞİMİNDE EĞİTİMSEL NÖROBİLİM YÖNTEMLERİYLE KARMA YÖNTEM ARAŞTIRMASI

BİRGİLİ Bengi
Doktora, Eğitim Bilimleri, Eğitim Programları ve Öğretim Bölümü
Tez Yöneticisi: Prof. Dr. Hanife AKAR
Ortak Tez Yöneticisi: Dr. Öğr. Üyesi Tuna ÇAKAR

Ekim 2022, 391 sayfa

Araştırmanın temel amacı, ortaokul matematik öğretmenlerinin eğitim politikası değişikliği sonrası öğretim ve ölçme-değerlendirme süreçlerinin kalitesini ekolojik bir yaklaşımla ve öğrencilerin farklı madde türlerine çoklubiçimli karma yöntem eşzamanlı baskın durum deseni ile verdiği tepkilerinin kalitesini incelemektir. Bu amaca ulaşmak için üç aşama işlendi. Aşama 1, doküman analizi yoluyla öğretmenlerin otantik sınıfiçi öğretmen yapımı soru maddelerinin (N = 380) kalitesini ortaya çıkarmaya yönelik uygulamaları içerir. İkinci olarak, Aşama 2, Türkiye'nin üst orta sosyo ekonomik seviyeye sahip bir ilçesindeki devlet ve özel okullarda görev yapan ortaokul matematik öğretmenlerinden (n = 350) oluşan bir örneklem ile onların öğretim yöntemlerini belirlemek üzere maddelerin teorik olarak oluşturulduğu ÖYÖDT-Anketinin uygulandığı anket aşamasıdır. Aşama 3, beşinci sınıf öğrencileri (n = 32) ile onların üstbilişsel ve duyuşsal süreçlerinin aktif kullanımı yoluyla olarak tepkilerini değerlendirmek için açık uçlu ve çoktan seçmeli matematik sınav maddelerinin uygulandığı göz izleme ve biyometrik araçlar içeren çoklubiçimli aşamadır. Son olarak, bu çoklu biçimli verilere dayalı olarak ülkemize özgü bir derin veri modeline teorik olarak ulaşılmıştır. Sonuçlar, öğretmenler matematik sorularını çoğunlukla geleneksel nesnel testlere dayalı olarak hazırlama eğilimindedirler. Öğretmenlerin çoğu, işbirlikli öğretim yöntemi ve değerlendirme süreçlerini uygulayabilmek için gereken temel becerilerden yoksundur. Ortaokul öğrencilerinin üst düzey bilişsel becerileri ölçen soru türleriyle karşılaştıklarında üstbilişsel becerilerini ve duyuşsal süreçlerini daha güvenilir bir şekilde yansıtabildiği bulunmuştur. Öğretmenler tarafından yapılan öğretim, sınıfiçi değerlendirme ve soru maddelerinin kalitesi, öğrencilerini gelecekteki yeterliliklere yeterli düzeyde hazırlayamadığı tespit edilmiştir. Öğrencilerin açık uçlu maddeleri cevaplarken üstbilişsel becerileri kullanmalarının teşvik edilmesi önerilmektedir.

Anahtar Kelimeler: Matematik öğretim programı değişimi, Sınıfiçi değerlendirme, Üstbiliş, Karma yöntem araştırması, Eğitimsel nörobilim

To My Family

ACKNOWLEDGMENTS

I did my best to design this long-term dissertation with an innovative perspective of its time and interdisciplinary approach. I hope it will be inspiring for dear readers and graduate students who have just started to study on their dissertations. Be hopeful. Of course, your dissertation will end one day; always remember "A good dissertation is a done dissertation. A great dissertation is a published dissertation. A perfect dissertation is neither."

I would like to express my sincere thanks and gratitude to my supervisor Prof. Dr. Hanife Akar for her guidance, support, and encouragement throughout intense process of this study. The knowledge and skills she taught to me was invaluable. Her belief in me has kept my motivation high during this process. I also would like to express my sincere thanks to my co-supervisor Asst. Prof. Dr. Tuna Çakar for his unconventional perspective and for his engineering intelligence and skills that he used while guiding me to create the multimodal phase. In addition, I am very thankful to my thesis monitoring and defense committee for their extensive suggestions. I greatly appreciate mathematics teachers and administrators in public and private schools in Istanbul who kindly helped me in the data collection process.

My dear friends have been a great supporter for me. It was challenging process, but they were always at the end of a phone line to answer my questions and give their expert opinions. Thank you for supporting me and allowing me to keep my joy. I am very grateful for Asst. Prof. Türker Toker who assisted the initial data collection process. I would like to express my sincere gratitude to Assoc. Prof. R. Didem Taylan for her help in the main data collection and all the academics of MEF University in the department of mathematics and science education. Especially, I am very obliged to Assoc. Prof. Utkun Aydın for her valuable mentorship in my academic progress, her motivation, and inspiration. I am very thankful Asst. Prof. Ömer Demir who always share innovative research ideas with me, and encourage me to reach this study's destination line. I am thankful to doctoral candidate Sercan Çelik for his valuable feedback on how to use

English as a foreign language. I am thankful to Dr. Mustafa Baştürk for facilitating the preparation of my dissertation for publication. And I am very thankful to Dr. İpek Saralar-Aras from the Ministry of National Education, for her expert opinion.

My sweet friend, it was a great chance to work with you: my dear friend, psychological counselor Gizem Öztemur, you have always encouraged me! I am so lucky to be working with you! And Ragibe, my eternal friend, thank you for your presence! I am very grateful to Ömer Cihan Karabulut, who know the value of science best, for all his motivation and support. So glad I have you! Finally, I would also like to commemorate Dr. Sevinç Tunalı, my dear [deceased] colleague, who listened to the first sparks of this study with her great enthusiasm, she was as excited as I was. And I would like to thank my dear teachers and friends that I may not remember right now. This process was very challenging yet enjoyable with you.

My dear family; I am very grateful to my brother Buğra Birgili, my mother Cemile Birgili and my father Turhan Birgili. I felt you behind me at every moment and you were my biggest supporter.

As closing my words, a big congratulations to myself...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABIDE	Project of Monitoring and Evaluating Academic Skills
AI	Artificial Intelligence
AERA	American Educational Research Association
APA	American Psychological Association
CAT	Computerized Adaptive Testing
CEC	Commission of the European Communities
HETE	Higher Education Transition Examination
HoTs	Higher Order Thinking Skills
ILE	Intelligent Learning Environment
LoTs	Lower Order Thinking Skills
MES	Measurement and Evaluation Strategy
MoNE	The Ministry of National Education
NCME	National Council on Measurement in Education
NCTM	National Council of Teachers of Mathematics
OECD	The Organization for Economic Cooperation and Development
PISA	Programme for International Student Assessment
TIMSS	Trends in International Mathematics and Science Study
TM	Teaching Method

CHAPTER 1

INTRODUCTION

1.1. Background of the Study

Schools prepare children for life and inevitably for the jobs of the future. If we think of school as the first system where education is systematically given, we assume that all schools are in this way equal and fair. They can prepare children for jobs that have not yet been created, the technological inventions that have yet to be discovered, and problems that have not been predicted (Rios et al., 2020; The Organization for Economic Cooperation and Development [OECD], 2018a, 2019). However, major factors that affect education in this competitive world are economic, political, social, and technological trends. Under these environmental, economic, and social factors, schools should help every learner to develop as a whole person and at the same time consider their well-being. Not only should they provide new opportunities for human advancement, but they should also provide formal or informal learning environments to equip them with the skills they need to accomplish in this new world. More importantly, they should create a teaching and learning atmosphere which can help this generation confidently utilize these skills and demonstrate competencies (Rotherham & Willingham, 2010; Van Laar et al., 2020; Voogt & Roblin, 2010).

Entering school, children for the first time experience an environment other than that of their families, and they must adjust by becoming part of a social group with new rules. Predictably, there are many differences between children as they attend school for the first time. Some are ready for school while others are not and their prior learning and the skills they come with differ, as does their basic levels of skill acquisition, interests and attitudes. Having different social levels and socio-economic status means that they initially try to adapt to the school environment in different ways. Their skills related to learning objectives differ. Likewise, their degree of desire for learning differs (APPG, 2017; Hansen, Heckman & Mullen, 2003; Nunn, 2014).

Measurement, evaluation, and assessment are integral parts of a curriculum since they are systematically used to determine whether students have acquired quantitative and qualitative aspects of each behavior students are intended to acquire (Gelbal, 2013; Özcelik, 1998). There should be diversity in assessing, measuring and evaluating students based on how the teaching method differs. Not only via paper-pencil measures but also via portfolios or project-based assessment types, there should be diversity in the dynamics of the measurement and evaluation process (Barootchi & Keshavarz, 2002; Grammatikopoulos, 2012; Tiekstra et al., 2016). It is important to have a variety of formats as this offers several advantages. Assessing students' comments, questioning skills, perspectives, and metacognitive skills, as well as evaluating their content knowledge and skills, enables educators to evaluate their students in a multi-dimensional way. At the same time, in a globalizing world, raising individuals and generations who can express themselves, solve problems, see things from a different perspective, and have developed social skills are one of the significant developmental steps of a literate society.

Although curriculum and instruction concepts are examined separately, they are interdependent. While the curriculum is thought of as a planned learning experience, instruction is thought of as methodology, instructional strategy, methods, and techniques, in short, as a teaching role (Mayer & Alexander, 2011; Rose, 2004). What is important about addressing such differences is that this can create a successful educational environment, namely, fruitful teaching and learning processes that enable children to grow as happy individuals (Turkey's 2023 Education Vision, 2018). This is the biggest role of schools. The goal of the educational system is to provide the most suitable school environment, class climate, teaching and learning processes (Goodlad, 1994) for the development of healthy and happy individuals. In this way, children become individuals who understand the world and confidently develop and compete within it (Nyman & Kaikkonen, 2013).

One of the most vital elements of schooling is teachers (Darling-Hammond & Lieberman, 2012; Ticha & Hospesova, 2006). Teaching is a significant profession (AACTE, 2010; Darling-Hammond, 1999, 2006; Özcan, 2011, 2013). At the same time, teaching is a polymorphous concept (Rogers, 1969; Peters, 1970). Every individual within this system adapts the facilitation of change and learning. On the subject of shaping a child's personality, especially in the context of Turkey, the main responsibility is believed to fall on teachers who are conducive to helping students gain new knowledge and develop their skills, building on family values. The quality, experience and qualifications of teachers are related to the quality of their teaching (Darling-Hammond & Lieberman, 2012). In teaching, quality can be achieved by creating learning opportunities to achieve the learning and teaching goals and by consistently aligning these processes with student needs (considering their pre-requisite knowledge, skills and attitudes). The quality of a school's equipment and qualifications and the ability of teachers to design their own tools for the benefit of the learning process can be used to benefit students. For a qualified teaching process, colleges of education and in-service training opportunities should provide professional development to teachers in line with their needs (Darling-Hammond & Lieberman, 2012; Darling-Hammond, 2017; Darling-Hammond & Oaks, 2019). The better educated and qualified teachers become, the better they will educate their students to become young scientists who interact and compete with the world and keep an open-minded viewpoint.

One aim of schools is to prepare children for the next level of schooling. To do this, it is helpful to review how the goals of education have been achieved through the determination of learning objectives, feedback from the process, and whether the students have gained a skill in such a way so as to carry it to the higher grade. This is called “the curriculum” (Bobbitt, 1918; Goodlad, 1968;

Ornstein & Hunkins, 2004; Tyler, 1949). Based on the relationship between curriculum and instruction, it is necessary to prepare appropriate programs and instruction according to students' developmental characteristics. The curriculum should guide teachers. Instruction is part of the learning process. Although the instructional approach is a process representing the distribution system of the content dimension, the source of the teaching and its components as well as the elements of the teaching method cover the basic structures related to this process. These two concepts must be compatible with each other in terms of approach and design. The most important element that gives feedback on how these structures are tested and the way in which students reach their goals is the assessment process of an instruction. The relationship between curriculum and instruction is so intimate that instruction may exist independently of the curriculum but would not be operationalized regarding direct purpose to maximize students' learning and development (Flake, 2017; Hilebowitsh, 2005; Random, 2016; Yates, 2013).

In Turkey in 2017, the implementation of a change in the large-scale examination system was announced. Recently, the Turkish Ministry of National Education (MoNE) has expressed the intention of enhancing higher order thinking skills by including open-ended questions and, thus, has initiated converting the testing style to a more thought-provoking one in evaluating students' progress (Berberoglu, 2009; Çıkrıkçı-Demirtaşlı, 2010; TEDMEM, 2013, 2017; MoNE, 2017a). The MoNE has become the center of attention due to rumors that the Transition from Basic Education to Secondary Education (TEOG) high-stakes examination system may be immediately abolished in Turkey. Following an announcement that the new examination system replacing TEOG would be coordinated under the title of National Monitoring System ("Milli İzleme Sistemi", ODSGM, 2017), it was thought that the MoNE had implied that open-ended questions would gradually be implemented as a change to the question format based on the recent publication of sample questions with short answers. The MoNE stated that the examinations to be developed in the context of the Project on Monitoring, Research and Development of Assessment and Evaluation Applications [ABIDE] would initially be implemented for 5th grade students. Current studies have been ongoing with local workshops and discussions under MoNE (MoNE, 2017b). Many other countries have similar concerns and are attempting to develop effective knowledge construction and assessment systems. For example, in Finland, the main goal of assessment is to monitor the progress of students, improve their learning, and give feedback on their progress in a supportive way rather than force students to race for being the world’s strongest performers on the standardized high-stakes testing (Hendrickson, 2017). In the Netherlands, the examination system is twofold; students take school examinations, and they take a central examination (OECD, 2012). This means that the high-stakes testing approaches changed from students’ qualification to school accountability. As for the school exams, the Netherlands has issued many innovative item types in high-stakes testing. In many educational settings it is believed that the only effective way to understand whether knowledge construction has occurred is through understanding individuals' interaction with question types. However, there should be consistency between utilizing scientific truths revealed from item types and the governing political power of centralized curricula which identifies the competencies.

The gap between traditional approaches and community needs is growing (Jorgensen et al., 2017). It is fundamental to update the knowledge, skills and dispositions of the communities in line with 21st century competencies. The primary institution that shapes a society is the school education system. To improve the productivity of the education system, technological advances need to be incorporated. The children of this new age are called digital natives and are believed to be acquainted with new technological developments. Technological advancements are developing every day in areas including communication, digital competences, literacy, numeracy, and transversal competences. New skills such as learning how to learn can support the development of autonomous and collective individuals. These competencies will be encouraged through new educational pedagogies and a flexible curriculum created to assist the all-around growth of welfare, active involvement in community, and the innovative skills of creativity and invention (Commission of the European Communities [CEC], 2008; Drew & Mackie, 2011; Yalçın, 2018). These developments show us that students' cognitive skills, as well as non-cognitive skills (affective and psychomotor skills), are important and becoming increasingly so as the acquisition of these skills during school years is decisive for a country's future competitiveness. These skills can be developed in the school environment and it is important that students learn these skills before they are put into practice. Business world and educational policy leaders want schools to develop students' skills in collaboration, communication, learning to learn, problem solving, technology literacy, and self-management in their learners; in short, 21st century skills (National Research Council, 2012). In order to assess these skills, one type of measurement is insufficient. Rating scales, performance evaluations, computer simulations, and portfolios are used, as well as different types of question format (multiple choice, computer based and open-ended items) (Soland et al., 2013; Yalçın, 2018).

In general, there are two types of examination items: closed-ended and open-ended. Students' writing abilities, conceptual knowledge, and high level thinking skills, such as assessing, analyzing, and problem solving, are routinely assessed via open-ended items (Reilly et al., 2014). Moreover, their assessment takes time and is likely to be influenced by the assessors' subjectivity. To address these concerns, automatic assessment approaches for shorter narrative replies have been explored for more than a decade. These strategies work effectively for items with a single or small number of correct responses. Some open-ended items, on the other hand, require students to articulate their rationale, making it impossible to establish a reference response. A review of past studies shows that earlier academic work concentrated on methods such as Latent Sematic Analysis (used mostly till 2017) and Natural Language Processing in general, and recent improvements have an increase in interest in *Deep Learning* applications for automatic grading of open-ended items. The framework for evaluating open-ended question responses automatically is still in its infancy. The majority of scientific research has been carried out over the previous five years, and there is a lot of interest in trying out new ways. Examining these studies by country shows that those with the highest number of studies on automatic grading in open-ended questions between 2004-2021 were the China, Germany, India, the United States of America, the United Kingdom and so on (Casalino et al., 2021). Innovative item types, generally preferred in large-scale assessments in Europe (e.g., Pearson), are computer-delivered items interacting with test takers. Such items offer many benefits to students and school systems compared to traditional, paper-and-pencil items (Strain-Seymour et al., 2009). Numerous studies have looked into the increased construct validity, effectiveness, and capability to evaluate a wide range of skills and competences of new items. Item types are either traditional item formats (e.g., close-ended, open-ended, fill in the blanks, multiple-choice) or innovative item formats (e.g., computer-based) (Sireci & Zenisky, 2011). However, in Turkey, most questions consist of multiple-choice formats in the initial stage, whereas open-ended items would also be used. In other words, the changes to question format which has been discussed since 2014 have not completely been implemented. One of the foremost reasons for this is that educators are required to be fully qualified to apply open-ended items in examinations. Such types of questions allow students to make use of metacognitive, critical, creative, and analytical thinking skills. However, the experts evaluating such open-ended questions are required to be well-educated, their evaluation should be independent and objective, and attention should be given to keep the identity of students and evaluators confidential (Schleicher, 2017). Many global research initiatives, researchers, and international organizations still have a common perception that Turkey's education system is not up to global standards (OECD, 2018a, 2018b; Schleicher, 2017).

The attitudes, content knowledge, and skills that will be provided to students in a planned manner are organized by the curriculum as part of a holistic strategy (Doll, 1996; Ellis, 2004; Miller, 2007; Ornstein & Hunkins, 2004). In this sense, the preparation of curricula in Turkey and the process of developing or changing the education system should be considered in conjunction with other sub-components. One of these components is the development of educational policies in a way that includes quality, equality, effectiveness, and universal values based on national and social values (Ho, 2018; Hopman et al., 2014). Policies determine the aims, principles and strategies to be achieved in Turkey. There are intersections of theory, policy, and practice in the field of education (Apple, 2018). In line with these policies and regulations, the MoNE began to renew its curricula in 2018. This change has been shaped according to the needs of the education system and following the developments in education and teaching both in Turkey and abroad. It is reflected and updated in the newly developed curricula by taking into consideration the principles of transparency, scientific enquiry, and participation. The mathematical competencies are described as "the development and application of mathematical thinking to solve a range of daily life problems" (MoNE, 2018 p. 6). The competencies include process, activity, and knowledge built on solid arithmetic skills, thinking skills, (logical and spatial thinking) and the ability and eagerness to use mathematical modes of presentation (constructs, formulas, graphs, models, and tables) to varying degrees. As stated in Turkey's Education Vision 2023:

> Indeed, "curriculum" is one of the most contested words in the conceptual map of contemporary education. Under pressure from standardized tests, the curriculum ceases to be a means and instead emerges as an end. This tension builds up due to serious discrepancies between our nation's schools. Our vision for the future converts the curriculum from a collection of information to a source of skills, and then to positive ways of living. It also trains teachers to relieve the pressure caused by tests. In a system with well- trained teachers, a curriculum framework alone would suffice (p.11).

The Ministry of National Education prepared the draft curricula (TTKB, 2017) for the education system within the framework of basic skills and competences, and can therefore be said to have been prepared according to the European Qualifications Framework (2008), National Education Qualifications Framework, and Turkey's Qualifications Framework (2016), which entered into force in the Official Gazette. However, in terms of the continuity of the curricula development process, the draft curricula differ from the previous ones. The MoNE highlighted that curricula, which were believed to contribute to a happy and successful life for our students, were prepared based on basic core values (e.g., values education) and competencies (e.g., mathematical literacy, self-awareness, basic life competencies). For mathematical literacy, it is not sufficient for students to only write events and situations as mathematical formulas. They are also required to explain, justify, and interpret the facts. The goals of the curricula include learning to use not only cognitive but also self-awareness processes. However, the reasons for the changes made in the dimensions of philosophy, aim, teaching process, and measurement and evaluation were not clearly understood and have been criticized by some studies researching teachers' views (Çolakoğlu, 2018).

The MoNE states that the mathematics curricula valid until 2017-2018 academic years are aimed at ensuring middle school students (grade 5-8) take on national values, as well as mathematical skills. However, it is essential to consider the spiral curricula learning outcomes, in which the number of outcomes decreases. In the 5th grade curriculum, the total number of learning outcomes (n = 56) and course hours (180 hours) remain the same, and, likewise, learning and sub-learning areas have not changed. Yet the number of competencies (e.g., digital competency) and skills planned to be acquired by 5th grade students has increased (İlhan & Aslaner, 2019). While both curricula aim to develop basic skills and competencies, taking responsibility for one's own learning comes to the fore in the program valid since 2018. In particular, it aims to make students learn to learn and gain self-awareness. Therefore, students should be given the opportunity to think about their own thinking, to follow their own learning needs and processes.

The literature shows that the role teachers play in various aspects of the curriculum development process is still relatively unknown. Current data on teachers' experiences are limited (Altun & Akkaya, 2014; Çetin & Ünsal, 2018; Çolakoğlu, 2018; Yayla & Yayla, 2018). Also, concrete evidence of what kind of changes teachers are able to adapt to while implementing in-class innovative educational approaches, measurement and evaluation processes are open to improvement (Shuilleabhain & Sery, 2018). Studies have often demonstrated possible challenges of application of a curriculum into teaching environment due to the complexities involved (O'Shea & Leavy, 2013) and the way teachers are challenged to change their practices (Fetters et al., 2002). In addition, Turkish students show very low achievement in worldwide international test results, such as the Programme for International Student Assessment (PISA), organized by the Organization for Economic Co-operation and Development (OECD), and the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), both coordinated by the International Association for the Evaluation of Educational Achievement (IEA) (OECD, 2016, 2018b; PISA, 2015). One of the reasons behind these failures is that Turkish students are not accustomed to the restriction of fully open-ended examination items, and thinking about processes and concepts cannot be reflected in how they understand (ABIDE, 2016; Çolakoğlu, 2018).

The Turkish education system is much like a land of exams. There are several national central examinations. Teachers are not offered many alternative assessments due to the central examination so that they are not willing to use alternative assessment techniques in class examinations. Even from the lower grade level of primary school, testing and test books are a bitter reality and an integral mode of the teaching process. Although it is emphasized in the curriculum that teachers can use various other measurement and evaluation approaches than those recommended, the pressure of exams makes it inevitable to turn to objective testing. Worldwide research evidence shows that the systems that utilize automatic, intelligent and innovative items, measure the knowledge and skills of students, and make meaningful inferences from multimodal measurements of student performance are not created as a whole, but by fragmented and scattered small-scale studies. The acceleration of those studies on all these intelligent systems depends on teachers updating their professional skills according to the curricula, developing an effective in-class assessment system, and preparing students for their achievements with quality items, and on students' ability to express themselves metacognitively (i.e., by cognitive strategy and self-checking) and affectively (i.e., worry and effort). There are many measurement, evaluation and assessment processes that necessitate students practice different item types, drill and practice what they have learned so far, solve problems, and transfer their current mathematical knowledge and skills. Problem-solving is a dynamic process that interlinks cognitive, metacognitive, and affective dimensions of learning in a multidimensional way (Mangaroska et al., 2021). In this process, students' readiness for the lesson, their prerequisite knowledge of the subject and their cognitive levels can be tested. During problem solving, students use both cognitive and metacognitive thinking processes and affective processes in problem solving in-class assessments or examinations. The role of cognitive strategy and self-checking (i.e., metacognitive subskills) as well as worry and effort (i.e., affective process) in in-class assessment and problem-solving processes become passionate predictors to see "how well they did" and "what they could have done better".

Considering the attention on the dichotomy of policy and practice, the aim of the study was primarily to reveal the dichotomy between educational policy change and curriculum enactment in terms of teaching method and assessment, and students' readiness in relation to some skills and competencies highlighted in the curriculum. The study contributes to the educational field by providing research on the quality of middle school mathematics teachers' adoption of curriculum change and by assessing middle school students' readiness levels through educational neuroscience methods. The findings show that mathematics teachers' adoption of curriculum change significantly serves students' needs.

1.2. Purpose of the Study

Following the new educational policies (e.g., Turkey's Education Vision 2023) of the Turkish education system, the first aim of this study is to investigate the current quality of mathematics teachers' authentic teacher-made items, their in-class teaching method preferences and measurement-evaluation strategy preferences after the curriculum has been implemented in schools. A further aim is to investigate middle school students' reactions and responses to open-ended items in terms of usage of metacognitive skills and affective processes when provided with an opportunity to respond to quality item types. As an overall perspective, the current study aims to present a theoretical deep data model that is able to measure and evaluate the responses of middle school students to open-ended items in terms of metacognitive skills and affective processes. In line with these purposes, this study attempts to assess the following: The extent to which the middle school mathematics curriculum (dated 2015 academic terms) is compatible with the proposed assessment procedures within the curriculum in preparing middle-school students for learning outcomes; the teaching methods and measurement-evaluation strategies used by middle school mathematics teachers in the classroom after the maths curriculum change (Official Gazette, Article 2726, Dated 19/01/2018) compared to the previous curriculum (valid until the end of 2016-2017 academic terms (Official Gazette, Article 55, Dated 28/07/2015); middle school students reflection on their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels in their responses to different item types; the difference between the amount of reflection of students' metacognitive skill levels to their responses to multiple-choice and open-ended items; the neuro/biomarkers needed to measure students' responses to open-ended items to evaluate their metacognitive (cognitive strategy and self-checking) and affective processes (worry and effort) through *deep data modeling*.

1.3. Research Questions

Five main research questions guided this study:

1) To what extent is the enacted middle school mathematics curriculum (dated 2015 academic terms) compatible with the proposed assessment procedures within the curriculum in preparing middle-school students for learning outcomes?

2) Do the teaching methods and measurement-evaluation strategies used by the middle school mathematics teachers in the classroom after the maths curriculum change (in 2018) compare to those used for the previous curriculum (dated 2015)?

2.1) Do middle school mathematics teachers' change their teaching methods with the implementation of new curriculum policy?

2.2) Do middle school mathematics teachers' change their measurement and evaluation strategies with the implementation of new curriculum policy?

3) How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types? Is there a significant difference between the amount of reflection of students' metacognitive skill levels on their responses to multiple-choice and open-ended items?

3.1) Is there a significant difference between the amount of reflection of students' cognitive strategy skill levels on their responses to multiple-choice and open-ended items?

3.2) Is there a significant difference between the amount of reflection of students' self-checking skill levels on their responses to multiple-choice and open-ended items?

4) What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?

4.1) Is there a significant difference between the amount of reflection of students' affective process levels on their responses to multiple-choice and open-ended items?

4.2) Is there a significant difference between the amount of reflection of students' worry levels on their responses to multiple-choice and open-ended items?

4.3) Is there a significant difference between the amount of reflection of students' effort levels on their responses to multiple-choice and open-ended items?

4.4) Do total time on task and gaze shifts have an impact on predicting reading or not while responding to items?

5) What neuro/biomarkers are needed to measure students' responses to open-ended items to evaluate their metacognitive (cognitive strategy and self-checking) and affective processes (worry and effort) through deep data modeling?

1.4. Significance of the Study

Education has emerged as a global policy concern (European Strategy Framework, 2015; Meyer et al., 2017; Tiven et al., 2018). The evaluation issues comprise very diverse perspectives including models of evaluation, analysis of evaluation systems, types of evidence, types of assessment, the source of data, quality criteria, and existing paradigms among evaluation theories. Nevertheless, no reform in other areas would be sustainable without radical education reform, without raising generations that are productive, entrepreneurial, practical, visionary, respectful of ethical values, and without reference to international and national values, reason and science.

In formal education all over the world, technology and human-technology interaction demands on the educational policy agenda (Tuomi, 2018). It is applied in as many educational and research settings as possible. When a new promising advancements and technological tools emerge, these technologies may appear to open up entirely new possibilities for tackling both old and new issues more quickly. In addition, student learning outcomes are still vital to successful global digital exchange. The research and diagnosis of students' learning outcomes and their performance in qualified testing environments should be the basis of both design and evaluation of new curricula and instructional renovations in particular and quality of education in general. One of the important steps of globalization is to follow the world not only in terms of philosophy and approach, but also in the use of technology and methodology, to follow current developments and to test them in national, local and regional adaptive studies. This idea is also supported with the Maastricht Global Education Declaration, highlighting “the methodology of Global Education focuses on supporting active learning and encouraging reflection with active participation of learners and educators” (p. 4) and “develop, or where developed improve strategies for raising and assuring the quality of Global Education” (p. 5) (European Strategy Framework, 2015).

Comprehensively, the literature on evaluating students’ deep-thinking processes (see page 64 for detail) during restricted open-ended items is sparse and does not offer field standards or a conceptual framework for measuring and evaluating students’ thinking processes. One significance of this study is that it attempts to present a new and expansive framework for how to use multimodal tools (e.g., eye-tracking) and to measure and evaluate students’ metacognition and affective process. While teachers' in-depth knowledge of students’ thinking helps them to structure their teaching processes in accordance with student needs and readiness level, it will also assist with making in-class teaching and reliable in-class assessment compatible with the curriculum. Its multimodal evaluation will also guide teacher to make the right decisions. With the advancement of technology, multimodals would facilitate teachers’ work. In addition, as captured in *The*

Blind Men and the Elephant (John Godfrey Saxe, 1872), teachers should be aware of the different insights in the students' knowledge and skills. Thus, multimodal tools can enable us to look at students and the learning environment from a range of angles. They can look at their own classroom practices, not from a simple point of view, a simple direction and perspective, but from different points of the picture with more accurate tools. Multimodal tools help students not only by capturing and interpreting their cognitive and metacognitive knowledge but also their affective process, their area of interest and their thinking levels at the same time. It is of critical importance for students to be measured in a valid and reliable way and to compare their metacognitive skills and affective processes. Therefore, it may be thought the different insights feed off each other. Teachers can manage processes more consistently and achieve more efficiency.

Another significance is the exploration of the policy-practice dichotomy. Exploring teachers' adoption of the curriculum change into their daily mathematics classrooms while teaching and evaluating students, and trying to capture the students' metacognitive skills and affective processes through a multi-model tools in understanding the relation between classroom practice and student readiness level is thought to increase the quality of classroom practices and in-class assessment. Measuring and evaluating students deeply in their problem-solving process is considered important in terms of teachers knowing these and preparing students for new systems that would change. The study tries to show how students' metacognitive skills can be ready to deal with higher order thinking skills in relation to their age levels. Teachers need to invest higher expectations in their students and teach and assess them accordingly so that they can be competitive in the global context. It is also possible that the study reveals how much policy change can be reflected in classroom practices, teachers' enactment ability from instructional and assessment points, and students' readiness level related to affect and metacognition for educational systems that would change in the future.

Examination of master’s theses and dissertations in Turkey shows that, as of 2010, there was a limitation of PhD dissertations compared to master’s thesis regarding multimodal tools such as eye-tracking. Most of the studies have been processed in the fields of computers and instructional technology, English language teaching, primary education (i.e., CEIT, ELT, childhood education) in particular rather than the field of educational sciences in general. This suggests that there is a need to look at the phenomenon through the eyes of educational scientists and mathematicians. There are more studies that utilize mixed-method research designs than qualitative research designs, and the theses focused on higher education and secondary school levels. The studies mostly selected students as participants, while teachers, university students and faculty members least frequently were employed less frequently. The subject focus of these studies was mathematics, geometry, social sciences, English language, and Turkish language, but they recently did not look at the usability of eye-tracking tools in the department of educational sciences, which points to the need for this study.

There are few interdisciplinary studies between educational sciences and other disciplines in Europe, the UK, the USA, Canada, Asia (Catrysse et al., 2018; Fleming & Frith, 2014; Li & Wang, 2020). Interdisciplinary studies (i.e., related to collaborative skills among 21st century skills) in Turkey, conducted between educational sciences and cognitive sciences [a.k.a. Neuroeducation] are currently limited (Nunez et al., 2020). In this study, the assessment-evaluation process and question types were focused. How an educational program is implemented in the classroom is seen to be as important as how it measures the student from different aspects. Feedback is provided to both the individual and the educational process on the student's knowledge, skills and attitudes. While authentic evaluation processes gain importance in national and international literature, the use of restricted open-ended items, which is considered an alternative to multiple-choice in classroom and large-scale exams, gives more explanatory information about the students’ solution steps and processes. Our work on students' mathematical problem solving, investigating the metacognitive and affective processes, is considered pioneering work in Turkey (Birgili, 2014; Birgili & Kiraz, 2017; Koyuncu, 2017). This innovative study points to a methodological as well as research gap in the literature, as there seems to be a limited number of empirical studies that attempt to determine middle school students' metacognition and affective process through a multimodal approach. Taken together, the study is reexplored through the lens of mathematics teachers' in-class authentic examinations, their teaching method and measurement-evaluation preferences under an ecological approach, and then students' metacognitive and affective process experimentation are adapted with a multimodal investigation. The aim is that this data uniquely forms the theoretical basis and input for the deep data models that are intended to be accessed at the end of the study and that evaluate students' responses to restricted open-ended items. Its contribution to the field and society may well be unique in terms of an interdisciplinary approach.

1.5. Definition of Important Terms

The following definitions of the key terms are employed thoughout the paper:

Curriculum Change is to make the curriculum different in some aspects, to alter its philosophy by way of its aims, goals and objectives, to review the content, to revise its instructional methods, and to re-think its evaluation procedures (Priestley et al., 2015)

Assessment Change is to make adaptations in measurement, evaluation and assessment processes and types, to meet key objectives by including adoption of a new technology, changing a key process or restructuring the organization (Carless & Zhou, 2015).

Authentic Teacher-Made Items is teacher-made examination items constructed on the basis of a carefully planned table of specifications, generally with the same type of items and providing clear directions to the students, and used for formative assessment in classrooms (McMillan et al., 2002).

Metacognition is defined as the cognition of cognition (Flavell, 1979) and expanded as the method by which people reflect on their own thoughts in order to create solutions to subject-based problems. This process has been broken down into several sub-categories such as awareness, cognitive strategy, monitoring, planning, self-checking and so on (O'Neil & Abedi, 1996; O'Neil & Brown, 1998). Among several metacognitive subskills, **cognitive strategy** is a goal-directed and consciously regulated that supports or facilitates performance as learners create internal processes that provide them the ability to carry out desired skills. It can be domain-specific as well as general. In other words, whereas domain-specific strategies are effective in a particular context and can be applied in other situations, broad strategies indicate problem-solving procedures spanning a wide range of situations (Mcewen et al., 2009); **Self-checking** is "self-monitoring one's performance when engaging in a task" (Shaughnessy et al., 2008 p. 117). For instance, while responding an item, students' findings their own error on the solution, asking questions to stay on track may activate their self-checking subskills.

Affect is a "psychological term which states the experience of feelings and emotions. It is a generic term for emotions and other mental states that have the quality of pleasant-unpleasant, such as feelings, moods, motives, or aspects of the self, e.g., self-esteem" (Forgas, 1994 p. 57). In addition, *affect* is "a physiological reaction of one's to testing experiences such as fear, physical discomfort or nervousness" (Lufi et al., 2004, p. 177). As a cognitive phenomenon intimately link to affect, **worry** relates to one's self-evaluation of the suitability of the assessment type to measure the one's cognitive capabilities and the cognitive aspects of their anxiety (Awang-Hashim et al., 2010). It is a cognitive distress related to testing (Lufi et al., 2004)**.** While **effort,** relates to "the willingness to keep trying and the mental strength or willingness to persist to complete a task" (Awang-Hashim et al., 2010 p. 343).

Eye-Tracking is to track eye position and movement to follow visual attention during performance. It helps researchers to measure one's eye movements in real time and know where one's focus is at any given time. For a variety of purposes, academics and researchers use data on eye movements and fixations, such as to evaluate one's attenton, contrast group behavior, track responses to stimuli visually, and more. Researchers can use eye tracking to discover what grabs people's attention right away, what they ignore, what order they notice things in, and how some things compare to others. It can be screen-based, with glasses and VR (Holmqvist, & Andersson, 2017; IMOTIONS, 2021). **Gaze shift** is the realignment of the line of one's vision so as to bring the image of a new object of interest to the central retina where receptor density and hence visual resolution are the highest (Binder et al., 2009). **Area of Interest (AoI),** also referred to as an AOI, is a tool to select regions of a displayed stimulus and extracting metrics specifically for those regions. It defines the area by which other metrics are determined, despite not being a measure in and of itself (IMOTIONS, 2021). **Total Time on Task** is the total performance time spent by a student during assessment or testing experience, including instructional time as well as time spent for performance such as studying and completing assignments (e.g., group projects, reading, writing, and thinking aloud process).

CHAPTER 2

LITERATURE REVIEW

This chapter introduces the philosophy behind study as the teacher agency in curriculum and assessment practices change after policy change through ecological approach and then overview of various educational assessment concerns globally, and analysis of in-class assessment. Then, it gives a brief summary of the relevant literature about the analysis of authentic teacher-made items related to some cognitive taxonomies. This review of literature enriched by cognitive diagnostic models in general and student monitoring systems in Turkey, and then the importance of measuring students' metacognition and affective processes through eye-tracking tools and other biomarkers. The literature encompasses imperative need of artificial intelligence systems in education, the most recent developments on how deep data models are used in the students' learning process, and especially the effect of innovative item types on the learning process and students' success, and the essential needs to evaluate the knowledge and skills of our country's children with very comprehensive intelligent systems are presented.

2.1. Teachers' Role to be Change Agents in Curriculum Development Process: An Ecological Approach

Agency is a quality of engagement of actors, namely teachers in educational contexts. It is not related to what the teachers have but what they do and achieve (Biesta & Tedder, 2007; Leijen et al., 2020). The ecological understanding of agency (Emirbayer & Mische, 1998), which has been used as a framework, has three dimensions. They are 1) iterational, 2) projective and 3) practical-evaluative by considering the influences of past experiences, engagement with present and so orientations toward the future.

Teachers as being the agent of curriculum change (Priestley, Biesta & Robinson, 2015) have become an accepted phenomenon which was believed to close the knowledge gap between purpose and practice of curriculum making process. In a more flexible way, it puts the teachers at the heart of curriculum development process and are seen as independent developers of a curriculum (Çelik & Kasapoğlu, 2014; Nation & Macalister, 2010; Priestley & Biesta, 2013; Rahimi & Alavi, 2017). This approach has been aimed to prepare teachers with appropriate knowledge content and pedagogy after a curriculum change in a country. Rather than relying on specification of content knowledge, teachers have an opportunity to study active forms of pedagogy, facilitator of learning, being aware of accountability and professional developers of a curriculum.

There are three dimensions focused on the ecological approach: individual, structural and cultural. Accordingly, not only autonomy, but also other factors affect how teachers decide during application of curriculum change in the classroom. For instance, a really agentic teacher would like to do adaptations in their first-hand practices, yet structural or multitude of factors may not allow it (Bascia et al., 2014; Wallace & Priestley, 2017). The ecological approach is not only a theoretical but also a methodological framework. These dimensions can be shaped by observations and following other qualitative approaches. How many objectives are achieved and how the decision-making process works are critical questions which have been answered in the ecological approach. "How effective are the beliefs of teachers in individual factors in the decision-making process? How do the pressures of the centralized system, structurally, reflect on the teachers? How do the ideas, habits guide teachers in a school?" are thought provoking questions which may dig up teachers' practical experiences in the light of ecological approach. Agency can be achieved under particular ecologies. Therefore, teachers can develop their agency relying on the collaboration of the competences and ecological conditions (Biesta & Teddler, 2007; Priestley & Drew, 2016).

Priestley and his colleagues (2012) studied with two high school science teachers regarding their curriculum enactment. They discovered that the teachers carry out the curriculum in different ways depending on their belief system. One teacher, for instance, implemented lecturing because he thought of his students as passive consumers of knowledge. His teaching method in instructional approach was quite didactic. He carefully followed the documents outlining the official curricular policy. The study, however, also discovered that the second teacher, another participant in the study, acted out a very different course experience because she thought life science education should encourage creative thinking. So, she used many ways and her teaching method was very promising which engaged her students with course content. It was evidenced that the teachers generate positive beliefs about new curriculum reform with the help of their professional development. Yet they keep their newer beliefs more minor, while keeping their core belief sets.

Mesker et al. (2018) used ecological approach in relation with professional development teacher candidates in the international arena. Socio-cultural differences have been the main concern that may somehow challenge during action or interaction in teaching and learning. According to their understanding of ecological perspective, Mesker and his colleagues believed that ecological approach considers past, present, and future dimensions. So, the impact of the personal histories, experiences, and interpretations of their student teachers on professional development has to be better understood by teacher educators. Through a qualitative approach, they used case study design to investigate the boundaries noticed by eight student teachers while they were doing their international internships for their professional learning. They were aimed to explore where this learning experiences originate. Four data sources were used during the study. These include each student teacher's unique history, two individual inventories that include perspectives on sociocultural differences, and individual interviews. According to research findings, they found 15 main boundaries. For instance, boundary 1 and 14 is about “Discontinuity is related to existing pedagogical approaches and student teachers perceive boundaries in teaching aspects" (p. 7) whereas boundary 7, 8 and 9 is about "Discontinuity is related to specific school type or culture. Students perceive a boundary in their teaching practice, which is the result of the school culture or school type of the school where they are interning" (p. 7). They concluded that students' teachers' professional experiences abroad have multiple dimensions including also cultural or personal aspects. Therefore, it was suggested that teacher educators should not merely concentrate on gaps in teaching knowledge or abilities.

As Zhang and Shen (2012) stressed, Teachers who are focused on their professional growth and have clear plans for it can gain more "agency" than those who do not. Teachers who are experts in their field can demonstrate greater agentic behavior than those who are novice. Structures and cultures of their professional life should be considered to deduce from their practices.

2.2. Bloom's Revised Taxonomy and TIMSS Framework

Bloom's Revised Taxonomy has a framework which measures cognitive process and knowledge dimension of examinees. The original Bloom's Taxonomy was widely accepted to be used as a framework by many educators since 1956 (Bloom et al., 1956). This was used to define educational objectives in line with any curriculum so that educators better and more concretely assess the learners. According to Amer (2006), the taxonomy is so rigid in terms of its principles which points out simple to complex, having cumulative hierarchy. In this cumulative hierarchy, each level from knowledge to evaluation has organized to growing complexity and each level was assumed to consist of all behaviors of the less complex level while increasing in levels. It was designed as unidimensional which aimed to define objectives of what students intend to learn and to be able to do objective based evaluation, and categorize those objective statements written by educators in teaching and learning process. This taxonomy became the rule for mastery learning (Bloom, 1956, 1964, 1968, 1971, 1974, 1976, 1985; Bloom & Rakow, 1969; Guskey, 2010) because mastery of one low level category became a requirement to achieve mastery in the next high-level category. These categories were: *Knowledge*, *Comprehension*, *Application*, *Analysis*, *Synthesis,* and *Evaluation*.

However, once the research and theoretical studies conducted on Bloom's original taxonomy have been enlarged, it was criticized for measuring a generalized knowledge domain. Several drawbacks and practical limitations have been discussed and disputed (Amer, 2006; Bloom, 1987; Forehand, 2005; Krathwohl, 2002). A team of testing and assessment experts, curriculum and instruction researchers, and cognitive psychologists made the decision to update the first classification (Anderson et al., 2001, as cited in Amer, 2006). Since they believed that original taxonomy did not allow users to assume overlapping between categories; that Knowledge level sometimes can be more complex than certain skills for Analysis or Evaluation; that Evaluation is not more complicated process than Synthesis; Synthesis involves Evaluation (Amer, 2006). Recent developments in Constructivism (Slavin, 2003), Metacognition (Zimmerman, 1998), Self-regulated learning, Self-awareness, Self-monitoring triggered to reshape and redefine behavioral approach to measure knowledge and skills in learning and assessment. Therefore, the dimensions of cognitive demands became significant (Forehand, 2005).

Secondly, this taxonomy did not adopt learners' levels of higher order thinking, which seem to be significant for enhancing the quality of instruction and assessment. Therefore, Anderson, Krathwohl and their colleagues created the two-dimensional revised Bloom's Taxonomy. 1) Knowledge Dimension included: *Factual Knowledge*, *Conceptual Knowledge*, *Procedural Knowledge*, and *Metacognitive Knowledge*. 2) Cognitive Process Dimension included: *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, *Create* (Anderson, 2005; Anderson et al., 2001; Anderson & Krathwohl, 2001; Krathwohl, 2002). Since then, the Revised Bloom's Taxonomy has been broadly used through an international perspective in mathematics (Ernest, 1999; Radmehr & Drake, 2017, 2018, 2019), science (e.g., Lee et al., 2017; Lee et al., 2015; Tan, 2019) and other educational research studies (Cheong, 2018; Porter, 2006; Reynolds, 2019; Uymaz & Çalışkan, 2019)

TIMSS Framework, on the other hand, is known as the most prolonged, large scale international assessment of mathematics and science education in the world (Mullis & Martin, 2019). Mullis and Martin (2019) explain the cognitive domains:

> The first domain, *knowing*, covers the facts, concepts, and procedures students need to know, while the second, *applying*, covers the capability of students to apply knowledge and conceptual understanding to solve contextual problems or response items. The third domain, *reasoning*, goes beyond the solution of routine problems to include unfamiliar situations, complex contexts, and multistep problems.

The content domains and cognitive domains in TIMSS assessment framework (Table 2.1.) are depicted as:

Table 2. 1. Content and Cognitive Domains in TIMSS Framework

Grade Level (Content Domain)		Grade Level (Cognitive Domain)		
4th Grade	8th Grade	4th Grade		8th Grade
Number: 50%	Number: 30%	Knowing:	40%	35%
-	Algebra: 30%	Applying:	40%	40%
Measurement and Geometry: 30%	Geometry: 20%	Reasoning:	20%	25%
Data: 20%	Data and Probability: 20%			

Note. From TIMSS 2019 Assessment Framework: TIMSS 2019 Mathematics Framework (pp. 13-25), by Mary Lindquist et al., 2019, Boston College: TIMSS & PIRLS International Study Center, Lynch School of Education. Copyright 2019 by the International Association for the Evaluation of Educational Achievement (IEA). Reprinted with permission.

As stated by TIMSS Mathematics Framework, Restricted Use Items for Grade 4 (pp. 105- 110) (see. Table 2.2.):

Table 2. 2. Sample Items in TIMSS Framework

Sample Item #	Label	Content Domain	Topic Area	Cognitive Domain
M041291	Subtract 428 - 176	Number	Whole numbers	Knowing
M051140	Expression for Jenny's age	Number	Expressions, Simple Equations, and Relationships	Applying
M041298	Which rectangle is 1/4 shaded	Number	Fractions and Decimals	Knowing
M041124	Use the rule to complete the table	Number	Expressions, Simple Equations, and Relationships	Applying
M051093	Perimeter of the given shape	Geometric shapes and measures	Two- and Three-dimensional Shapes	Reasoning
M041264	What is the area of the triangle	Geometric shapes and measures	Two- and Three-dimensional Shapes	Applying
M041191	Peter's height	Data Display	Reading, Interpreting, and Representing	Applying
M051077	What students did after school	Data Display	Reading, Interpreting, and Representing	Applying

Last but not least, when the authentic teacher-made examination items inspected related to measuring the amount of LoTs and HoTs, McREL International (2017) proposed the trend of item analysis findings from an international perspective and stated that teachers prepared the items mostly based on LoTs such as remembering (11.5%), understanding (27.6%), applying (32.1%); whereas in a limited portion of HoTs such as analyzing (16.2%), evaluating (5.1%) and creating (4.7%). In addition, the fact that the teachers mostly tend to ask LoT questions than HoTs were in support of earlier studies on the estimation that 70 to 80 % of all items require the simple recall of facts (i.e., LoT items), while only 20 to 30 % require the HoT items which triggers thought processes of clarifying, generalizing, and making inferences from the items (Haynes, 1935, as cited in Khan & Inamullah, 2011; Gall, 1970). The findings of Gall (1970) stated that teachers tended to prepare items having 60% recall, 20% procedural and 20% thought provoking. Those of Lee (2015) revealed that 79% of the total items asked were lower-order questions whereas only 5% were targeted to measure higher-order items.

2.3. Teachers' Role on the Curriculum Development Process in Turkey

Curriculum development process for curriculum change is a team work. Although developed under the National of the Board of Education and sent to all schools in Turkey, national curricula have been developed according to top down approach (Bascia et al., 2014; Rahimi & Alavi, 2017), experts from the Ministries several academics, educators and teachers are labor intensive and efforts in this process. The draft curricula were evaluated with great precision in a workshop attended by 360 people who participated in working groups including academicians and teachers (TTKB, 2017, p.13). However, 2017-2018 National Education Statistics (2018) depicts that a total of 1 030 130 teachers currently work in formal education. It is a matter for teachers in Turkish schools who do not take part in curriculum development in a way that they can assimilate these changes in their schools. How they use all changes effectively within the classroom currently has a gap.

The program development process starts with a short idea of education, namely philosophy. The approach to be developed according to this philosophy is determined. The most appropriate program development model to be shaped by the theory and practice of education is selected and appropriate steps are followed. Then, it continues with the determination of the general and specific aims of the program. In order to reach the target learners, learning objectives specific as aims to the course and the subject, in other words learning outcomes are determined. At the beginning of the process, if the needs of the target learners are not defined, they should be subjected to a needs analysis process.

After this step, the program design and development process begin. At this stage, the content of the subject to be taught according to the knowledge, skills and affective aspects of the learners is selected. The tools, content, material, technology etc. to be used in the teaching process is detected. The measurement tools are prepared to determine the process and how much the student has learned in order to give feedback to the educator. At this stage the program is developed. The implementation process of the program continues with the formative and summative evaluation steps. The program is evaluated. Experts of other branches from whom get benefit in this education process, for example; education technologist, education economist, psychologist, etc. In the intellectual background of this large picture, it is worth to be highlighted that it is actually associated with the social, cultural, political and economic reality of its society (Bascia, et al.2014; Bümen, et al., 2014; Demirel, 2012; Kerkez, 2018; Nohl & Somel, 2016; Oliva, 2009; Ornstein & Hunkins, 2004).

Teachers, who are the most significant cornerstone of the education system and how active they are in the program development process, have always been open for research (Bascia et al., 2014; Oliva, 2009; Priestley et al., 2012; Stein, et al., 2007). We assume that they are the one who know the subject matter knowledge in the best way and teach it to students. In performing their profession, they are also those whom we expect to choose the right method to teach the students according to the curricula and the best way to teach them in their teaching-learning process (Tan-Şişman & Karsantık, 2017). Thus, why they should be deprived of a curriculum development process when it is expected to have such an active role is still a debatable and hot topic in the context of Turkish education system.

One of the interesting studies which payed attention to teachers' role in the curriculum development process in Turkey was conducted by Kerkez (2018). Kerkez's thesis explored, as followed by a phenomenological research design, the teachers' opinions about the theory and practice of curriculum development process in vocational and technical education in Turkey. 19 teachers who had attended in the curriculum development commission within the years of 2012-2016 were conducted semi-structured interviews while they were also working at the state vocational and technical high schools in Ankara. The results investigated that renewal of a curriculum is a systematic process and this process necessitates cooperative group work from different stakeholders. Some teachers were found to have inadequate knowledge of curriculum even though they know the relation between curriculum and instructional activities. They have a general awareness of the traditional program development process. Besides, it has been determined that the participant teachers have some incomplete and incorrect knowledge about the core elements of a curriculum. These deficiencies, specifically, were found to be about aims, content, and instructional aspects. This study stressed that even though some teachers have a general overview of the curriculum and gain experience in this process, they may still have incorrect knowledge in some elements. Hence, this may cause the inability to accurately reflect the objectives within the classroom. It is also important to question the views about these elements in the enacted curriculum. What they know and to what extent they can adopt the enacted curriculum should be enquired in the study together with its basic elements.

2.3.1. Curriculum changes reflected on practice of education

A recent study found in the extent literature about the effect of curriculum changes on the practice of education was conducted by Bümen and others (2014). They also agreed and stated that there is a centralized education system in Turkey. While implementing the curriculum (enacted curriculum) developed by MoNE, teachers in Turkish schools make many adaptations for their classes depending on its region and changes based on their own preferences or on their students. On the other hand, some scientific studies remarked that the renewal of curricula does not assure to be the renewal of classroom practices (Atila, 2012; Bümen et al., 2014; Öztürk, 2012; Yaşar, 2012). From this perspective, Bümenand the colleagues (2014) showed us the importance of opening out how teachers adapt the renewed curriculum as enacted in their classroom environment as well as their authority and degree of freedom. They found out the effect of factors on the curriculum as teacher qualities and training, program properties, and institutional features. In addition to the literature, other variables worth to investigate are students' characteristics such as learning differences, achievements (Kaya et al.2012), high stake tests (Güneş & Baki, 2011), centralized education system itself, also regional, social, economic, and cultural differences (Goodlad, 2007). Most importantly, they drew attention to the results of some significant studies on teachers' awareness and knowledge of the curriculum development process. Since 2014, the results have revealed that teachers in Turkey are not master sufficiently with "the curriculum" (Tekbıyık & Akdeniz, 2008) because they see the curriculum as a list of subjects or yearly plan rather than a lively and active process. They cannot perform the roles given to them. They think that the stakeholders create unrealistic programs and are generally developed by academics and the lesson process (Kaya et al.2012). Unfortunately, they have also serious difficulty in assessment (Yaşar, 2012). Following, there is a great need for their professional development to prepare them for in-class practices after curriculum changes. They struggle to fulfill the curricular mandates (Greene, 1995; Priestley et al., 2012; Pinar, 2012). All in all, these factors affect teachers' perception and belief system so that their application during in-class teaching and assessment cannot be matched with the aims of the curriculum. Thus, this triggers a gap between the official and enacted curriculum.

Foreign language teachers' perceptions on a top-down national curriculum change were investigated by Rahimi and Alavi (2017)'s study. They studied 127 teachers who had been enacting this curriculum change process for three years through a quantitative design. The teachers' perceptions were measured by a perception change survey with three focus; administrative support, new curriculum, and teacher practices. According to quantitative analysis, the findings showed that the experiences of teachers accounted for 9% percent of the variance in the results of curriculum change. The researchers also supported the results with a qualitative approach by structurally interviewing 10 experienced and 8 novice teachers in the field. The findings enlarged with triangulated data revealed the teachers were optimistic on the new syllabus but they thought their agency was overlooked in this curriculum change. On the other hand, the concerns of experienced teachers were regarding the practical side of this curriculum change, for instance, time issue and deficiency of audiovisual tools. In brief, this study expressed the need of looking at the actual practical impact of curricula changes. Experiences and perceptions of teachers, and how they are adopted in teaching are the main issues of a curriculum change.

Foremost among these, Burul's master thesis (2018) emphasized the importance of examining the commitment of teachers to the official curriculum. Because it guides us to measure the harmony between the official program and the enacted one. Although they are thought to be similar, it is likely that there will be differences in practice according to school, environmental conditions and needs of students. In this context, it is important to focus on the reality of the implementation process (Gerstner & Finney, 2013). In particular, the degree to which the program is applied and how it is applied should always be investigated. Teacher characteristics are one of the most important factors (Bümen et al., 2014).

Within this framework, Burul (2018) studied the association between the teachers' preferences of curriculum design and their curriculum fidelity. He selected 319 primary and secondary school teachers conveniently in two districts of Balıkesir, Turkey in the academic year of 2015-2016. The method was quantitative survey and causal comparison study that measured teachers' curriculum design approach preferences by using "Curriculum Design Orientations Preference Scale of Teachers." "Curriculum Fidelity Scale" was also developed to determine the fidelity of the teachers toward the enacted curriculum. According to the results, teachers preferred more student-centered design approaches. There was no statistically significant difference between the group of teachers except their level of teaching. It meant that those working in primary school seem to prefer subject-centered design compared to those in secondary school. Primary school teachers tended to share more ideas and collaborate with their colleagues about curriculum than secondary school teachers.

Eriş and Kılıçoğlu (2019) conveniently selected 222 teacher candidates from the department of classroom teaching in order to investigate their competencies in the curriculum development process. They used a scale to determine their curriculum development skills which then analyzed statistically thorough SPSS to reach the descriptive and inferential results. This quantitative designed study revealed some variables such as gender, class and university were found statistically significant in context of curriculum development competencies. The findings showed that the teacher candidates were poor at preparing course plans, "using teaching approaches to raise the students' capacity, and differentiating the teaching and learning process for students' individual differences. In addition, they showed moderate competency regarding development of a curriculum and measurement and evaluation in educational context. On the contrary the participants had good competence in founding the relationship among educational program items such as purpose, content, instruction, evaluation, and determining the goals at different domains as Cognitive, Affective, Psychomotor. Through the curriculum change in 2017 in Turkey, very few studies have been conducted to gather teachers' opinions on draft curricula. For instance, Özcan and Düzgünoğlu (2017) searched for science teachers' views on draft curricula through a qualitative research paradigm. They conducted individual interviews with ten science teachers who had been selected by criterion sampling method. The qualitative data obtained were analyzed and presented descriptively. The findings described that the common point of teachers highlighting the draft curriculum period has some weaknesses. The stakeholders could not prevent the process from negative consequences such as subject selection, adding or removing some subjects from the draft curricula. It was seen that there was no change in the evaluation of the current curriculum, its educational philosophy, teaching methods and techniques and measurement-evaluation. However, when we look at the findings from who were the practitioners of the curriculum, it was understood that they have not exchanged their views with their groups during teacher meetings in the schools. However, considering the content knowledge of science teachers across the country is not the same, it is seen that the achievements on application process of draft curricula within this form do not provide enough guidance to science teachers. One of the findings of the study remarked that without fully implementing the curriculum, the teachers and authorities cannot explore the basic deficiencies. They also suggested that authorities should identify the deficiencies with opinions and suggestions by taking all stakeholders' opinions. Thus, it can be inferred that searching for in-class lived experiences of teachers are highly significant to reflect more on the enacted curriculum.

Likewise, Shawer (2017) shared that the more teachers make changes and adaptations in line with their students' needs, the higher the performance of the students due to the received and taught curriculum would become different. Otherwise, teachers may be inclined to follow "teaching to test" by taking large-scale assessments as their principal referencing point. What materials they can use in instructional implementation may be manipulated by the preparation for and expected findings of these tests (Boardman & Woodruff, 2004).

2.4. Assessment as a Global Concern

Assessment can provide a wide range of benefits to educators, teachers, parents, policy makers, researchers. Through assessment, we can get information whether an educational program is effective; descriptive or functional models about psychological attributes; a rationale to support for international education; measure and evaluate students' not only knowledge but also skills and dispositions, understand learning and developmental processes, and even diagnose problems (Sternberger et al., 2009). When planned and implemented effectively, assessment can examine factors such as generally content knowledge; mostly cognitive function, cognitive ability, executive functions, and fair extended motivation, attitudes, values. However, if educators or policy makers are aimed to evaluate all aspects of education delivered to students, a good global, international or national, in-class assessment should include multiple tools and methods integrated with the enacted curriculum. We can examine not only direct evidence of student learning (via tests, papers, portfolios) but also indirect evidence (via students' perceptions of their learning) for their school performance. However, by applying robust curriculum development and evaluation systems with outcome-based evaluation strategies, educators or policy makers can infer from limited use of assessment results. In global education, the education systems require beyond traditional movements of inputs and outputs to embrace the complex impact of international learning on students' happiness and academic success (Deardorff, 2007).

Studies have often evidenced possible challenges of application of a curriculum into the teaching environment due to its complexities (O'Shea & Leavy, 2013) and teachers become challenged of how they should change their practices (Fetters et al., 2002). In addition, Turkish students show very low achievement between participatory countries in international test results such as Programme for International Student Assessment [PISA], International Mathematics and Science Study [TIMSS] (OECD, 2016, 2018b; PISA, 2015) even though in Turkish students' level was in the last place below average, called the third group; now it's very close to the average in 2018 results (Ministry of National Education [MoNE] PISA Pre-report, 2019). Some reasons behind these failures is that our students are not accustomed to solve the restricted open-ended items during in-class assessment or authentic teacher-made exams; thinking about processes and concepts cannot be reflected in how they understand (ABIDE, 2016; Çolakoğlu, 2018).

Therefore, in recent years, the differences between students' academic achievement, test-taking efforts in the classroom and their success and motivation in international exams are striking (Andrews et al., 2014; Eklöf et al., 2014; Hopfenbeck & Kjærnsli, 2016; Michaelides et al., 2020). Thus, it has emerged that students should be measured in multiple methods and techniques. It is essential for improvements in the teaching and learning process to follow authentically what students think and express while solving questions (i.e., process of responding). Then, educators, teachers, policy makers can make evidence-based decisions with multiple data in decision-making processes. In the re-development of the national curriculum and in-service education for teachers, data-based decisions and scientific findings can be shared and more valid guidance can be made. Studies have emerged showing the need to investigate in-depth what teachers teach but contrary what students understand. However, there is little evidence (Unal et al., 2020; Azevedo & Aleven, 2013; Van Gog & Jarodzka, 2013) in the related literature including interdisciplinary studies in Turkey.

2.4.1. Measurement and evaluation system in Turkey

The national context in Turkey is familiar with sudden changes in terms of educational reforms and innovations. Changes to curriculum, measurement, evaluation and assessment including examination systems have often become a hot topic due to such sudden changes. Since 2014, phrases such as “open-ended questions are considered for SBS”, and “a new system to replace SBS” have become a current issue to point out the changes considered to be made in large-scale examination system (Haberport, 2019; see Özkaya 2021 for curriculum comparisons). Another matter of debate on such changes having press coverage is that open-ended items would replace multiple-choice in the TEOG examination system applied during transition from middle school to high school. Even the possibility of including open-ended items in large-scale examination systems has caused a dilemma among educators, academicians, and instructors. Discussions are still ongoing about possible consequences of implementing open-ended items in large-scale examinations, the quality of open-ended questions, evaluation of the examination results, and whether such evaluation would be objective.

As a matter of fact, multiple-choice and open-ended items are inherently used to measure various forms of knowledge and skills categorized under different taxonomy (i.e., comprehension, application, analysis etc.). Differential effects of these two common question formats have been examined from various perspectives in previous studies (Bilgeç, 2016; Cohen et al., 2007; Eren, 2015; Gültekin & Demirtaşlı, 2012; Karadeniz, 2016; Koyuncu, 2017; O'Neil & Brown, 1998; Stepankova & Emanovsky, 2011). Since 2014, such differential effects have been a thesis subject within the context of what kind of metacognitive skills could be more efficiently measured. Also, these differential effects were investigated which affective behaviors could be triggered by these question formats whereas efforts have been made to address potential dilemmas and questions by tangible outcomes in the light of scientific methods (Birgili, 2014; Eren, 2015; Koyuncu, 2017). Hence the current research investigated students' and teachers' in-class experiences regarding open-ended question formats because since 2014 there has been a limited amount of research, which explored this phenomenon with a holistic perspective. The scope of research was very narrow and revealed what has been going on in-classes in the schools after the question format dilemma in the Turkish examination system.

The results from various international and national large-scale assessments of students indicated that the tendency of assessing students' performance toward the use of MC. The Turkish examination system has been totally based on MC at various levels of schooling as in OKS, SBS, TEOG, University Entrance Exam, ALES etc. Although most students demonstrated acceptable performance in MC, many students' performance has been low. In PISA, examination results clearly showed that Turkish students' mathematics and science performances were lower than those of many other participating countries (OECD, 2012, 2014, 2016; PISA, 2012, 2015). According to PISA 2015 results, Turkey was ranked nearly last in the mathematics, science and reading sections. Among 72 countries, China, Japan, South Korea, Singapore, Denmark, Sweden etc. participated in the PISA 2015. Turkey was ranked 50 with 420 points in mathematics (X_{mean} = 461), 54 with 425 points in science (X_{mean} = 465), and 50 with 428 points in reading (X_{mean} = 460). Several studies (e.g., Berberoğlu & Kalender, 2005; Eraslan, 2009; PISA, 2015; TUSIAD, 2013) inspected possible underlying reasons of our student's failure. One of the significant findings from the national exposition of The PISA 2015 international report indicated that learning activities in the school are more effective than learning activities outside the school in the students' learning. Then, the better the quality of teaching and the variety of teaching methods in the classroom, the more successful the learning activities during in-class teaching (Kyriakides et al., 2018). This brings the positive aspects of an effective teaching method into the classroom. It is seen that Turkish students did not successfully express themselves while solving OE question formats. It seems to us that this failure of Turkish students in international examinations is due to the fact that teaching and assessment methods can not prepare our students for these examinations. Each experience of students during in-class teaching should not only prepare for recall level of information, in other words "knowing" level but also for "applying" and "reasoning" level. They should be exposed to in-class experiences so that they are able to show their level of knowledge in international arenas where the depth of knowledge changes.

The curriculum development process and the change of measurement and evaluation decisions in Turkey have mostly failed to include teachers' perceptions in decision-making. Because the Turkish National Board of Education is the only certifying authority in curriculum development and decision-making process. On 13th of January, 2017, Ministry of National Education announced that 53 different draft curricula for primary, secondary, high school level had been published in the website of the Ministry. Respective people from the Ministry investigated the perceptions regarding new curricula from every segment of the society so that the new curriculum could be applicable as of 2017-2018 in all Turkish schools. MoNE organized several national workshops to get feedback and views from stakeholders, teachers, academicians, educators, non-governmental organizations (MEB, 2017; TTKB, 2017). After the announcement, a great number of public, private institutions and non-governmental organizations (AÇEV, 2017; Ankara University, 2017; Bogazici University, 2017; ERG, 2017; METU, 2017) published their own views and suggestions. These organizations all agreed that there is a lack of process regarding who the stakeholders were in the new curriculum development process. These views and suggestions indicated that they could not evaluate because of the fact that they were not sure about the effectiveness and qualification of the group who had worked on the subject-matter and curriculum development process. However, it is an international belief that teachers should be at the heart of developing curriculum policy into practice (Priestley, 2017). Hence a critically significant suggestion was that the curricula should be reviewed by teachers who are qualified in the subject matter, experts, scientists and other stakeholders. In addition, it was frequently reported that guidance for teachers was not sufficient about how the instructional strategies can be arranged by them and how general goals, learning outcomes can be evaluated. Teachers have recently been reinvented as change agents and professional curriculum and assessment experts through an ecological framework in global education policy. (Alvunger, 2018; Biesta et al., 2017; Priestley et al., 2015; Priestley, 2016, 2017).

One of the important points in achieving the targeted level of success with the draft program is that the program can be implemented effectively. One of the most important responsibilities at this stage is to allow teachers to be well-trained on various assessment types, programmed, to have a say in the program or measurement preparation process. Teachers should be given the right and the duty to be involved in the decision-making process in order for the outputs of the program to be measured effectively. However, in Turkey, where the test system is so coherent that the use of the curricula is controversial, it remains unanswered how and to what extent teachers are involved in the process of the change in national curriculum-making (Priestley, 2010). In this context and in light of the discussions, the study is focus on the extent to which the teachers play a decisive role in the assessment and evaluation system in Turkey.

Intelligent systems that are thought to facilitate human life where technology develops in such a way, and even in some areas can be traced to human beings, are called Artificial Intelligence. It is argued that artificial intelligence systems are important for education, what roles they will take, and they are working on their designs rapidly in engineering sciences. “Artificial Intelligence” Systems is thought to play an important role in increasing the quality of education and economics in particular. Some of the developments in the world since 2017 that are expected to affect the national context include: 1) Analysis and categorization of students' written exams in detail; 2) Virtual reality systems for an immersive education; 3) Simulations and games for rich and deeper learning analysis; 4) Intelligent teaching systems and developments in language processing ability etc. (Budak, 2017; Göçmez, 2017; Rossi, 2017; Yasar University, 2017). Analyzing, evaluating, interpreting and evaluating on open-ended questions used on large scale exams around the world via artificial intelligence systems (e.g., Pearson, PARCC, NAEP) (Bilgeç, 2016) is quite difficult, still being studied but characterized as a great need for the future. The use of open-ended questions has been accepted for large-scale assessments in Turkey in which millions compete every year. Work on specific projects has begun. However, since the existence of a valid and reliable system has not yet been identified, this rapid transition process has been suspended. For this reason, one of the sub-goals of this study is to investigate how open-ended questions can be evaluated through an Artificial Intelligence system. The current research also examines how an artificial intelligence system to be developed can make the closest, correct, valid and reliable prediction of cognitive skills, self-checking, worry and effort dimensions from open-ended responses of the students.

2.4.2. Authentic teacher-made assessments

Turkey is a land of exams. Around 50 different exams are delivered in a year (ÖSYM, 2019). For this reason, the examination system is ready to change at any time and within the framework of education policies, they can change even overnight. However, due to the decisions taken by the Ministry of National

Education through Turkey’s 2023 Education Vision (2019), the quality of the exams will be prioritized. This means a system that will enable children to fully develop and prioritize their development as a whole so that the exam-based system will be progressed gradually (MoNE, 2019). Multiple-choice exams, with which educators have been trying to measure many skills at the same time in a very short period of time for ages, can be replaced by the possibility of taking more authentic, progressive exams. Still students’ knowledge and skills in the classes are measured by short-term exams which are referred to as authentic teacher-made exams (i.e., teacher-made items) in the Turkish education system. For example, secondary school teachers determine a part of their students’ achievement by taking 2 or 3 quarter-term examinations.

Two fundamental questions guided by the teaching learning process at any education level are “what do I want that person to learn?” and “What evidence verifies that learning?” (Tyler, 1949; Gareis & Grant, 2015). Teaching and learning have a mutual relationship whereas assessment is like another leg of a table, integrally related with them. Teachers know what their students have learned and be reflective toward themselves about their teaching methods and learning objectives. For effective teaching, one of the important foundations is the integrated relation between curriculum, instruction and assessment (Marzano, 2003). Assessment or measurement and evaluation during teaching occurs via teacher-made authentic assessments (Foley, 1981). Assessment or measurement and evaluation is therefore an essential obligation to measure the success of a curriculum (National Council of Teachers of Mathematics [NCTM], 2000). In other words, the centrality of assessment in curriculum is not a neutral element. It is a live element and always evolving according to feedback, evaluation, assessment emerging from the teaching practices. “What is assessed determines what is taught” (Verhoeven & Verloop, 2002, p.91). Some studies may reveal varying degrees or amounts of misalignment between curriculum and teaching methods, such as between teachers' assessment practices and instructional objectives, and between teachers' views, beliefs, and practices. (Verhoeven & Verloop, 2002).

In the related literature, these assessment types are called as authentic assessments (Bolat, 2016; Darling-Hammond & Snyder, 2000; Frey et al., 2012; Hayati et al., 2017; Kılıç, 2014; Kinay, 2015; Nagel, 1993), teacher-made exams/examinations (McMillan, et al., 2002; Şahin, 2014; Walker, 2006; Verhoeven & Verloop, 2002), teacher-made assessments (Gareis & Grant, 2008, 2015; Sahlberg, 2017; Şimşek, 2016; White et al., 2018), teacher-made tests (Barootchi & Keshavarz, 2002; Boesen, 2006; Broekkamp et al., 2004; Foley, 1981; Delil & Özcan, 2019; Hartell & Strimel, 2019; Huang & Wu, 2013; Kirby & Oescher, 1989; Marso & Pigge, 1991; Ort, 1967; Özcan & Delil, 2017; Smawley, 1962; Zhang & Burry-Stock, 2003), classroom tests (Özcan & Delil, 2017), teacher-made model exams (Anteneh & Silesh, 2019), teacher-constructed traditional tests (DiDonato-Barnes et al., 2014) or in-classroom assessment (İnceçam et al., 2018). Authentic assessment procedures comprise several types of assessment methods teachers can use in-class teaching such as graphic organizers, performances, portfolios, checklists, rubrics, journals, logs, discussion forums, memories, reflections, and other student self-assessment procedures (Bullens, 2002; Sahlberg, 2017).

Einbender and Wood (1995) explored teachers' experiences and their past practices on authentic assessments through an autobiographical investigation. They picked out from some interesting autobiographies of teachers that students' whole skills and knowledge cannot be measured only standardized tests. Teachers stated that they required accurately more than test scores. Understanding this complex phenomenon, Sahlberg (2017) also reached the similar conclusion. As Marso and Pigge (1991) noted, data gathered from teacher made-tests which delivered by those graduated of over 10-year period showed teachers' test construction skills do not progress with their seniority year and subject-based item writing skills in their field probably need more experience, exercise and fulfill in their bachelor's year such as in measurement and evaluation courses. Additionally, this study revealed that creating test items that required higher cognitive levels from students was not an easy task, thus teachers must acquire the appropriate information and abilities and then be urged to use these talents while creating the test items.

According to Hartell and Strimel (2019), there are various assessment techniques but teachers are rarely provided with teacher training. They have a lack of knowledge from how to prepare assessment techniques to how to gather evidence of students' learning because they have been trained neither during at the department of teacher education nor practicing teaching or inductions (Lundahl, 2009). These researchers studied with 28 schools of which have a right to decide their own teaching and learning activities and these participating schools were asked to share their assessment and instructional documents in technology context. The main focus was analysis on testing materials. To illustrate, the questions prepared by teachers in each school were analyzed in terms of being MC, alternative response, short answers, restricted-response essay, extended-response essay, performance tasks and others or incomprehensible. The main results across 28 Swedish schools discovered that test items ranged from simple multiple-choice questions to essays whereas their quality ranged from well-written examples to having mistakes with some ambiguities. Of 413 assessment items, 48% of them (n = 199) were found to be short answers although %32 of them (n = 135) fell into the restricted-response and extended-response category. Then, MC items were found to be at the fourth level among test items. However, overall results showed that the tests mainly included specific terminologies and sentences and rarely aimed to assess higher-order knowledge of students. This study taught that teachers should provide better opportunities for every student during teaching, receive fair grading and better ability to assess their higher-order knowledge or thinking skills. They frequently require expert assistance to develop suitable learning activities and assessment methods, including but not limited to examinations. To sum up, it is important to determine the general approaches of teachers in preparing question formats and to take precautions when necessary.

In a similar context, Hayati and his colleagues (2017) conducted a quantitative research design regarding the analysis of authentic assessment in the 2013 curriculum. They distributed a questionnaire to 120 teachers who were randomly selected as a sample, in Padang, Indonesia. Since with this 2013 curriculum, education policy remarked that there is a shift on assessments from tests to authentic assessments. The results showed that 69.17% of the participant teachers expressed that they always used the effective competence assessment to their students. 75% conducted the observation for their students' performance. In terms of questions types, the results depicted that 56.67% prepared to fill in the blank questions in the tests, while 48.3% gave MC and 45.83% gave essay questions. The other question types used by teachers were found to be 20% oral tests, 7.5% matching questions and 2.5% True-False questions. It can be concluded that even if they maximized the use of self and peer assessment for affective aspects, they were using very diverse types of items for the assessments of cognitive aspects, dominantly MC, fill in the blanks and essays.

Analysis of teachers' authentic tests were analyzed in terms of their appropriateness to 8th grade middle school mathematics curriculum in Manisa province, Turkey (Özcan & Delil, 2018). For this purpose, the participants included 18 mathematics teachers of 8th graders from 13 different schools in the Manisa, Turkey. 30 exam papers including 548 questions were collected from the teachers. According to the content analysis approach, the findings revealed that 88% of the items were compatible with the 8th grade mathematics curriculum attainments. Some of the items were intended to measure multiple attainments at the same time. Some of the findings also drew very important attention that changes in curricula were not reflected in the timely learning and teaching process by some teachers and that mathematics curriculum were not examined before implementation. Therefore, many teachers tended to use the same mathematics questions in their exams, and to use the questions as similar as the items from internet sources. Finally, they suggested that mathematics teachers should be educated about the table of specification before preparing in-class exams. They should join teacher meetings.

İnceçam et al. (2018) realized that the exam items prepared by teachers were only evaluated according to the cognitive taxonomy in the literature. They also determined that the test items written by teachers usually measure basic skills. For this reason, there is no study that evaluates teachers' competencies to prepare open-ended items and these competencies according to teachers' different school types, branches, and genders. They conducted a survey research design in which they studied with 167 teachers from public middle school, public imam hatip middle and private middle schools. Maximum variation sampling was used in this study. Teachers were asked to evaluate their opinions about their ability to prepare questions with the Information Form for Preparing OE Questions between 0 and 10 points. The results were analyzed with descriptive statistical methods in parallel with the research questions. The form prepared by the researchers.

The results indicated that in the open-ended item preparation, at most 50% of the teachers were able to accomplish the 18 criteria. 90% of the teachers did not have sufficient knowledge at 12 of the criteria. They were found to be partially competent on the ability of "preparing a set of item specifications", "considering the item to be clear and understandable" and "considering the appropriateness of the items with age and grade." The other research question was about whether there is a significant correlation between the competency levels of teachers and their gender, field of study and their occupational school types. Regarding this research aim, the findings showed that there was no relationship between the teacher competencies and gender variable, field of study and school type in the criteria of "preparing a set of item specifications" and "considering the item to be clear and understandable." However, on the criterion "considering the appropriateness of the items with age and grade", it was seen that there was a significant relationship between teacher competencies and gender variables in favor of the males. Teachers whose affiliation was public middle school and imam hatip middle school were found more qualified than those in private schools. Regrettably, the study discussed the fact that teachers had an extremely low level of competency on preparing open-ended questions which may trigger the inefficiency of the measurement and evaluation process.

Moreover, Şimşek (2016) analyzed teachers’ and trainers’ test items in terms of cognitive, affective, and psychomotor domain. For this purpose, 120 instructors including 62 teachers and 58 trainers were selected as participants of this study. The researcher has served as an educational consultant for all schools and corporations for 18 months and school teachers attended 80 hours of this training program, of which approximately 20 hours about educational measurement and evaluation. A total of 6450 test items in various fields of learning were analyzed to make comparisons between these groups. The items were also analyzed descriptively with regard to how much of which had prepared for cognitive, affective and psychomotor domain. The findings uncovered that most of the items (95%) were prepared according to categories of cognitive domain in which 33% items were in knowledge, of 29% were in comprehension, of 16% were in application, of 8% were in analysis and of 8% were in synthesis and of 1% were in evaluation level. Other remaining 5% of items were prepared in categories of affective and psychomotor domains. Conversely, the results found no significant differences between school learning and corporate training in addition to between the grades such as elementary education and secondary education in terms of distribution of items with subject domains.

When we took into consideration teachers’ development on assessment in history, we thought we had to show that the picture above is differently evolved. For instance, a very old study conducted by Marso and Pigge (1991) analyzed 6529 items from 175 teacher-made tests. Almost similar results showed that the teachers' test items had differentiated as 20% MC, 19% matching, 17% short answer, 15% true-false, 14% problem, 8% completion, 6% interpretive and 1% essay type of items. Furthermore, cognitive domain of items prepared as in the level of 72% knowledge, 11% comprehension, 15% application, 1% analysis, and lower than 1% is synthesis and evaluation.

To sum up, good assessment practices help teachers to understand concrete evidence of student learning outcomes by using both performance-based or more conventional, objective assessments such as tests. Even if teachers' levels of teaching were remembering level, their assessment questions can measure higher-order knowledge (Gareis & Grant, 2015). On the other hand, teachers can instruct at levels to promote higher cognitive demand, however, their in-class assessments unintentionally measure students' learning at the levels of lower cognitive demand. In both situations, there would be misalignment between instruction and assessment. In addition, teacher-made assessments should be aligned with both content and actual level of cognitive demand which they intend to measure.

Consequently, looking at overall findings of the related literature, one of the purposes of this study is to investigate middle school mathematics teachers' authentic exam items who are teaching through enacted curriculum in Turkey as a beginning in relation with appropriateness with the objectives of mathematics curriculum, cognitive and knowledge domains, item types, and an international framework. Hence, it can be objectively discussed how to prepare our children for life, society, and future goals in the world.

2.4.3. Analysis of in-class assessment: Reflection of authentic teacher-made items

Authentic teacher-made examinations have a great role in monitoring students' achievement and making evidence-based decisions. Because, in-class examinations, teachers can examine at what level the student's learning outcomes and the achievements we aim at in daily lesson plans have been achieved. Mathematics, for which many people such as adults and students have anxiety (Richardson & Suinn, 1972; Tobias, 1987) in our country as well, is one of the disciplines with the lowest success in national and international exams (MoNE, 2021). In mathematics, the reflection of in-class examinations provides us with feedback on the teacher's instructional design process, curriculum objectives, instructional learning outcomes and the quality of in-class measurement and evaluation.

The most popular taxonomy preferred to be used by educators while preparing classroom examinations is the Bloom’s Taxonomy. First of all, it was scrutinized as a single dimension by Bloom and his colleagues (1956) together with a group of measurement and evaluation experts. Currently, it continues to be preferred by many educators in terms of its use and keeping it up-to-date. Bloom's Taxonomy was originally conceptualized to assist curriculum specialists and planners in setting goals, planning educational experiences, and preparing assessment tools (Bloom et al., 1956, p. 2). While writing the objectives, it included the hierarchy of knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al., 1956, p. 17). The writing of educational objectives and their association with instructional activities and assessment tasks has become an important exercise for policy makers, curriculum designers, test designers, and teachers. However, with the evolution of complicated knowledge, skills and competencies over time, the emergence of higher-order cognitive skills and the necessity of evaluating them, the taxonomy has been divided into two dimensions: knowledge dimension and cognitive process dimension. Revised by Anderson and Krathwohl (2001), while the knowledge dimension was sequenced as factual, conceptual, procedural, metacognitive; the cognitive process dimension was revised as remembering, understanding, applying, analyzing, evaluating, creating as from noun to verb (see. Anderson & Krathwohl, 2001 for the Revised Bloom Taxonomy) (Anderson et al., 2001; Anderson & Krathwohl, 2001)).

It is recommended that teachers, while preparing their in-class authentic examinations, first should make a table of specifications and match the curriculum outcomes and/learning outcomes with the prepared examination items. Because the more examination items are aligned with the students’ grade level, curriculum goals and learning outcomes, the better they can cover the content. In addition, it is seen that in-class examinations include items from all cognitive levels, reflecting the professionalism of teachers, and allowing them to make more valid and reliable assessments. However, despite all these, the difficulties of clustering classes in line with the Bloom's taxonomy (Long et al., 2014) are also asserted, and other taxonomies launched to be used in the international literature are also recommended to be used in in-class examinations and measurement and evaluation processes [e.g., Mathematical Task Assessment Hierarchy (MATH) Taxonomy (D'Souza & Wood, 2003; Smith et al., 1996), Structure of the Observed Learning Outcome (SOLO) Taxonomy (Biggs & Collis, 2014), TIMSS Framework (Long, et al., 2014; Usiskin, 2012)].

Although there are many studies investigating the level of in-class examinations according to Bloom's taxonomy, there are also studies using an international framework (e.g., TIMSS) in the literature. However, there is little evidence on this topic (Delil & Özcan, 2019; Long et al., 2014). To take a closer look at a few of the studies, for example, Çağlar and Kılıç (2019) conducted a descriptive study on the content validity of central exams and secondary school mathematics teachers and teacher-made examinations in Düzce, Turkey. In addition, the scores of the students in these examinations were scrutinized in terms of various variables. Descriptive, quantitative, and qualitative processes were used in the research method. 40 schools included in the study were selected through the maximum variation method, and the teacher-made examinations and central exams of these schools and the test scores of 1848 students who took these tests were included. The qualitative data of the research were collected with the prepared specification table, question-achievement table, exam achievement table data collection tools and analyzed with the document analysis technique. Quantitative data, on the other hand, were collected with a personal information form and transcript and analyzed using correlation by Spearman Brown Correlation, Linear Regression, Mann Whitney U and Kruskal Wallis techniques. If we look at the extent to which 40 teachers measure the scope in the classroom examinations, this rate is between 60 and 80%, and it has been revealed that the average content validity percentage of all teachers is 72%. It was observed that the content validity of teacher-made examinations and central exams were at the same level, there was a high correlation between the scores of the students from the two tests. The most preferred content on which the test items developed was 'multiplication in exponential numbers'. Some of them prepared items on the coordinate system from the analytical geometry unit. Unfortunately, few items were encountered in the data analysis unit. The effect of the gender variable on the scores of teacher-made examinations was revealed. It has been suggested that faculties of teacher education, teachers in-service training institutions and centers that prepare central examinations should give more importance to content validity, and that alternative measurement tools other than multiple-choice tests should be used in order to accurately and fully measure the skills and competencies to be developed in the secondary education mathematics curricula and the desired outcomes. The importance of asking items that would rely not only on cognitive but also on psychomotor skills has also been considered.

Correspondingly, Çevik (2009) analyzed the High School Entrance System (LGS) (a.k.a. SBS) and teacher-made examinations within the scope of the social studies course and concluded that the examination items were not sufficient to measure all the achievements and the distribution of items was disproportionate. Aldım's (2010) study on English as a foreign language course items examined English teachers' in-class examinations and LGS items and found out that the items were not fully aligned with the curriculum, important learning outcomes were ignored, and the items solely measured morphology rather than course content knowledge. İnci (2014), on the other hand, in her thesis study, concluded that only 68 items which are aligned with learning outcomes were asked in the examinations out of the 137 learning outcomes in the science and technology curriculum, which should be addressed in the science and technologies part of TEOG examinations, and the content validity was low. When the learning outcomes related to the general examinations delivered in the schools are examined, it has been determined that they are only aligned with certain learning outcomes and that the general examination items were concentrated on lower-order cognitive skills. In the interviews conducted with the teachers; They stated that they found the number of items asked in the general examinations related to the science and technology course to be sufficient, that the items were compatible with the curriculum, but only some learning outcomes in the curriculum were emphasized in the examination and no items were asked about other remaining learning outcomes.

Delil and Özcan (2019) emphasized that in-class assessment is still a problematic concept, and they analyzed 548 item types from 30 in-class examinations shared by 18 mathematics teachers in 13 schools in Manisa, Turkey. Teacher-made examinations analyzed by content analysis method were classified in terms of TIMSS-2019 cognitive domain, item types and test construction errors. Following this, the results determined that 50% of the examination items were based on level of Knowing, 43% were level of Applying, and only 7% were based on level of Reasoning. The least frequency represented in each examination paper was regarding level of Reasoning. While the majority of the examinations (83%) consisted of multiple-choice items, a few of them (17%) included constructed-response item types. Unfortunately, it was determined that the teachers did not prepare the authentic items in line with their own knowledge and skills, they generally tended to copy and paste from various sources, and they mostly tended to use the internet as a primary source. In fact, it has been emphasized that teachers prepare in-class examinations without considering the cognitive domains known to be related to the quality of exams.

To sum up, related research on in-class examinations and authentic teacher-made items (Aldım, 2010; Birgili et al., 2021; Çevik, 2009; Çağlar & Kılıç, 2019; Delil & Özcan, 2019; Güvendir & Ozkan, 2021; Hartell & Strimel, 2019; İnci, 2014) and the properties of the examinations showed that there is still a need to analyze the in-class examinations of mathematics teachers in detail, at national and international level. In-class examinations cannot measure the learning outcomes of the national mathematics curriculum in a comprehensive way, but the fact that they are related to the prospective exam scores reveals its importance once again.

2.4.4. Cognitive diagnostic models

The development of computer and instructional technologies calls for the development of new technological tools which has taken place owing to today's needs. New technologies developed not only affect fields such as medicine, engineering, informatics, finance, data science or cognitive sciences but also education directly.

Cognitive diagnostic models (CDMs) are an area of psychometric research which entails mathematics and statistics behind them. It emerges against the idea that overall scoring on an individual cannot explain everything about the success of that individual. Therefore, as the technology evolves, new assessment models serve our purposes. With this purpose, cognitive diagnostic models evaluate an individual's intended skills or traits (attributes) to provide complete feedback and enhance his/her learning -skills (Ayan, 2018; Başokçu, 2011; Wang & Jiang, 2018; Wang et al., 2018). In other words, these models classify test takers' response patterns on a test into a set of attributes which is related with different hierarchically defined mastery levels (Briggs & Circi, 2017). For instance, a certification test evaluates and finds an overall score which decides whether to pass or fail even though the aim of a formative assessment in education is used to give feedback to students and teachers their strengths and weaknesses. Moreover, a mathematical item "105+205" requires an additional attribute to answer correctly. Another mathematical item "12x15" requires only the multiplication reasoning whereas an item "21: 7 – 8 x 4" differs from others that requires more than one attribute. This new tool depends on a different paradigm called latent class theory than classical item response theory (IRT) or classical test theory of the 1980s. According to latent class theory, it accepts an examinees' single score to assign sets or groups which are categorized along a number of axes (Yamaguchi & Okada, 2018).

Our children need to identify their knowledge and cognitive development according to different skills rather than a single score. Providing new learning analytics to their needs has been the necessity of this era. The tool of this new era designed to determine the level of cognition is cognitive diagnostic models in the child's cognition (their learning status). Especially when we consider the need to evolve the measurement and evaluation strategies in large-scale assessments towards a multiple-choice open-ended question, these tools may be the solution to this new problem. In addition, CDMs provide valuable diagnostic information about education which may enhance teaching and learning (de la Torre & Minchen, 2014). Especially, the assessments of any learning at school are both for making decisions and for providing feedback in the process (Çıkrıkçı-Demirtaşlı, 2017).

To illustrate how CDMs help the item mapping process, de la Torre (2012) remarked on the scores of 2013 NAEP Grade 8 mathematics tests (U.S. Department of Education, 2013). In this item mapping study, students' scores corresponded to a different cognitive level. For example, 331 points correspond to at the top of Proficient level and its measured skill was "calculating the area of an inscribed square". 296 points is at the top of Basic level and its measured skill was "using average (mean) to solve a problem". On the other hand, 257 points is below Basic level and its measured skill was "solving a problem involving rates." These descriptions specified some evidence about the types of problems students can and cannot response (de la Torre & Minchen, 2014).

Yamaguchi and Okada (2018) are two researchers who studied CDMs model from Kyoto University. They believed in recent years development of CDMs to diagnose students' achievement and skills. They ground their study on the idea that each CDM model has different assumptions about students' achievement so that they empirically investigated which CDM model better fitted the actual data. They examined the problem comparatively by using the representative CDMs to TIMSS 2007 assessment data from seven countries: USA, Hong Kong, Singapore, Slovenia, Armenia, Qatar, Yemen. They stated that the major result emerged was CDMs had a better fit than the IRT models in keeping with former studies because of IRTs restricting students' latent ability to be a unidimensional trait or at most few dimensions. They came up with the idea on working with this model that in order to succeed in the TIMSS mathematics assessment, students expect more than one skill.

Another study which investigated the quality of a teacher-made test and diagnosed their students' learning misconceptions on fractions and decimals was conducted by Huang and Wu (2013). Teacher-made mathematics tests included 22 items and were taken by 32 4^{th} grade students in Taiwan. For this purpose, they used a CDM named as BW model. This model worked on four personal attribute indices (psychological response aberrances) such as carelessness, capability, guessing, misconception, and also four item-facet indices such as difficulty, disturbance, hint, and indistinctness. They assumed that analyzing students' misconceptions and learning mistakes help to enhance their conceptual grasps. They also explored underlying content knowledge in items. By using the BW model, Huang and Wu found that the BW model showed a good indicator of agreement between the middle school fourth grade students and this teacher-made test on fractions. It meant that the students' abilities on fractions and decimal concepts were well assessed with the test. Items in the test demonstrated the levels of item difficulty from .13 to .94 and item discrimination from .00 to .88. The Q matrix of items related to concepts showed concept of *equivalent fractions* was most mastered and the concept of *transforming fractions into decimals* was the least.

Also, computerized adaptive testing (CAT) changed the testing paradigm. The idea of computerized testing was proposed by Weiss (1973) (Kalender & Berberoğlu, 2017). It equates the test difficulty with the test taker's ability. It has higher reliability with fewer items compared to paper-based tests (PBT). Glas and Van der Linden (2001) stated CAT can decrease the test anxiety of test takers by providing more than one chance of taking the test. It also helps administrators to prepare different new item formats, such as interactive items, multimedia items, etc. There are many practical advantages of using CAT. Not only individual researchers but also various professional organizations such as the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) (AERA, APA, & NCME, 2015) compared PBT and CAT.

Kalender and Berberoğlu (2017) aimed to investigate the difference between simulation analysis and live test administration to determine the best practice for Turkish university admission system. For this purpose, in order to simulate Turkish students' higher order cognitive processes, a total of 5,000 students who had taken the university admission tests (in other words, higher education transition examination [HETE]) were randomly selected from each school type in the database pool. A total of 15,000 students formed sampling in the simulation. For the live application, 37 students volunteered to take the CAT version were selected as the research sample. As a content, science tests were selected from the HETE applied in 2016. It was made of 45 MC items with the alternatives. According to the findings, the researchers found that the use of CAT as an alternative to the admission system operated a similar role on the PBT version. CAT was able to generate ability estimates with standard errors below .30. But, if fixed- length CATs contained 10 and 15 items, they were not able to generate almost no ability estimations with a standard error of less than .30. It meant that using very few numbers of items did not explain a test taker's ability in a reliable way. The correlation between science and math tests was validated to estimate CAT's ability. In the live CAT admission, the results were the opposite. In the simulation findings, CAT's ability estimations were lower than those of the PBT by school types. On this finding, the researchers discussed that the items were quite simple for the higher group. That's why they did a good job on PBT. Both tests seemed to have the ability to classify a high proportion of examinees into the same percentiles. The researchers suggested further researchers to use large item banks and think about sample representativeness as limitations.

Finally, a very recent study used cognitive diagnostic computerized adaptive testing (CD-CAT) conducted by Lin and Chang (2019). They were targeted to investigate their proposed item selection method more adapted to regulate attribute balancing, exposure status, and precision. They constructed a simulation. The independent variables were item selection method, test length and number of attributes. Briefly, the results showed that SWDGDI method for item selection successfully balanced attribute analysis in CD-CAT because this method had its weighing scheme and the capacity to combine a variety of non-psychometric constraints. Besides measurement and evaluation features, in terms of educational purposes the researchers suggested that to use classroom assessment, the CD-CAT model should not cover a broad range of topics. Classroom assessment may be suitable for CD-CAT formative assessment type. Hence, CD-CAT can be inserted inside the teaching and learning process. Educators can benefit from CD-CAT to specify their learning objectives and instructional strategies after deciding what their students learned and struggling areas. Then, specific suggestions toward a student can be given to improve their higher level of learning.

Last but not least, Doğan and Tatsuoka (2008) studied a CDM model in order to reveal Turkish middle school students' profile, basically mathematics skills, in the TIMSS-R international assessment. They intended to analyze Turkish students' performance with a diagnostic model called the Rule Space Model. For this purpose, firstly, they determined students' mathematical and cognitive skills (attributes) which had been measured by the test. They determined students' master profile by using students' response data and an incidence matrix; the Q-matrix. 62% of Turkish students' profiles found to be lower quartile in the international assessment whereas only 1% were in the top. Through this problem, their aim was to obtain more detailed information about these students' profiles in order to make more accurate international comparisons. In this study, they used data from two samples of eight grade students (2900 Turkish and 4411 American) who participated in the 1999 TIMSS-R. Some of the results uncovered that US students outperformed Turkish students on 17 of all 23 content attributes. Largest differences found to be on the attributes related to quantitative reading and estimation in favor of US students. The attributes of Turkish students were weak in quantitative reading, estimation, patterns and relationships and solving open-ended items. From these attributes, it was inferred that Turkish students comparatively did not perform well during uncertainty, develop rules and construct unique answers as opposed to selecting from given alternatives, and grasp suggestions by using logical thinking.

After that, when Dogan and Tatsuoka (2008) determined the learning paths from the analysis, they found that most of both Turkish and American students tended to learn skills for responding problems prior to any other skill. Interestingly, Turkish students tended to learn geometry subject skills first whereas American students primarily tended to learn number skills. They noted that the Turkish students outperformed the attributes “basic concepts in geometry.” These results and discussion might be one of the reasons derived from nearly ten years ago that our Turkish students did not even know how to deal with open-ended questions and construct their responses. They cannot perform well to answer open-ended questions and still have many discussions going on.

As a recent study- up-to-date – Ayan and Çıktıkçı (2021) created a test to assess the four key cognitive skills in the fractions sublearning domain of mathematics. A total of 1380 students from six to eight grades were given forms. This multiple-choice test forms included 89 items in 5 forms, whose learning area was “numbers” and sublearning area “fractions.” Hierarchical CDM was used to estimate latent classes. CDM-based diagnostic result reports were created, with detailed outputs on the cognitive development levels of the individuals. Members of each latent class were also given specific instructions on what they could and couldn't do. Some striking points revealed from the study were the students were either largely or totally lacking in basic learning skills related to the fractions, regardless of their grade level and it was concluded that, regardless of the student's grade level, learning remained incomplete, and that this severely hampered future learning. It is suggested that in-class assessment and evaluation activities be used more frequently as a method for directing and learning weaknesses than quantitative and outcome-based processes. National monitoring systems based on the CDM can be established, particularly for diagnosing and monitoring learning in the field of mathematics.

To sum up, unlike IRTs, CDMs consider a variety of students' cognitive abilities with a probabilistic approach. Teachers can benefit from their teaching and learning process (Huang & Wu, 2013). New generations' diagnostic tools- CDMs- help teachers review their curricula by providing more concrete evidence. Hence, they can only focus on unperceived specific attributes learnt from the CDM results during instruction. CDMs are suggested to be used in international competencies (Dogan & Tatsuoka, 2008; Lee et al., 2011) and to be able to provide individual outputs rather than collective outputs or single score reflecting the state of the country. The literature has been limited in the knowledge that will support practical applications for direct use of CDMs, CAT or CAT-CDMs within the class (Ayan, 2018; Briggs & Circi, 2017; Dogan & Tatsuoka, 2008; Erdoğdu, 2009; Lin & Chang, 2019; Wang, 2021). But they can be projected to be valuable new tools which may somehow determine the new generation's intended skills not only within the class but also international assessments, including dealing with open-ended questions, in the future.

2.5. Student Monitoring Systems in Turkish Education System and the Possible Challenges

Various student monitoring systems (The Project of Monitoring and Evaluating Academic Skills [ABIDE], CITO-Turkey, PISA, TIMSS) have been issued for primary and secondary school students to determine the relationship between their achievement and in or out of school variables. One of the three modules of CITO-Turkey Student Monitoring System was Student Social Development Program (ÖSGP). It was aimed to determine the relationship between student academic achievement and social, educational, affective factors. CITO-Turkey Report (Issue of October-December 2010), in addition, focused on elementary school students' higher order thinking skills in the mathematical content area. The report warned us that the enacted mathematics curriculum had mostly content-focused learning outcomes, so the teachers did not focus on their students' thinking process. The students tended to memorize the subject on an algorithmic level. Hence the level of use of memorization strategies increases as the class level increases (İş-Güzel et al., 2010). The findings remarked that the in-class teaching and learning activities should be always designed to promote middle school students' not only low-level but also higher-order thinking strategies in order to protect them from memorization.

The popular usage of MC in large-scale assessments has achieved by chance and lack of ability to test out higher order cognitive skills. Therefore, another innovative aspect of the student monitoring system is testing the variables that teachers cannot experience and measure in order not to disrupt teaching within the classroom. For example, one of the objectives of CITO-Turkey is to allow experts to test open-ended question formats. Students' responses to the open-ended questions were measured and evaluated. The system was adapted to automated scoring. In the report it was indicated even though the automated scoring of OE was subjective and time consuming that might be one of the challenges (Çıkrıkçı-Demirtaşlı, 2010). Gültekin and Çıkrıkçı-Demirtaşlı (2012) also supported these claims by indicating that the answers from OE items gave more information about the students' mathematical achievement than MC. In this context, the inclusion of MC as well as limited OE in large-scale assessments may increase the level of thinking skills measured by qualified questions, which can prevent guessing strategy of MC, and according to the results of these tests it will be ensured that the selection or qualification decisions given are more valid and reliable.

The last and most updated report regarding determining students' ability to answer OE and relation between several variables was ABIDE (2016). In this monitoring system, human evaluators assessed students' answers to OE mathematics, Turkish language, science and social sciences according to the given rubric. The results revealed that 60% of the 8th grade students (N = 20780) were found to be at the below baseline and at the baseline level, about 29% (N = 9956) at moderate level and 11% (N = 3922) at intermediate level and advanced level in mathematical content. The students found to be at below baseline and baseline level were able to do remembering, routine calculations, using knowledge from given instruction, understanding principles and rules.

When the possible challenges have been considered, as the level of difficulty of OE and the number of steps required for solution increases, the diversity of student responses increases. Regarding errors that students make at different stages of different solutions, more information can be obtained about the misunderstandings. It is possible to get rich feedback on students' learning by OE. Parallel to this, consistency between evaluators is decreasing that might contribute to a possible challenge (Bilgeç, 2016). However, the results from these several student monitoring systems lack much of the difficulties teachers or evaluators had in preparing open-ended questions. All in all, it is expected that this study will be modeled on the creation of new systems to evaluate OE questions better by focusing on the difficulties and minimizing the challenges.

2.6. Differential Effect of Open-Ended vs. Multiple-Choice Item Formats

The most typical item-formats used in assessments are multiple-choice (MC) or constructed-response (CR), in other words open-ended (OE). However, once MC has been introduced to educational testing, many studies showed advantages and weaknesses of this format and various studies have been conducted to examine these claims (Bonner, 2013; O'Neil & Brown, 1998). Some studies conducted in this context vary in terms of research design such as quantitative (e.g., experimental, survey, correlational) (Bonner, 2013; O'Neil & Brown, 1998; Dutke et al., 2010; Gullie, 2011; Mulkey & O'Neil, 1999), qualitative (e.g., Birgili, 2014; Burfitt, 2019; Duran & Tufan, 2017; Koyuncu, 2017) or mixed (e.g., Herman et al., 1997). Some of them differentiated in content domain such as problem solving (e.g., Bonner, 2013), mathematical reasoning (e.g., Gullie, 2011; Herman et al., 1997; Wang, 2002), test-takers use of different strategies (e.g., Haladyna et al., 2002), science achievement in relation with gender and ethnicity (e.g., Dimitrov, 1999), self-efficacy and worry (e.g., Mulkey & O'Neil, 1999).

For instance, Bonner's study (2013) was an experimental design research conducted with 64 undergraduate examinees. The participants were applied think-aloud to examine their cognitive processes in mathematical problem-solving situations. At the same time, the study examined the relationship students' strategy use on items, test format and gender through the students' verbal reports on their mental processes. The findings revealed that CR found to be more difficult than is the MC format. MC was related with more varied solution strategies, guessing approach and backward strategies. On the other hand, there was no main effect of gender on students 'performance. Interestingly, disproving the hypothesis, this study failed to find format differences in use of metacognitive self-regulatory strategies. Gullie's (2011) thesis was aimed to study the predictive ability of students' responses to OE and constructed questions on their fifth-grade mathematics achievements by considering their third and fourth grade mathematics content sub-categories. The results of this study implied that OE predicted fifth grade mathematics performance as proficient/non-proficient outcome levels. As well as investigating the implication of OE/CR response questions, the study also evidenced once more that CR is an important predictor for analysis on student outcomes on achievement tests in mathematics.

Response format comparisons in the related literature have included the studies that explore the relationship, cognitive and affective and metacognitive aspects (Dutke et al., 2010; O'Neil & Abedi, 1996; O'Neil & Brown, 1998). However, the findings revealed that OE seems to be more preferred than MC (Sole, 2018), at the same time, many students do not necessarily like challenges that OE items introduce. The reason why some of those prefer the traditional assessment format- MC- was because of the relative novelty of open-ended items. The results also signified students' misunderstanding of assessment of their performance on OE items. The discussion and scientific investigations on this area seem to be still going on and worth pursuing (Ryan, 2001).

Measuring knowledge through alternative ways has become a necessity in educational processes today as studies show that MC and OE measure separate constructs (Beller & Gafni, 2000; Hoogerheide et al., 2019). For instance, validity is high in OE as students do not tend to do guessing whereas reliability is high in MC. One of the most important reasons for this is the possibility that more questions can be written. Heck and Stout (1998) claimed that MC necessitates recognition and recollection. The main strength of MC is that these tests can evaluate a high number of abilities with fast grading and objectivity. Also, there is less necessity for crosscheck and more efficiency in marking. MC tests are also thought to be highly reliable considering their fairness. However, MC is criticized for its limited focus and tendency to decrease the subject specific knowledge to only verifying facts (VanSledright, 2008). Thus; OE, which enables examinees to organize and generate their own answers, has become popular. Still, OE has some weaknesses such as being complicated, time consuming, and less reliable. Shavelson and colleagues (1992) note that rapid shifts should be clarified carefully. It should not be overlooked that some topics are very sensitive to assessment methods.

Instilling meta-cognitive dimensions is not a novice concept; however, testing them through national or international large-scale assessment is an innovative notion. In prior research, O'Neil and Brown (1998) worked on eight graders' meta-cognitive and affective processes during a large-scale mathematics assessment. Three-factor mixed model design (Kirk, 1982; as cited in O'Neil & Brown, 1998) attempted to investigate the format's effect in terms of gender and ethnicity as between-subjects factor and meta-cognitive and affective variables as within-subjects repeated measures. Mathematical items in MC and OE forms were administered to the students. It showed that MC caused greater self-checking than OE due to its novelty. Also, OE encouraged more cognitive strategy usage and less self-checking behavior than MC. Sole (2018) also agreed that OE questions should be part of instructional strategies in the classroom because students can solve a problem in a variety of ways. Assessment of students learning through OE makes them use their not only prior knowledge and level of proficiency but also self-confidence to use different solution procedures and accurate decisions on how to approach mathematical situations (Sole, 2016). More than one correct solution can be constructed as a response. They use thinking skills beyond procedural knowledge (Sole, 2018).

Moreover, a study on a similar context conducted by Eren (2015) who studied on a descriptive study which investigated the relations of the scores obtained from different test types of question formats (only multiple-choice question-MC test and mixed test formed with MC and restricted open-ended questions-RQ together) and determined the students' and teachers' view about different test types. This study discussed that mixed tests stated a more reliable measurement than other formats according to reliable coefficients of both tests. Also, the results from students' views about question formats stated that students tend to prefer taking MC tests more and the reasons for their preferences vary.

Finally, Öksüz and Güven-Demir (2018) conducted research on the differential effect of open-ended questions and multiple-choice tests with regard to some psychometric features and student performances. The items were prepared in both forms according to two units of 4^{th} grade science and social sciences curricula. The study was conducted at 2016-2017 academic semesters in Samsun province, Turkey. The group were 102 5^{th} grade students including 52 females and 50 males from a public middle school. For content validity of tests, the researcher studied with 8 experts from the academics. OE and MC items were compared with regards to cognitive levels, item difficulty, item discrimination, reliability, and student performance. As stated by the results, the researchers found out that there is a significant difference between OE and MC in terms of item difficulty. The MC tests items prepared in both science and social sciences were found to be easier than OE ones. The type of test had a small effect on this item difficulty difference. In the science course, the application-level questions were found to be mid-difficulty in the open-ended and easy in the multiple-choice. On the other hand, in the social sciences course, the items significantly differed cognitively in comprehension, application and analysis levels. In these three cognitive domains, multiple-choice were found to be easier than open-ended.

Moreover, in terms of item discrimination findings revealed that multiple choice tests for science and social studies courses had more item discrimination compared to open-ended, test type variable had a great effect on the difference in science course and a small effect on social studies course. However, they found no difference in reliability between open-ended and multiple choice. They discussed that different variables such as handwriting, simplicity of a question might affect reliability even if there might be some significant difference in favor of MC in the literature (Bağcan-Büyükturan & Çıkrıkçı-Demirtaşlı, 2013). While tests in science courses showed a significant difference in the student performance in favor of MC, it was not significant in the social studies. Accordingly, the study showed that as the scores of the students in the open-ended cognitive domain increased, the scores in the multiple-choice also increased.

It can be concluded that the Turkish education system using open-ended questions in large-scale assessments is believed to create a new dilemma in the educational context. In particular, how do the teachers prepare students to gain the necessary knowledge, skills through the teaching methods in the classroom? how does he/she measure and evaluate a small prototype and simulation of these exams in their in-class teacher-made tests? How do students construct their knowledge on both cognitive and metacognitive levels? Most importantly, if we have an assessment system that aims to include millions of open-ended students, how should an artificial intelligence system be evaluated? stand as a big picture. In this context, it is tried to look at the big picture by using the following methodology. For this reason, the search for a system to evaluate open-ended questions has been countered and final and the main goal of this current study is to contribute to this search.

2.7. Modeling Metacognition and Affective Processes with Eye-Tracking Tools and Biomarkers

Metacognition was defined as the cognition of cognition (Flavell, 1976, 1979) and expanded as the method by which people reflect on their own ideas to come up with solutions to problems; it was also defined by Nelson (1996) as a model of cognition. Efklides (2001, 2006) defined metacognition as having more aspects such as metacognitive knowledge, metacognitive skills and metacognitive experiences. As asserted by her study (2009), metacognition has multiple facets, each of which leads to self-regulated learning. For instance, “metacognitive knowledge through the affective regulatory loop, which involves affect and motivation, lead to both the short- and long-term self-regulation of effort and persistence; through the cognitive regulatory loop, which involves metacognitive knowledge, metacognitive skills, and metacognitive experiences such as metacognitive judgments, also contribute to the short- and long-term self-regulation of cognition” (p. 81). As indicated, there are a huge number of sub skills of metacognition. Metacognition are also found to be related with such skills; they are reflecting, recognizing, monitoring, reflexive thinking, acquiring, retaining, transferring, being capable thinkers, cognitive strategy, self-checking, planning, evaluating, goal setting, continued monitoring identifying what you know or what you do not know, adapting as necessary, self-questioning, annotated drawings, self-explanation, concept mapping, making checklists, reciprocal teaching, organizing one’s own thinking, being responsible learners of one’s learning process.

However, metacognitive knowledge and skills that occur in people's inner cognition, the use and experience of these skills cannot be measured as easily as it is thought, and continues to be the focus of various research studies. Skills grounded in emotion and cognition are measured and evaluated with psychological instrumental measurements and physiological instrumental measurements, and even participant groups (i.e., adults or school children) are expected to reflect their inner speech by being exposed to the think-aloud process (e.g., Björn et al., 2019; Öztürk & Kaplan, 2019; van Gog & Jarodzka, 2013) because merely knowing about a subject-specific/content knowledge and procedural skills does not always mean using it consciously. For this reason, measuring it with valid and reliable measurement tools becomes significant in terms of testing the accuracy of the responses given and reflected by the research participants to the researcher's questions.

There is burgeoning interest in using eye-tracking technologies, galvanic skin response (GSR) tools, smart watch tools, graphic/bamboo tablets, GoPro high tech cameras, log-files, physiological lab data, screen recordings, think-alouds, self-reporting/self-explanation sessions, facial expression of emotions, and linguistic analysis of discourses (Azevedo, 2002; Azevedo et al., 2017; Azevedo & Gašević, 2019; Jarodzka et al., 2017; van Gog & Jarodzka, 2013) in educational research studies. These tools were found to be affective and reliable to collect and analyze physiological data from participants so that they have been preferred in multimedia learning (e.g., Alemdağ & Çağıltay, 2018), problem solving situations in mathematics, reasoning, scientific reasoning, literacy skills, testing the effectiveness of cognitive and metacognitive techniques (e.g., Apaydin & Hossary, 2017; Öztürk, 2021; Zhang, 2018), testing a metacognitive-based training (e.g., Öztürk, 2021). The studies conducted in measuring metacognitive knowledge, skills and experiences, metacognitive self-regulation via innovative eye-tracking tools scattered in the related literature around commonly in experimental/quantitative research designs (e.g., Low & Aryadoust, 2021; Taub & Azevedo, 2018; Van Loon et al., 2020), lower number of qualitative research designs (e.g., Damayanti et al, 2020; Salvucci & Goldberg, 2000) and prospering field of mixed-methods designs (see. Deluca & Lari 2013, Stark, et al., 2018 for a mixed-methods or Öztürk, 2021 for an embed mixed-method design). The subject field mostly rely on, up-to-date, reading (e.g., Aryadoust, 2019; Zhang, 2018), mathematics education (e.g., Norqvist et al., 2019; Strohmaier et al., 2020), science (e.g., Ariasi & Mason, 2011; Tsai et al., 2019), chemistry (e.g., Muna & Bahit, 2020), economics (e.g., Low & Aryadoust, 2021), engineering (e.g., Elliott et al., 2019), early childhood (e.g., Marulis et al., 2020), foreign language learning (e.g., Šafranj, 2019) and other uncategorized fields (e.g., Azevedo & Gašević, 2019; van Loon et al., 2020) such as computer-based learning environments.

Theoretical discussions, review papers and even professional conceptual papers began to assert that metacognition is a groundbreaking discussion point when it has been searched under the roof of innovative technological tools. Theoretical, conceptual, methodological, analytical, and instructional issues have been conspicuous research fields on metacognition (Azevedo, 2009; Azevedo & Aleven, 2013; Ching-En, 2018). Owing to the fact that cognitive, motivational, and behavioral attributions may impact learning, the most-often studied areas resolve cognitive development, metacognitive subskills, patterns of information processing, and the effects of instructional strategies on learning. The studies conducted within the context of the eye-tracking technologies highlight that studies (e.g., Bannert & Mengelkamp, 2007; Deluca & Lori, 2013) were mostly conducted with college students and eye-movements were collected for make inference about the cognitive process of selection, organization, integration skills. The eye tracking measurements and learning performance of the students can lead to the relationship between those cognitive processes. Metacognition, and emotions (i.e., affective processes in some studies) were the potential influences that can affect their eye movement measurements. In line with the current literature, Azevedo and colleagues emphasize the prominence of using multimodal data to examine the complex functions of cognitive, affective, and metacognitive processes among students during learning process or problem-solving situations. The dilemma of valid metacognitive skill diagnosis has sometimes been neglected in metacognitive research; however, this question has recently been put on the research agenda by some researchers (Azevedo & Aleven, 2013; Jarodzka et al., 2017). The theses in Turkey, for instance, as of 2010, most of the thesis were presented in master’s degree (80%) (e.g., Akgün, 2018; Coşkun, 2019; Malcı, 2021) and minority of them were presented in PhD dissertation (20%) (e.g., Ayhan, 2019; Bayraktar, 2014; Yılmaz, 2019). The faculties of studies which focused on technology focused eye-tracking tools were held in CEIT (53.55%) (e.g., Coşkun, 2019; Malcı, 2021), English language teaching (33.33%) (e.g., Akgün, 2018), Turkish language (6.67%) and Childhood education (6.67%). In addition, the selected studies mostly utilize a qualitative research design (e.g., case study) (f = 7, around 46%) and some studies employ a mixed-methods (f = 3, 20%). Around 7% of the theses (f = 1) did not specify their research design approach. Concerning educational stage, the theses seemed to be conducted frequently in research settings of higher education (f = 8; 53.33%) while similar number of studies were conducted at the secondary school levels (e.g., middle school and high school) (f = 7; 46.67%) levels. Most of the theses in the Higher Education Council theses database (YÖK, 2021) selected students as participants (e.g., Dağlı, 2014; Malcı, 2021; Yılmaz, 2019); least frequently teachers, university students and faculty members (e.g., Akgün, 2018; Coşkun, 2019). The number of minimum participants were two and of maximum participants were ninty-five. Finally, the theses were held in the subject area of mathematics (e.g., Yılmaz, 2019), geometry (e.g., Malcı, 2021), social sciences (e.g., Dağlı, 2014), English language (e.g, Akgün, 2018) and Turkish language (e.g., Ayhan, 2019). Hence, the minority of doctoral theses in Turkish literature and the fact that eye-tracking tools have never been used in the department of educational sciences up-to-date revealed and supported the need for this study. The fact that there are studies conducted with middle school students in the subject area of mathematics also supports my participant group.

Thus far, considering challenges of doing interdisciplinary areas of research and encouraged by the advancement of the field of metacognition and affective processes, the current study would make progress in the field of neuroeducation; progress on the research findings or propositions which need to be empirically verified. Last but not least, the current research study would need to fill the gulf for empirical evidence by using many innovative tools such as an eye-tracking tool, a bamboo tablet, GoPro camera, GSR smart watch in a holistic manner; for conducting research with middle school students who are preferred rarely due to accessibility; for methodological gap by designing a mixed-method research; for practice gap by measuring middle school students' affective process aside from metacognitive sub skills as cognitive strategy and self-checking (Deluca & Lori 2013).

2.8. Imperative Need of Artificial Intelligence Systems in Education

As mentioned above, children will face more open-ended questions in national and international exams. Although it is an inevitable fact, how and by whom it should be read and evaluated is a possibility and the most obvious tool in the world of possibilities should be artificial intelligence (AI) systems. In fact, when Alan Turing had named them as "thinking machines", "electronic computer" or "digital computer" in the 1950s.

Because, when we consider independent evaluators and machine scoring, artificial intelligence systems have preferable features in terms of time, economy, accountability and objectivity (Hernandez-Orallo, 2016; Üstkan, 2007). Systems that work with artificial, intelligent, deep learning algorithms, which can keep some of these variables under control, are increasing day by day. Robotics, coding, driverless vehicles, natural language processing, engines, maps, social media, cognitive sciences, medicine, mathematics (Conati, 2016; Lemaignan et al., 2016; Russell & Norvig, 2009) as a great hand that will cure almost every field. However, their task is specialized in each discipline. For instance, robotic navigation of a Mars traveler can share some of the techniques with the AI designed for a driverless car on Earth, but their final application is extremely specialized in both cases (Hernandez-Orallo, 2016). In addition, they have mastered image recognition. They have started to learn about product and advertising suggestions, medical support, imitating or even defeating the athletes of the years, and defeating people in games (Turing, 1950). As Prof. Cem Say (2018) says "…Now computers know a lot about people. We have already passed the point of no return in the IT revolution; we can no longer return to a world without a computer.” Computer systems are systems and environments where human thinking and ability can be simulated (Thompson, 2003 as cited in Demir, 2004). They can do a job in a shorter time with higher performance than a human-being. Because, contrary to what everyone thinks as a black box of the future, AI systems exist in order not to make our lives harder or not to get our jobs out of our hands but to make it easier. Thus, for people, the time to focus on one’s own happiness and existence seems to be shortened.

2.8.1. History of AI: Computing machine vs. intelligence

First known eponym of “artificial intelligence” in history was John McCarthy (Allahverdi, 2002; Human-Centered AI Institute, 2019; Russell, 2019; Say, 2018). He is both a computer and cognitive scientist. He is the designer of the lisp program, which is used as a written language in the development of the first artificial intelligence programs. He defined “AI is the science and engineering of making intelligent machines” (McCarthy, 2007, p.1). Whereas Minsky defined this phenomenon as “[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]” Minsky (1968 as cited in Hernandez-Orallo, 2016).

Turing (1950) mentioned that in order to produce a program, we should design a machine which imitates a child’s mind instead of an adult’s mind. An adult’s mind has the following thinking processes; “a) The initial state of the mind, say at birth, b) the education to which it has been subjected, and c) other experience, not to be described as education (p. 455).” Then, he described a child’s brain by the analogy of “child-brain is something like a notebook as one buys it from the stationers. Rather little mechanism, and lots of blank sheets.” (p. 456) He inferred that the programmers cannot reach a well child-machine in the first try but one can try to educate it while experimenting and to see to what extent it learns and simulate a child’s thinking. He proposed that this was the way to teach the machine to be developed. Again, if a child learns a behavior at the most basic level and learns this behavior with the reward and punishment approached, the machines can be taught by increasing the probability of repetition of its intended output (for it, it's a behavior) by the same method. From the machine perspective, he says "the machine has to be so constructed that events which shortly preceded the occurrence of a punishment-signal are unlikely to be repeated, whereas a reward-signal increases the probability of repetition of the events which led up to it" (p 457). By following the basic studies of Alan Turing, the first artificial neural network-based computer named SNARC was performed by Minsky and Edmonds working at MIT in 1951.

In general terms, the word intelligence is defined as an inorganic tool which is able to simulate human thinking, reasoning, perception and comprehension. However, learning and adaptation to new situations are also included in intelligence. Therefore, we can call artificial intelligence a system that does not like humans, but it has the ability to think as human. In another definition, people are communicating with each other, reasoning, decision making, thinking, so some similar behaviors are considered to be taught to computers in the sub-field of computer science called artificial intelligence (Allahverdi, 2002; Karadayı, 2004; Russell, 2019). However, some problems arose after the problem-solving systems emerged. For example, the fact that these programs only work by syntax cannot make sense in the context. Therefore, it is believed that it cannot be used to solve real life problems. As Russell (2019) highlighted the more we discover how the mind works via experimental process and scientific research and innovative tech tools, the easier it will be to rediscover the mind's skills and competencies so that we will have taken a step towards artificial intelligence. It's skill, not consciousness that counts. The design studies of artificial intelligence (i.e., narrow intelligence) systems that act rationally for humanity but in a way that maximizes the expected benefit require long processes.

As the investments increased with the development of the industry, the studies continued and the context problem was reconsidered. Scientists have designed expert systems that are specific to an area and can be an expert in that field. It was determined that it would provide more accurate operations and results if it was equipped with expert knowledge in a specific field. The term **expert system** has begun to be used. Expert systems as a type of artificial intelligence programming differ from it in some ways such that artificial intelligence systems are preferred to be used to explore misunderstood problems because they do not need algorithms to solve these problems. In that case, languages such as Lisp and Prolog are mostly preferred rather than traditional programming languages such as Pascal, Fortran, C. Some definitions for expert systems were as follows:

> An expert system is regarded as the embodiment within a computer of a knowledge-based component, from an expert skill, in such a form that the system can offer intelligent advice or make an intelligent decision about a processing function. A desirable additional characteristic, which many would consider fundamental, is the capability of the system, on demand, to justify its own line of reasoning in a manner directly intelligible to the enquirer. The style adopted to attain these characteristics is rule-based programming (As quoted in Connell, 1987, p. 221 as cited in Omoteso, 2012).

Expert systems differ from more traditional decision aids in two fundamental ways. First, they place emphasis on knowledge, typically generated as rules, rather than algorithmic solutions. Second, they provide access to this knowledge base to the user of the decision aid. In addition, sophisticated expert system software gives numerous capabilities for enhancing the dialogue between the user and the system'' (Eining et al., 1997, p. 5 as cited in Omoteso, 2012).

All in all, the importance of having enough knowledge on the history of artificial intelligence, expert systems and its development over time from an educator's eye is undeniable. Since, **knowing the past means designing the future correctly**. Knowing its mistakes and deficiencies, machines based on artificial intelligence (or expert systems) will be designed for our children to measure and evaluate them most accurately.

The next title relied on what are the expert systems needed for teaching, learning and assessment in the educational context.

2.8.2. Current trajectory of AI in education

Artificial Intelligence (AI) system developers use information processing or reinforcement learning teaching approaches. Interestingly, nearly all educators are familiar with these approaches because they are the basic teaching approaches.

Many studies including master theses and doctoral dissertations conducted in Turkey have been on AI in Education. According to preliminary studies in the 2000s, Demir (2004) developed an artificial intelligence software to be used in computer-aided instruction in the master thesis and this software was evaluated by experts. This program was at the most basic level at that time; The aim of this course is to teach how to open and save Microsoft Office Word files with user guidance by detecting Turkish words. In limited circumstances he has imitated the characteristics of a teacher. In the first phase of the study, the researcher designed the artificial intelligence program of a particular feature. The second stage of her research process was about evaluation by academics, having at least a doctoral degree, from various universities in Turkey. Evaluation contents were specified as general features of the program, instructional feature, whether having appropriate artificial intelligence feature and screen design feature. At the end of the study, experts were conducted on a 56-item questionnaire. According to the results of this study, the program has an experimental character and exhibited artificial intelligence at a sufficient level at that time.

In similar years, Karadayı (2004) examined the developmental characteristics of pre-school children in the context of learning theories and investigated the contribution of computer-based instruction to child development. In addition, he reviewed computer-based pre-school education, how to use artificial intelligence methods, examples and problems revealed the use of. The situation in Turkey has been examined on two factors as hardware and software, and suggestions were made on them. Finally, the computer-based pre-school education studies in Turkey were presented. Some of his conclusions revealed that computer education to preschoolers should be started with games and possible problems should be given by early intervention. It should be kept in mind that computer-based pre-school education is just a tool and should be used for the right purpose. In 2014, in addition to focusing on solely programming approaches, Erümit (2014) studied the effect of artificial intelligence-based learning environments developed in accordance with Polya's problem solving steps (1957, 1973, 1990) on students' problem-solving processes because he grounded this framework on the idea that children should be good problem solvers to be good mathematicians and to keep mathematical thinking active. For this purpose, Erümit identified the most difficult situations students faced in the problem-solving process. Then, an artificial intelligence-based distance learning environment was made to eliminate these difficulties. System graph theory was used in the design process of AI. Logical inferences were made by the forward and backward chain method. The study was designed on the movement problems of 9th grade mathematics and conducted on 60 students in Trabzon province, Turkey. Within the scope of pilot study, the first system and interface were evaluated by taking student and expert opinions. In the main study, ARIMAT (artificial intelligence and math) was applied to the groups by using pretest posttest in a quasi-experimental design. Students and teacher interviews, and field notes were also used in the process. The results of the study showed that the students have developed problem solving methods, have increased their academic success. The ARIMAT provided a significant benefit to the teachers by successfully carrying out their assessment and evaluation activities.

In 2015, Schmoelz and his colleagues researched an intelligent tutoring interface which is based on inquiry-based (IBL) and multi-stage learning (MSL). They proposed how learning activities and pathways, are validated. Adaptive e-learning systems were used. In order to be able to use the knowledge as meta-data for the system, they explained the Pedagogical Ontology (PO), Web-Didactics (WD) metadata, ontology writing language (OWL) and pedagogies to link pedagogy and technology. They stressed the usage of these programs. They provide the flexibility such that predefined sequences can automatically adaptive to learner behavior. For this purpose, in their studies the researchers described the IBL and MSL with the combination of their metadata. They explained what the main differences between computational IBL might be and MSL.

They discussed that until now computers cannot react to semantically rich and individual questions as a requirement of the IBL approach. They cannot understand and react to them through students' thinking. Therefore, in the INTUITEL project working with the structured IBL pathway, the teacher submits a sum of optional research questions and their student can pick the one according to his/her personal interest. They explained the limitations of their studies from IBL to computerized IBL as "the new system cannot read research questions that are novel to the machine." The system at the beginning needed more participation and dialogue between teachers and students so that it can be fed from semantically rich inputs (Schmoelz et al., 2015).

Polat et al. (2016), a professor of mathematics at Yaşar University in İzmir, Turkey conducted an artificial intelligence study to evaluate students' responses to the test technique used in the assessment of teaching. They believed the importance of integrating AI to educational support systems such as learning management systems. For this purpose, they called it "supervised learning models." It was a concept mapping project on Intelligent Tutoring Systems in which Polat and colleagues worked on an intelligent system design that understood the content of a teaching document, drew concepts, and automatically identified the relationships between these concepts. This system with its most distinctive feature was able to determine the shortcomings and direct students to the right place for tutoring. The program they developed found the words given as input and the words that were closest to it. The system acquired artificial intelligence and language processing skills. The intelligent tutoring system was designed to share "what to teach" to students in the form of a map instantly (BTTO, 2017; Günel et al., 2016; Interpress, 2015; Polat, 2016; Sözcü, 2015).

Some researchers agreed that AI and mathematics have interrelated branches (Garrido, 2010; Günel et al., 2016). Problem solving is not only a sub-topic of mathematics but also a part of daily life skills (NCTM, 2000). The better our children can solve the problem, the better they can deal with problems and develop critical thinking skills in their lives. Thus, both in the sub-topics of mathematics and in other areas of the subject, problem solving must be involved. These skills are measured and evaluated intensively via both national and international large-scale assessments.

When all of these studies were examined, mathematical thinking processes, analogy-based classroom teaching tools in math (Besold & Kühnberger, 2014) were taken into consideration, the background of learning theories as "reinforcement learning" were used while developing a preliminary expert system. Algorithms were developed and evaluated by using them on the students in educational context. However, although some developed AI programs (or expert systems) have been evaluated in the research process, AI studies are relatively low specifically in the measurement, evaluation, and assessment perspectives (Arieli-Attali, et al., 2019; Falmagne, et al., 2006; Hand, 2004; Heffernan et al., 2006; Hernandez & Orallo, 2016). The important AI classroom assessment programs have been ASSISTment project developed collaboratively by the Worcester Polytechnic Institute, MA, United States and Carnegie Mellon University (Heffernan et al., 2006) and ALEKS. ASSISTment project is to provide cognitively based assessment of students while tutoring them. At the same time this program was created with the adaptation of OE question formats and provided assessment through students report to teachers. Instead, ALEKS [Assessment and Learning in Knowledge Spaces] (Falmagne et al., 2006) was developed by Falmagne and colleagues in the University of California, Irvine. It was an intelligent tutoring system to teach introductory statistics course outside class. It was based on active learning. The majority of ALEKS problems were OE rather than MC through which students gave authentic input appropriate to the discipline (McGraw Hill Education, 2017).

While using interactive and assessment programs, the evaluators and educators can collect more process data on students' strategies, cognitive and motivational aspects (Arieli-Attali et al., 2019). Grasping and exploring the fact that what kind of variables and inputs are collected from human-being and how to develop these AI systems seem to be significant in terms of repeating, developing and even sailing new horizons regarding the future of AI in education. For this reason, in this doctoral thesis research, a measurement-oriented framework was drawn especially in education.

2.8.3. Future of artificial intelligence in education

Just as IQ tests were used to measure human-being intelligence, discussions and measurement studies on the intelligence of AI systems have begun. For example, using an IQ test framework, Google's AI showed that it has an IQ of 47.28 based on the tests throughout 2016. From this result, Google deduced that Google's IQ score is relatively the same as a six-year-old human's IQ score. Liu et al. (2018) focused that higher the IQ score, the higher possibility to approach a human-being and its thinking while working in the industry.

From automated keyword estimation (Driscoll et al., 1991) to assessment cognitive and metacognitive skills (Conati, 2016), AI developments in the industry and especially in educational context showed us a dramatic increase in AI systems, in other words expert systems. In fact, the development of systems that can imitate human-based rather than logic-based systems, in accordance with a specific rule, will play a true role in solving real-life problems. The development of artificial intelligence through the monitoring and observation of people and how children learn throughout history, the machines have been taught in the same way. Similarly, in the future it is expected that expert systems will be designed by imitating their behavior in accordance with the data received from human-being. Until now, throughout literature review it has been found that teaching and learning approaches (instructional strategies) used in AI development: reinforcement learning, information processing, inquiry-based learning, case-based reasoning, multi-stage learning, cooperative learning, active learning, self-explanation. An expert AI has the following basic features: must be programmable and adaptable, clearly reveal the relationship between concepts and the concept space that logical inference will make, the mechanism of result output must be robust and complete, compliance between program type and algorithm should be equal to being feasible (Erümit, 2014).

To add, Conati (2016)'s study was on student-adaptive learning experiences. The goal was to develop an intelligent learning environment (ILE) [called as SE-Coach] which can support an adaptive tool working with domain independent meta-cognitive skills rather than domain-dependent knowledge. They designed a system which can automatically monitor students when they study examples and provide them with adaptive interventions so that they can make self-explanation. It means that this study focused on not only cognitive aspects but also metacognition in terms of self-explanation. In this process the aim of self-explanations was to help students improve their domain knowledge. In other words, the key innovative point of this study was the system they designed personalized to student individual differences both at the cognitive level (e.g., existing knowledge) and meta-cognitive level (e.g., tendency to self-explain). The feature of the system has two folds: "1) dependency network that models how each solution step derives from previous steps and 2) interface tools to scaffold the target self-explanations" (p. 185). The SE-Coach was evaluated in a controlled study with 56 college students while they are attending an Introductory Physics course.

As some researchers highlighted, the progress on theoretical issues can be approached by using analytical tools and techniques from deep data models as a subset of AI. Analysis of environmental requirements is a must. Thinking about the ways to implement some features in a robot begins with generating a specific and testable hypothesis for behavioral tests (Conati, 2016). As such, if we want to develop a deep data model that will meet the needs of the target audience such as students and to serve humanity easily, we must first start collecting from people and we must measure their behaviors correctly.

Figure 2. 1. A Summary of the Literature Review related to Curriculum Change, Assessment Change, Metacognition and Affect in Neuroeducation

In the light of the abovementioned literature review, it has been reiterated that the main purpose of this research was to examine the preferences of middle school mathematics teachers' teaching method and measurement-evaluation processes after the curriculum policy change in Turkey through an ecological approach and the quality of middle school students' responses and responses to different items types, using multimodal mixed methods concurrent dominant status design. While setting out to achieve these aims, the literature scrutinized at all the comprehensive phenomenon one by one and does not approach it holistically yet. For example, individual studies examining the quality of teachers' in-class examination items, what kind of changes and adaptations they could make in their teaching methods and measurement-evaluation strategies are mostly found to use qualitative research methods. In addition, individual studies examining students' readiness and what kind of cognitive and affective sub-skills they could respond when exposed to qualified and higher-order item types use qualitative and quantitative approaches separately. With the improvement of innovations, those studies were supported by multimodal tools in the current research. Based on these wisdoms, the research questions attempted to be answered with deep data analysis through utilizing a mixed-methods approach with melting different point of views.

CHAPTER 3

METHOD

In the method section; research design, research questions, research context, research settings, population and sampling procedures, data collection instruments, data collection procedures, data analysis, trustworthiness, validity and reliability of the study phases are explained for each phase. The chapter ends with researcher experience.

3.1. Research Design

By the nature of the aim of this study and research questions, the study had a mixed methods research design. A mixed methods study involves "the collection or analysis of both quantitative and qualitative data in a single study in which the data are collected concurrently or sequentially, are given a priority, and involve the integration of the data at one or more stages in the process of research" (Creswell et al., 2003, p. 212). The current research is based on mixed methods research design named with an innovative concept displaying complexity of research design as *Multimodal Mixed Methods Concurrent Dominant Status Design* as an adaptation of Creswell (2018)'s *Partially Mixed Concurrent Dominant Status*. The nature of the study began with the document analysis of in-class authentic teacher-made items that dealt with qualitative approach. Then, research flow required the quantitative approach to explore middle school mathematics teachers' preferences through a questionnaire. Finally, the research was followed with explanation of 5th grade students' reactions and responses to different item formats that required collection both qualitative and quantitative data in the research procedure. In the final phase, multimodal tools (e.g., eye-tracking, wacoom bamboo tablets, smart watch) triangulated the data for multiple modes of communication with fifth graders reactions and responses to item formats.

The research flow is seen in Figure 3.1. and the integration process of the qualitative and quantitative results in Figure 3.2. As depicted in the Research Flow, I started with document analysis of teacher-made items (i.e. Phase 1) in which mathematics teachers were the study group, then I continued with a quantitative survey (i.e. Phase 2) in which I authentically developed a questionnaire for teachers and administered to 350 mathematics teachers. After that, I extended the steps with multimodal phase to examine middle school students' metacognition and affective processes via qualitative reflection and quantitative evaluation (i.e. Phase 3 and Phase 4). Lastly, I integrated the findings to draw a deep data model.

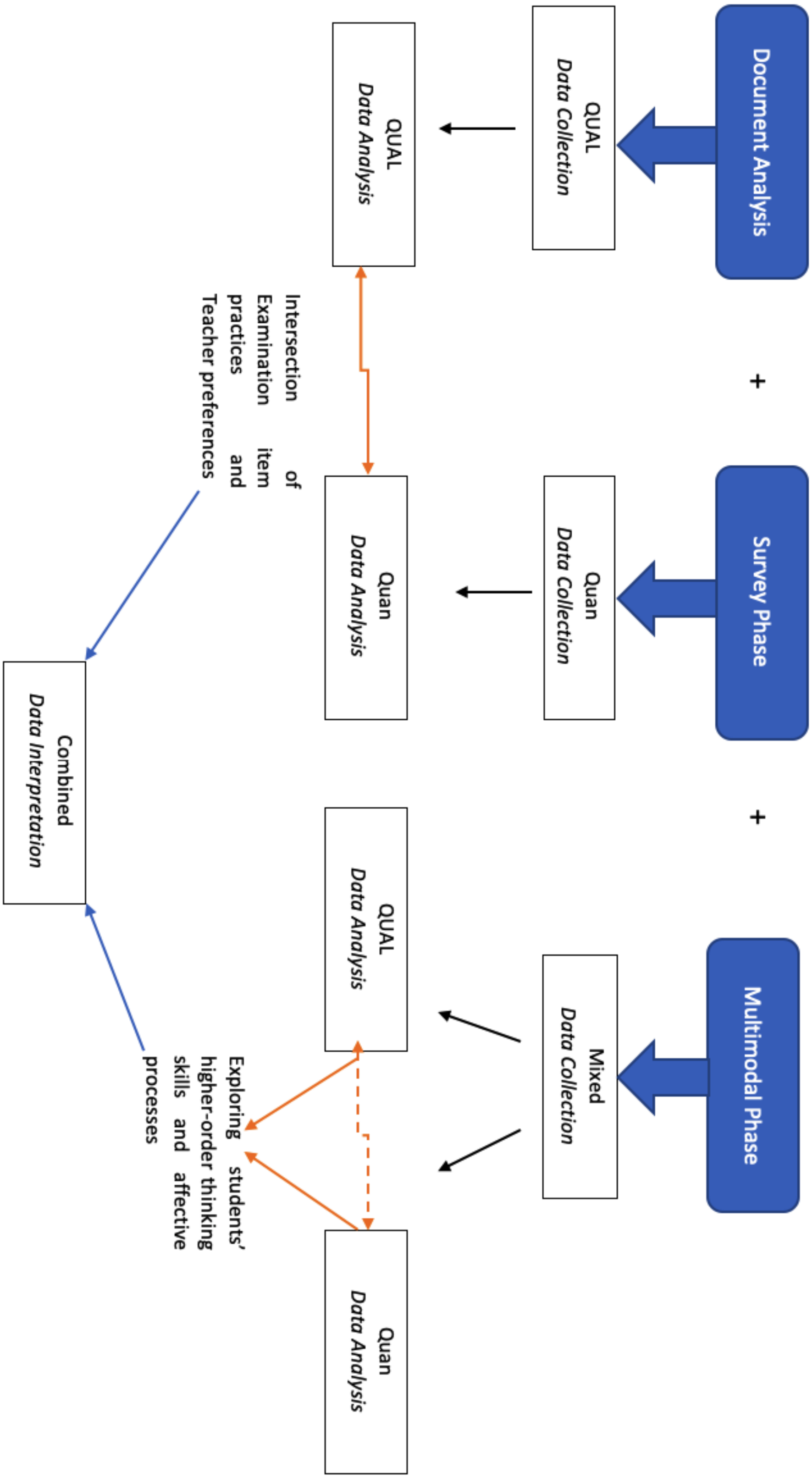


Figure 3. 1. Research Flow

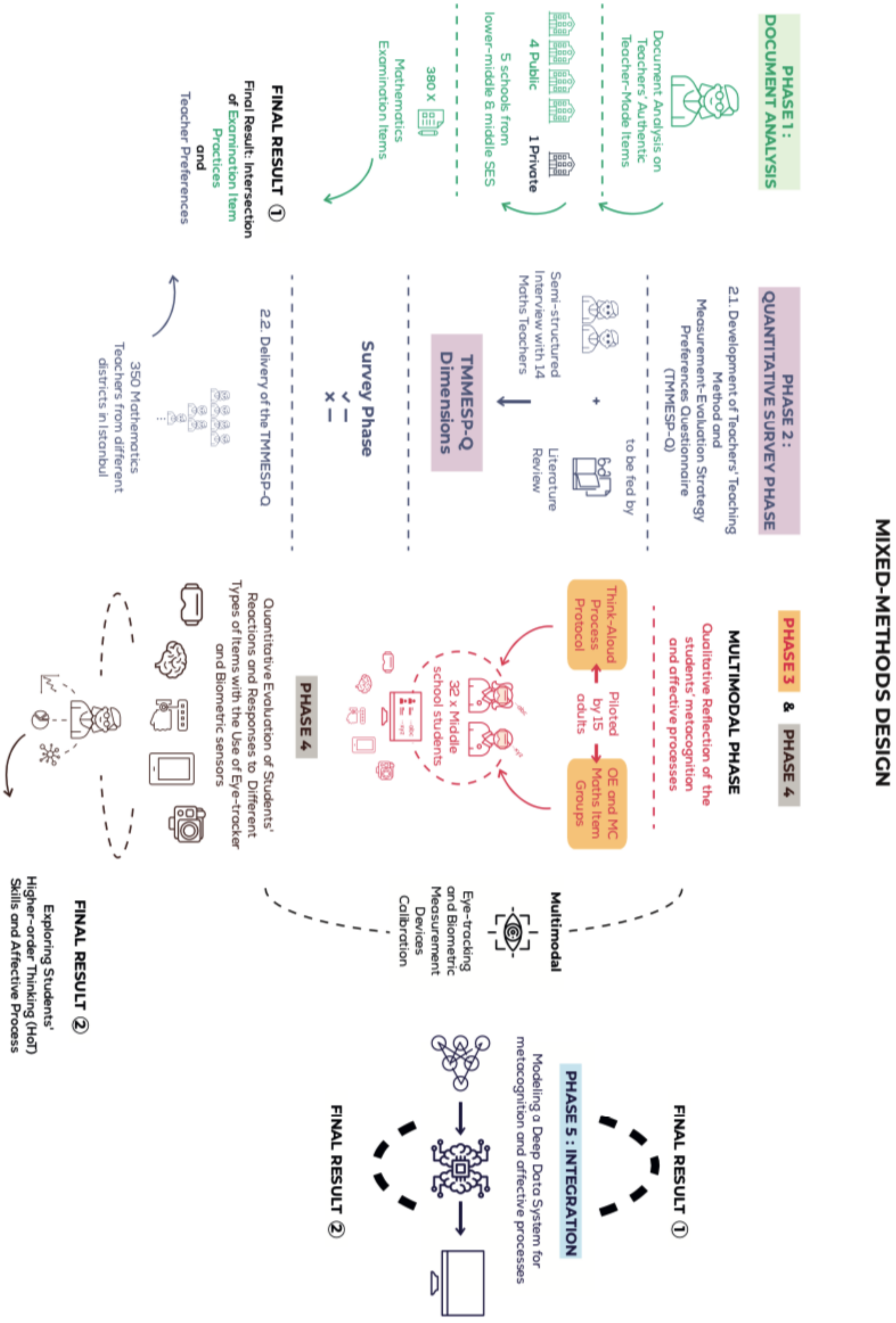


Figure 3. 2. Integration Process of Mixed-methods Research

3.2. Research Questions

The study relied on three big dimensions and five phases. Its phases were *Phase 1*. Document Analysis: Examination of authentic teacher-made items; *Phase 2*. Quantitative Survey: Investigation of teachers' teaching method and measurement-evaluation strategy preferences; *Multimodal Phases* including *Phase 3.* Multimodal Phase: Reflection of students' metacognition and affective processes to different types of questions; *Phase 4.* Multimodal Phase: Evaluation of students' reactions and responses to different types of questions with the use of eye-tracker and biometric sensors; (Neuroeducation Process); *Phase 5.* Integration: Modeling a Deep Data System for metacognition and affective processes.

The following research questions guided my study:

I aimed to explore 1) To what extent is the enacted middle school mathematics curriculum (dated 2015/valid until the end of 2016-2017 academic terms) compatible with the proposed assessment procedures within the curriculum in preparing middle-school students for learning outcomes?, 2) Do teaching methods and measurement-evaluation strategies used by the middle school mathematics teachers in the classroom after the mathematics curriculum change (in 2018) compared to those used for the previous curriculum (dated 2015)?, 3) How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types? Is there a significant difference between the amount of reflection of students' metacognitive skill levels on their responses to multiple-choice and open-ended items?, 4) What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?, 5) What neuro/biomarkers are needed to measure students' responses to open-ended items to evaluate their metacognitive (cognitive strategy and self-checking) and affective processes (worry and effort) through a deep data modeling?

3.3. Research Context

Based on the *multimodal mixed methods concurrent dominant status design*, the present study began at the fall semester of the 2017-2018 academic year. Then, the research process continued in the 2018-2019, 2019-2020, 2020-2021 academic semesters which were detailed in this study.

3.3.1. Research setting(s)

Data were collected in the Sarıyer district of Istanbul where I informed school administrators about the research process and shared the sample instruments with them. There were 8 schools in total where teachers are willing to contribute to the study. These were 2 private and 6 public schools. A foundation school informed that they have an intensive busy schedule in school tasks. Therefore, they were asked to be withdrawn from the study. It was decided to work with another private school instead of the former one.

There was a total of 89 schools in Sarıyer, 33 184 students and 2 324 teachers. There were 50 middle schools and among them, there are 37 public middle schools, one of them is for visually impaired students and 13 private middle schools. The public schools as representative of each unit of analysis in neighborhoods were also selected as purposefully as high, middle, low according to end of year success/TEOG results. Sarıyer district, Istanbul, Turkey is representing upper, middle and lower-middle class SES levels according to The Government of Istanbul Socio-Economic Analysis (2020). Mathematics teachers and their 5th grade students from these schools participated in the study. Mathematics teachers in Phase 1 participated in the study from the district. Mathematics teachers of survey development process in Phase 2 also participated in the study from the district. Mathematics teachers participated in the survey delivery process in Phase 2 from various districts in Istanbul. Middle school students who participated in Phase 3 and Phase 4 were from three districts (e.g., Beşiktaş, Fatih, Sarıyer) in Istanbul since there was an announcement conducted. The students from similar SES districts were also welcome. Ultimately, neuroeducation lab study with the volunteer students took place in Brain Dynamics Laboratory settled in a foundation university, Sarıyer district. I took neuroeducation settings apart below in section 3.3.3.

3.3.2. Population and sampling procedures

The sampling strategies employed for each phase of the study are explained successively based on the phases of the study.

a. Phase 1. Document analysis

The sampling strategy was purposeful sampling. "The power of purposeful sampling lies in selecting information rich cases for study in depth" (Patton, 2002 p. 169). The unit of analysis was each school settled in the neighborhood of Sarıyer District (see. tables in Appendix B and C). There are 38 districts in Sarıyer. The unit of analysis for this study was reaching a public school in each neighborhood because the unit of analysis was the major entity that I was analyzing in the study and the level of generalization. If more than one school in a neighborhood, the schools were selected as a rate of ½ to be representative. Major parameter was number of public and private schools. One private and 4 public schools volunteered to participate in the Phase 1.

b. Phase 2. Quantitative survey

The sampling strategy utilized for the survey was purposive and convenient sampling. A convenience sample is "a group of individuals who (conveniently) are available for study" (Fraenkel et al., 2014, p. 99). It was purposive at the beginning because I had to interview with the teachers from the schools in

Sarıyer for survey instrument development before reaching out bigger sample size of mathematics teachers for survey delivery. For survey development, eight schools in Sarıyer were participated. After the survey development, the mathematics teacher for the survey delivery part were selected according to convenience sampling method.

Before the new academic semester of state schools started, I collected data from the mathematics teachers while delivering the survey. MoNE Director of Istanbul organized a seminar for all middle school mathematics teachers for their professional development. I learnt the day and time of this seminar from MoNE Director of Istanbul and after that used this opportunity to reach a larger sample size.

c. Phase 3 and Phase 4. Multimodal phase

The sampling strategy was purposeful sampling (Fraenkel et al., 2014) and then snowball sampling to gather support from participants and their families. I made an invitation announcement to candidate schools, administrations, and parents (see Figure 3.3 for invitation poster) with parent consent forms. I would not simply study whoever is available but rather use my judgment to select a sample that I believe, based on prior information, would provide the data I need. For these phases, I aimed to work not only with regular fifth grade children allowed by their schools and their parents, but also with talkative ones who are also awareness of research, who did not have any optical problem, of contributing to the study.

Figure 3. 3. Invitation Poster to Neuroeducation Study

d. Introductory of schools participating in the research phases

***Private Schools in Sarıyer*.** Two private schools participated.

School 1. Philosophy: the world of tomorrow; population growth, economic and social dynamics created as a result of the rapid depletion of natural resources brings more rapid and compelling changes. Schools should embrace these challenges and change; students should be educated as individuals who will manage and analyze these phenomena without compromising the universal values. In School 1; we believe in the power of knowledge, the unification of universal values, and our responsibility to improve the existing conditions of human life and nature. Values: Honesty (Consistency, fairness and accuracy), Respect (to other people, cultures, living things, environment and individual differences), Responsibility (at personal and social and social levels), Productivity (individual who adds value to society), Awareness (in terms of personal and human conditions). Moreover, *School 2 is* developed by a public University Faculty of Education, the education is implemented by expert educators, who monitor the students individually and regularly inform families and actively participate in the process. A contemporary, original and creative education approach is applied. The courses in which questioning and investigative methods are followed are student-centered. Students construct and create information themselves in a democratic environment. They play an active role in obtaining information; they question, investigate and reconstruct information using their creativity. They do this by comparing them with the information they have previously acquired and retained in their mind.

Public Schools in Sarıyer. Six public schools participated.

There is overarching national aim of education among public schools. Their common vision is being in line with the principles of the *Basic Law of National Education*, which adheres to the principles and reforms of Mustafa Kemal Atatürk (our founder), integrates with technology and adopts democratic life; The principle is to raise generations of intellectual and conscience, who have a culture, who work, work, give importance to reason and science. Their mission is to raise knowledgeable, resourceful and self-confident individuals, to raise individuals who have the power of decision-making of democratic life, which derives from country values.

School 3 mission is to provide qualified education to prepare all children of compulsory education age for life and higher education. There are 24 teachers, 1 guidance and psychological counseling teacher and 378 students. Art and cultural activities such as Drama course, Ebru workshop, Ceramic workshop and English Theater Works are also held in this school. *School 4* prepares all our students as secular, democratic, future-oriented, happy individuals who are open to research, innovative, learning using technology, looking to the future, hopeful, happy individuals. in. They aim at educating respectful, democratic, peaceful, and self-confident students and to be the highest quality school of the future. There are 40 teachers, 2 guidance and psychological counseling teachers, 1073 students and 29 classrooms. *While in School 5,* there are 35 teachers, 565 students and 21 classes; in *School 6,* there are totally 23 teachers, 624 students and 22 classrooms. *In School 7,* it is stated that there are totally 72 teachers, 3 guidance and psychological counseling teachers, 1190 students and 18 classrooms, whereas, in *School 8,* there are 31 teachers, 348 students, and 20 classrooms. It gives importance to educate pupils as self-confident, successful, principled, sensitive, open-minded, investigative-questioning, improved communication skills, self-evaluating, collaborative, creative and critical thinking, open to learning and innovation, intellectual, courageous, open to multi-faceted, assimilated national culture to educate globally thinking individuals.

The participants of the overall research in line with research phases were described in Table 3.1.

Table 3. 1. Flow and Participants of the Study

Phases	Research Questions	Study Group	Data Collection Process	Experts Opinion (Prof., Assoc. Prof., Assist. Prof., expert in the field)
DOCUMENT ANALYSIS **Phase 1.** Examination of authentic teacher-made items	1) To what extent is the enacted middle school mathematics curriculum (valid until the end of 2016-2017 academic terms) compatible with the proposed assessment procedures within the curriculum in preparing middle-school students for learning outcomes?	10 mathematics teachers from 5 different public (62%) and private schools (37%) located in lower-middle and middle SES districts in İstanbul, Turkey.	380 in-class authentic teacher-made examination items were collected	1.Advisor, the professor of department of educational sciences, curriculum and instruction program 2.An associate professor from the department of mathematics and science education 3.A PhD candidate from the department of educational sciences, measurement and evaluation program and research assistant in the department of mathematics and science education

Table 3.1. (continued)

Phases	Research Questions	Study Group	Data Collection Process	Experts Opinion (Prof., Assoc. Prof., Assist. Prof., expert in the field)
QUANTITATIVE SURVEY	2) Do teaching methods and measurement-evaluation strategies used by the middle school mathematics teachers	14 Middle School Mathematics Teachers	Semi-structured Interview	1.Advisor, the professor of the department of educational sciences, curriculum and instruction
Phase 2. Investigation of teachers' teaching method and measurement-evaluation strategy preferences	in the classroom after the maths curriculum change compare to those used for the previous curriculum (valid until the end of 2016-2017 academic terms)?	350 Middle School Mathematics Teachers (administered & returned)	Development and Delivery of Survey Instrument Teaching Method Preferences and Measurement-Evaluation Strategies Questionnaire [TMMESP-Q]	program 2.A professor the department of educational sciences, measurement and evaluation program
MULTIMODAL PHASE **Phase 3.** Reflection of students' metacognition and affective processes to different types of questions **Phase 4.** Evaluation of students' reactions and responses to different types of questions with the use of	3) How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types? Is there a significant	14 Young Adults for Pilot Study 32 5th Grade Middle School Students	Think Aloud Process Protocol MC and OE Mathematics Item Groups Eye-tracking and Biometric Sensor Tools	1.Advisor, the professor of department of educational sciences, curriculum and instruction program 2.Co-Advisor, The assistant professor of computer engineering & psychology, big data analytics,

Table 3.1. (continued)

Phases	Research Questions	Study Group	Data Collection Process	Experts Opinion (Prof., Assoc. Prof., Assist. Prof., expert in the field)
eye-tracker and biometric sensors	difference between the amount of reflection of students' metacognitive skill levels on their responses to multiple-choice and open-ended items? 4) What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?			information technologies and computing 3.Several expert opinions from the professors of the departments of measurement and evaluation, cognitive science, mathematics education

Table 3.1. (continued)

Phases	Research Questions	Study Group	Data Collection Process	Experts Opinion (Prof., Assoc. Prof., Assist. Prof., expert in the field)
INTEGRATION **Phase 5**. Modeling a Deep Data System for metacognition and affective processes	5) What neuro/biomarkers are needed to measure students' responses to open-ended items to evaluate their metacognitive (cognitive strategy and self-checking) and affective processes (worry and effort) through a deep data modeling?	32 5th Grade Middle School Students	Theoretical Design of a Deep Data Models with neuro/biomarkers	Computer Engineering, Cognitive Sciences, Measurement and Evaluation, Educational Sciences, Mathematics Education

Table 3. 2. Research Phases Related to Research Design

Phases	Research Questions	Design Types	Reason
DOCUMENT ANALYSIS **Phase 1.** Examination of authentic teacher-made items	1) To what extent is the enacted middle school mathematics curriculum (valid until the end of 2016-2017 academic terms) compatible with the proposed assessment procedures within the curriculum in preparing middle-school students for learning outcomes?	Qualitative Design	Analysis of in-class authentic teacher-made items when dealing with quality of items
QUANTITATIVE SURVEY **Phase 2.** Investigation of teachers' teaching method and measurement-evaluation strategy preferences	2) Do teaching methods and measurement-evaluation strategies used by the middle school mathematics teachers in the classroom after the mathematics curriculum change compare to those used for the previous curriculum (valid until the end of 2016-2017 academic terms)?	Quantitative Design	Dealing with numerical data collected by the TMMESP questionnaire
MULTIMODAL PHASE **Phase 3.** Reflection of students' metacognition and affective processes to different types of questions	3) How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types?		Dealing with words and meaning of 5th graders while reflection their metacognition and affective processes.

Table 3.2. (continued)

Phases	Research Questions	Design Types	Reason
MULTIMODAL PHASE **Phase 4.** Evaluation of students' reactions and responses to different types of questions with the use of eye-tracker and biometric sensors	4) What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?	Mixed Design (QUAL → Quan)	Dealing with numbers and statistics while exploring 5th graders reactions and responses to different types of questions through multimodal tools.

a. Participants of document analysis: Examination of authentic teacher-made items

To begin with the research, after the ethical approvals from Middle East Technical University and MoNE, I got in contact with all of the middle schools from Sarıyer and 5 schools (1 private and 4 public) accepted my invitation to participate in the introductory phase of the research. The authentic teacher-made examinations were collected from 10 middle school mathematics teachers of 4 public and 1 private schools in İstanbul. The distribution of schools and teachers were described in Table 3.3.

Table 3. 3. Schools and Characteristics of the Mathematics Teachers who delivered Examination Items for Document Analysis

School ID	School Type	Number of Examination Papers for pre-change	Gender	Years of Experience	Number of Examination Papers for after-change	Total Number of Examination Items
School 1	Public	2	F	11	7	117
			F	2		
School 2	Public	2	F	7	5	87
			F	5		
School 3	Public	2	F	4	2	45
			M	13		
School 4	Public	3	F	20	6	117
			F	11		
			F	15		
School 5	Private	1	F	5	1	14
TOTAL		10			21	380

Note: F = Female, M = Male

The study group of this study (see Table 3.2) were ten middle school mathematics teachers from public and private schools in Sarıyer, Istanbul. Due to İstanbul being a very complex metropolitan city, Sarıyer was selected as a study province because my affiliation has an already signed protocol with Sarıyer District Directorate of National Education. So, I was allowed to easily access the public schools to use as pilot schools.

The study group of this multimodal mixed methods research design was selected as purposeful sampling. “The power of purposeful sampling lies in selecting information rich cases for study in depth” (Patton, 2002). In this study, the unit of analysis was each school settled in the neighborhood of Sarıyer District (see. Appendix A). There are 38 neighborhoods in Sarıyer. The descriptive statistics for each school was shown in Appendix B and C. The unit of analysis for this study was a school in each neighborhood because the unit of analysis is the major entity that I am analyzing in the study and the level of generalization. If there was more than one school in a neighborhood, the schools were selected at a rate of ½ to be representative.

There is a total of 89 schools in Sarıyer, 33 184 students and 2 324 teachers. There are 50 middle schools and among them, there are 37 public middle schools, one of them is for visually impaired students and 13 private middle schools. The public schools as representative of each unit of analysis in neighborhoods was selected as purposefully as high, middle, low according to end of year success (i.e., TEOG results). (see. Appendix O for participating schools for teacher-made examinations)

Ethical Approvals from METU and MoNE were conducted as following:

- *Phase 1.* Examination of authentic teacher-made items (see. Appendix D for the ethical approval from METU and see. Appendix G from MoNE)
- *Phase 2.* Investigation of teachers’ teaching method and measurement-evaluation strategy preferences (see. Appendix E for the ethical approval from METU and see. Appendix H from MoNE)
- *Phase 3.* Reflection of students’ metacognition and affective processes to different types of questions (see. Appendix F for the ethical approval from METU and see. Appendix I from MoNE)
- *Phase 4.* Evaluation of students' reactions and responses to different types of questions with the use of eye-tracker and biometric sensors;

(Neuroeducation Process) (see. Appendix F for the ethical approval from METU and see Appendix I from MoNE)

- *Phase 5.* Modeling a Deep Data System for metacognition and affective processes.

b. Participants of the survey: Investigation of teachers' teaching method and measurement-evaluation strategy preferences

*Survey Development study***.** A total of 14 middle school mathematics teachers from two private and six public middle schools in Sarıyer District in Istanbul participated in the Phase 2 of the study. Only two teachers were participants from one of the top-level private schools in Sarıyer District. Participants were 11 female (79%) and 3 male (21%) between the ages of 28 and 42. Their year of seniority in teaching ranged from 2 to 20 years. Most of the teachers (n = 11) graduated from the department of mathematics education (middle school). Only three of them graduated from the department of mathematics (arts and science), yet they hold teaching certification after the graduation. Most of them were teaching two grade levels (e.g., 5, 7 or 5, 6) while few of them were teaching more than two grade levels (e.g., 5, 7, 8).

I explained the aim of this stage of the thesis and asked permission to study and interview with the mathematics teachers, and then observe their classes. These 14 mathematics teachers who volunteered to participate in the study are as follows (Table 3.4):

Table 3. 4. Characteristics of the Mathematics Teachers who were Interviewed for the Development of TMMESP-Q Items

Teacher #	The Teacher's Code	Gender	School Type	School Numbers (Middle Schools)	Year of teaching experience	Years in the school
1	Teacher A	F	Public	School 3	20	10
2	Teacher B	F	Public	School 3	11	4
3	Teacher C	F	Public	School 4	2	1
4	Teacher D	F	Public	School 4	11	7
5	Teacher E	F	Public	School 7	4	1
6	Teacher F	M	Public	School 7	13	2
7	Teacher G	F	Private	School 1	3	3
8	Teacher H	F	Public	School 5	11	1
9	Teacher I	F	Public	School 5	7	1
10	Teacher J	F	Public	School 5	5	3
11	Teacher K	F	Public	School 6	6	4
12	Teacher L	F	Private	School 2	5	2
13	Teacher M	M	Public	School 8	6	3
14	Teacher N	M	Public	School 8	11	2

*Survey Delivery Study***.** The participants for the survey study were selected based on a convenient sampling method. Firstly, the data were collected by Teaching Methods [TM] and Measurement-Evaluation Strategy [MES] Preference Questionnaire (TMMESP-Q) developed by the researcher. The independent variables for descriptive were 350 middle school mathematics teachers' gender, age, school type, branch, graduation program, the university of pedagogical formation certification program (if applicable), year of pedagogical formation, period of in-service education, year of seniority in teaching, year of working in their school, grade level of teaching, and classroom size.

The descriptive statistics regarding 350 mathematics teachers were analyzed in Table 3.5.

Table 3. 5. Descriptive Statistics of Teachers who Filled TMMESP-Q in the Main Study

Characteristics	TM_{Total}[a]		MES_{Total}[b]		TOTAL[c]		
Gender	*f*	*M(SD)*	*f*	*M(SD)*	*f*	*%*	*Cumulative %*
Female	203	80.04 (6.27)	224	51.08 (6.70)	246	70.3	70.3
Male	93	77.47 (6.71)	99	51.27 (7.13)	103	29.4	99.7
Age							
20-29	107	79.62 (6.05)	117	51.89 (6.32)	123	35.1	35.1
30-39	130	78.07 (6.62)	138	50.67 (6.90)	151	43.1	78.3
40-49	44	81.11 (6.86)	51	50.57 (7.41)	55	15.7	94.0
50+	16	80.94 (6.06)	18	51.39 (7.63)	21	6	100.0
School type							
Public	176	78.63 (6.78)	206	51.42 (7.04)	217	62	62
Private	109	80.14 (5.86)	117	50.77 (6.26)	132	37.7	99.7
Branch							
Elementary mathematics	77	78.36 (6.78)	84	52.02 (6.57)	85	24.3	24.3
Mathematics	218	79.58 (6.35)	238	50.89 (6.87)	262	74.9	99.1
Graduation program							
Mathematics education	203	78.71 (6.65)	220	51.49 (7.15)	233	66.6	67.0
Arts and sciences	93	80.47 (5.96)	102	50.51 (5.90)	115	32.9	100.0
PFCP							
31 different faculty of education	-	-	-	-	104	29.7	29.7
NA	204	78.65 (6.62)	221	51.51 (7.23)	234	66.9	96.6
TOTAL					350	100.0	100.0

[a]The total of TM part of TMMESP-Q is 100.
[b]The total of MES part of TMMESP-Q is 75.
[c]The total of participants who was able to fill out both parts.

According to the descriptive statistics of teachers which is reported in Table 3.4, number of female mathematics teachers (*f* = 246, 70.3%) are higher than the number of male mathematics teachers (*f* = 103, 29.4%). Most of the teachers' age were between 30-39 (*f* = 151, 43.1%) and 20-29 (*f* = 123, 35.1). The number of teachers who are working at public school (*f* = 217, 62%) than the number of teachers who are working at private schools (*f* = 132, 37.7%). The number of teachers who graduated from the department of mathematics education were higher (*f* = 233, 66.6%) than those who graduated from the department of arts and science (*f* = 115, 32.9%). The teachers who graduated from pedagogical formation certification program (PFCP) (i.e., called as Teaching License Program in the U.S. A) indicated 31 different names (e.g., Abant İzzet Baysal, Atatürk, Çanakkale Onsekiz Mart, Fırat, Gazi, İstanbul, İTÜ, YTÜ etc.) of faculties of education in Turkey (*f* = 104, 29.7%). Most of the teachers teaching in more than one grade levels such as 5th and 6th (*f* = 34, 9.7%), 7th and 8th (*f* = 44, 12.6%), 5th -6th- 7th and 8th (*f* = 56, 16%).

According to the descriptive statistics of continuous variables in Table 3.4, the average time of having PFCP was around 1 year (*M* = 1.37 *SD* = .83). Nevertheless, nearly all of them did not know the time of their in-service teacher education (per month, week, day or hour). Their seniority year in mathematics teaching was around 11 years (*M* = 10.90 *SD* = 7.53) whereas their years passed in the schools until the data collection process was found to be 4 years (*M* = 4.09 *SD* = 3.68). It was seen that most of the participating teachers are experienced in the profession. The results indicated the teachers' weekly course hours were 25 (*M* = 25.05 *SD* = 8.12) and the average number of students they have in their classes were 30 (*M* = 29.63 *SD* = 8.62).

c. Participants of multimodal phase: Neuroeducation

This section refers to Phase 3: Multimodal Phase: Reflection of students' metacognition and affective processes to different types of questions and Phase 4: Multimodal Phase: Evaluation of students' reactions and responses to different types of questions with the use of eye-tracker and biometric sensors.

The participants for the multimodal phase were selected based on a purposeful and then snowball sampling method. Among middle school students who were neuroeducation study participants, 17 children were female (53%) and 15 were male (47%). All of them were 10 years old and 5th graders in the public and private middle schools. Three students were studying in a private school in Ulus, Beşiktaş (9.38%), 9 out of 32 were studying in a public school in Maslak, Sarıyer (28.13%), 3 were studying in a public school in Kocamustafapaşa, Fatih (9.38%), 2 were studying in a private school in Tarabya, Sarıyer (6.25%), 2 were studying in a public school in Etiler, Beşiktaş (6.25%), only one student was studying in a private school in Levent, Beşiktaş (3.13%), and 6 were studying in another public school in Kocamustafapaşa, Fatih (18.75%). One student was studying in Çapa, Fatih, (3.13%) and one student was studying in Ferahevler, Sarıyer (3.13%), 3 were studying in a public school in Fatih (9.38%), only one student was studying in a private school in Üsküdar (3.13%). In total, nearly one fifth (22%) of the students were studying in private schools whereas most of them were studying in public schools in Beşiktaş, Sarıyer, and Fatih.

The districts where the students schooling was divided into three regions; 5 out of 32 students were studying in Beşiktaş (16%), 13 were studying in Sarıyer (40%), and 14 were studying in Fatih (44%) districts. The maximum total time on task they used while solving mathematics items was 31 minutes and the minimum time was 9 minutes. According to time on test data, the most time spent while responding to items was the 3rd item (Procedural, Evaluating) with 2.9 minutes while the question with the least time was the second question with .8 minutes.

3.3.3. Neuroeducation settings

Neuroeducation is an applied field. The Brain Dynamics lab is located at the end of the corridor on the fourth floor of Mef University (Figure 3.4). This laboratory uses various high quality equipment such as Gaze Point 3 HD Eye Tracker 150Hz, g.Nautilus PRO 32-Channel EEG/ERP, Gazepoint 3 HD Biometrics (GSR) System, Gazepoint 3 HD Biometrics (GSR) System, MSI GT75 Titan 8RG-092 TR i9-8950 HK 32 GB 1 TB + 512 GB (2X256) SSD GTX 1080 17.3" Full HD Notebook, Lenovo ThinkStation P330 Tiny Business Desktop Black (Intel core i7-8700T 6-Core, 16GB RAM, 512GB SSD, Quadro P620, WiFi, Bluetooth, 5xUSB 3.1, 1xHDMI, Win 10 Pro) to enhance the accuracy of the scientific results. The initiatives that will be completed in this lab will be interdisciplinary academic research projects. Numerous academic fields, including linguistics, computer engineering, philosophy, industrial engineering, and psychology, contribute to Neurolab. The major areas of interest include cognitive neuroscience, human-computer interaction (HCI), brain-computer interaction (BCI), user experience research (UX), neurolinguistics, learning, and applied neuroscience (neuroeconomics, neuroergonomics, and neuromarketing).

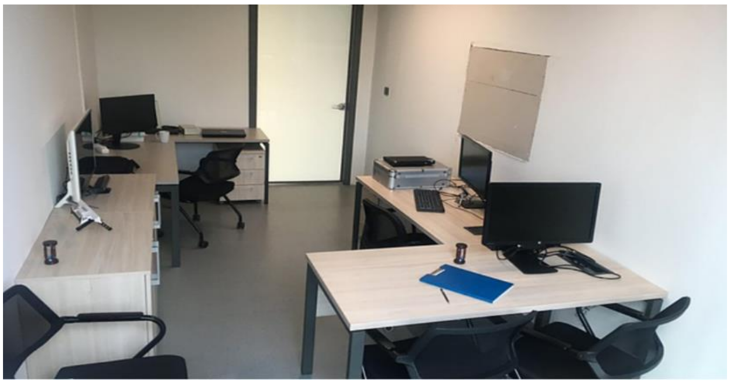


Figure 3. 4. MEF University Brain Dynamics Laboratory

Observation, surveys, interviews, and questionnaires are commonly used in current practices. Even if these methods have unquestionably improved our insight into students' educational experiences, they are limited to the student's memories and perceptions about what researchers can "externally" observe. Furthermore, data collected throughout this way can only be analyzed after the learning experience has ended, obviating the opportunity to scaffold students' learning in real time through feedback mechanisms (Giannakos, 2021). On the other hand, eye-tracking technology and wristbands allow for unobtrusive, continuous, and automatic data collection during students' experiences without self-report bias. For this reason, we decided to use Empatica E4 wristband (Figure 3.5) and Eye Tracker (Figure 3.6) in addition to the semi-structured interview for our research. The combination of gaze and physiological measurements (from eye trackers and wristbands, respectively) were used simultaneously while students were solving math problems.

Data collection occurred at the Brain Dynamics Laboratory at MEF University. First, participants were provided with informed consent. Parental approval was provided for all children. After providing consent, the E4 wristband was placed on the participant's non-dominant wrist, then data collection devices such as eye tracker, wristband, webcam for face gestures tracking system are connected and calibrated. The eye tracker was placed on a table that was adjusted to the height of the child's seat, and the child was positioned in a comfy chair at a viewing distance of 65 cm. The participants also had noticeable head movement freedom, which made the test environment feel more natural. Calibration procedure was completed before beginning the eyetracking. EDA and BVP data are stored on the E4 connect website (Figure 3.7), the E4 sensor is counted as a session from the time it is started until the time it is turned off. The Wacom Bamboo Slate (Figure 3.8) tablet, which has the ability to convert handwritten notes and sketches into digital files, was given to the children, and after writing their names, date and school, they were asked to answer math questions with this device. Children were given the chance to ask questions in order to get more information when it was needed. Children were asked to think-aloud after solving the problems/responding to the items. Children's verbal reports of thinking aloud were recorded by a Sony Audio Recorder (Figure 3.9), and the problem-solving process was recorded with a Gopro video camera throughout the entire session.

Figure 3. 5. Empatica E4 Wristband used for Affective Processes

Figure 3. 6. Eye-tracker used for cognitive processes

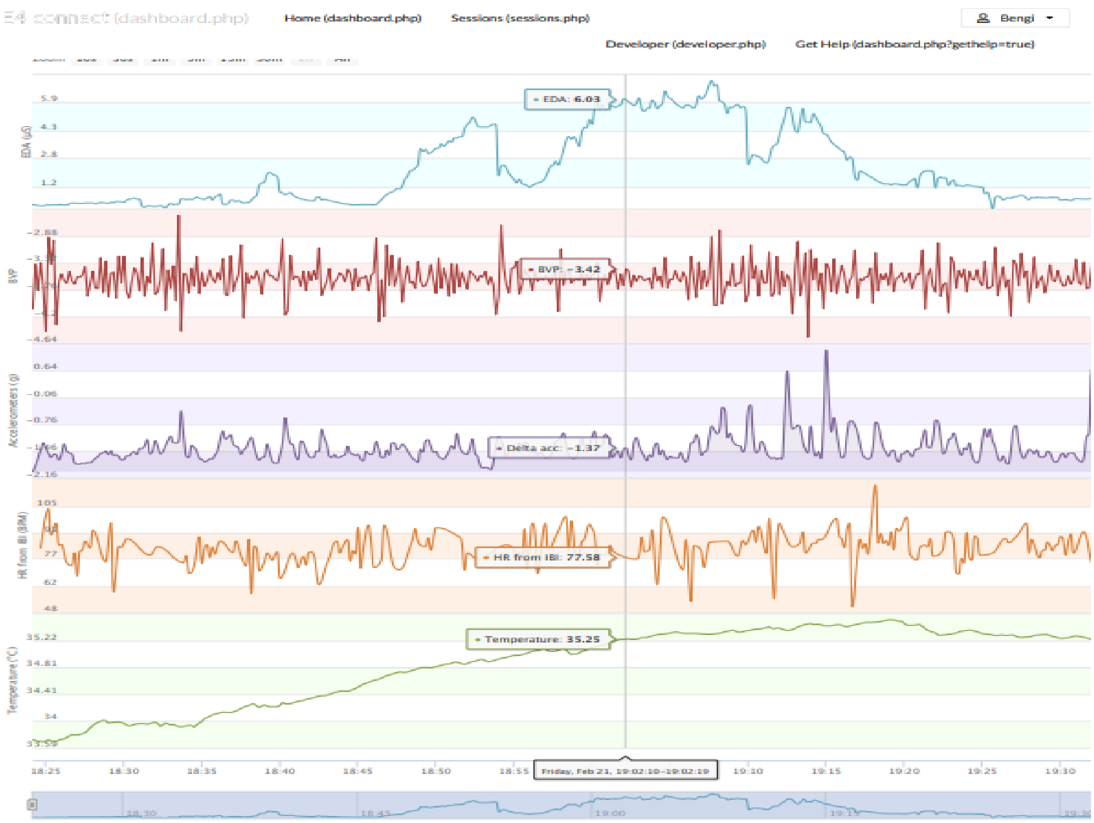


Figure 3. 7. EDA and BVP data stored in the E4 connect website

Figure 3. 8. Wacom Bamboo Slate used for handwritten problem-solving steps

Figure 3. 9. Voice recorder used for think-aloud processes

Sensors on the E4 are designed to collect high-quality data. It's the only wearable on the market that combines EDA and PPG sensors to monitor sympathetic nervous system activity and heart rate at the same time (see Figure 3.7). BVP, inter-beat interval (IBI), heart rate (HR), electrodermal activity (EDA) or GSR, skin temperature (ST) can all be measured with the E4 wristband. The Empatica device is designed to provide physiological parameters such as skin conductivity or heart rate that can be used to detect arousal changes. It does not, however, provide indicators for interpreting physiological data in terms of emotion, such as fear or joy (Sasshe & Leuchter, 2012). For this reason, in order to better detect the emotions of the students, after each math question, the students were asked questions to understand their feelings, such as how sure they were about this question, how much they tried, whether they felt worry or not. Worry and effort variables were created to better understand the affective process. It was then checked whether their discourses and physiological data matched and validated.

I combined the Think Aloud (RTA) method and eye-tracking technology. The practice of monitoring users' eye movements while staring at the location of an object is known as eye tracking (Figure 3.10 and Figure 3.11). After the session has concluded, an analysis tool generates a heat map that shows which sections of the website are being glanced at the most. It can be based on one participant's eye movements alone, or it can incorporate information from all of them (Hyrskykari et al., 2008). In our research, while solving the math question, the eye movements of the children were measured simultaneously. Then, under the sub-title of metacognitive skills, the area of interest and the number of times they looked back at the question were examined as variables by looking at the heat map. There may be a difference between where the child is actually looking and where the eye tracker reports that the child is looking, due to incorrect calibration. In cases where the caliber was incorrect or did not track the eye, I did not include it in the research data. During early development, abnormal fixation length patterns may act as an indicator such as Wass et al. (2015) recently showed that newborns in the early phases of autism had lower fixation lengths while viewing static images. To maintain internal validity, there was no child with a developmental disorder in our study. Lack of interest can cause the participant's data to be useless, loss of interest can be better understood and excluded from the research by looking at a student's E4 data, eye data, and think-aloud.

Figure 3. 10. Sample Item 2 from Group 1 Item Pool

Figure 3. 11. Monitoring students’ eye movements while staring at the location of an object

I also see technologies using a multimodal mixed-methods approach in other studies with children. Le-Cultura et al. (2021) use a combination of traditional video annotations and MMD to better understand how children interact with educational technology. They used a camera, wristband, eye-trackers, and Kinect to collect data while 26 children aged 10 to 12 played a Motion-Based Educational Game (MBEG). Children's skeletal data were collected using the kinect sensor, skeletal data is unnecessary for us as we do not make children play a game based on any physical activity. I preferred the screen-based eye tracking device because the participants were looking at the computer screen throughout the session. In this study, Tobii, which is eye tracking glasses with a built-in camera on the bridge of the nose, was preferred for the participants to move more freely, since it is a game-based study.

The study of Tóthová et al. (2021) looked at how well upper-secondary children could use the periodic table to solve problems. To analyze the logic behind the students' performance, eye-tracking and retrospective think-aloud approaches were applied to map the tactics they utilized and challenges they encountered when performing the tasks. The eye-tracker data was subjected to a quantitative analysis, which included a temporal fixation length evaluation on designated areas of interest. A qualitative study of the students' procedure was also thanks to the think-aloud approach, which supported the eye-tracking record as well as the students' transitions. In addition to the similarity of think aloud and eye-tracking processes, in this study, physiological parameters were not examined with the Empatica E4 device as in our study. Salmeran et al. (2017) looked at how high school students expressed skimming and deeper processing of information when answering questions using a Wikipedia page, as well as how their reading comprehension abilities and question type influenced these processes. Retrospective think-aloud methods and eye-tracking measurements were examined. In the coding of eye movement data, the area of interest was coded similar to what I did in my study. Then they computed the number of dwells and run dwell time based on eye-tracking data.

When I look at these studies, I observe that there are no physiological parameters other than an eye tracking device in studies with children using the think aloud procedure.

3.3.4. Data collection instruments of the research phases

A researcher-developed questionnaire, which is called as teachers' teaching method and measurement-evaluation strategy preferences questionnaire [TMMESP-Q] and its dimensions is primary data collection tool for quantitative survey after a semi-structured interview form was also used to collect initial pilot data for the questionnaire. The TMMESP-Q was developed to analyze teachers' teachers' teaching method and measurement-evaluation strategy preferences after curriculum change. Then, think-aloud process protocol and mathematics items pools consisting of open-ended and multiple-choice item formats were prepared for 5th grade students. Following sections provide a brief overview of data collection tools for the survey and the neuroeducation process.

a. Data collection instruments of quantitative survey: TMMESP Questionnaire

For the quantitative survey, TMMESP questionnaire was developed and used to determine middle school mathematics teachers' teaching method and measurement-evaluation strategy preferences after curriculum change.

Instrument Development Phase. A semi-structured interview protocol form (see. Appendix K) developed by the researcher containing eight questions was used to identify mathematics teachers' curriculum change practices (their view of curriculum, the similarities they observed, their instructional preferences for mathematics teaching and their preference for preparing students assessment and examination, constructivist approach) in Sarıyer middle school after the policy change in 2017, similarities and differences between the enacted curriculum (applied in 2018-2019 semester) and old curriculum (valid until 2017-2018), their experiences regarding in-class teaching method and measurement-evaluation, their observation regarding mathematics classes during doing mathematics exercises, their students' cognitive, metacognitive and affective experiences on open-ended items. In addition, they were asked to share their first-hand experiences revealed from in-class teaching and the reviewed literature about the difficulties of their experiences, their students' expectations, and improvement area. The form was also subjected to expert opinion from two mathematics teachers, one associate professor in the field of mathematics education and my advisor to be valid and reliable.

Firstly, I separately transcribed verbatim into a Word document from audio-recorded, semi-structured interviews conducted with teachers and the field notes are typed. I adopted an inductive approach for content analysis to discover patterns, categories and themes (Patton, 2002). Content analysis was separately conducted to these transcriptions. Widely used in social sciences, this method was used to scan transcriptions and to determine the patterns behind words and concepts (Krippendorff, 2004, 2011). I firstly read and reread the raw data to immerse into interviews and field notes. Subsequently, she started coding and generating categories and themes. As a result of content analysis of the interviews, 112 codes and 16 categories were reached.

Based on the codes, categories and themes from semi-structured interviews systematic literature review, I authentically and theoretically developed and decided the Teaching Methods [TM] and Measurement-Evaluation Strategy [MES] Preference Questionnaire (TMMESP-Q) dimensions.

b. Data collection instruments of neuroeducation process

There are many data collection instruments used in the neuroeducation process at the brain dynamics laboratory while experimenting with 5th grade students. They are think-aloud process protocol, Empatica E4 wristband, Gazepoint eye-tracking tool, Sony voice recorder tool and Wacom Bamboo Slate (see the detail of the tools pp. 132-134 above). Empatica E4 wristband was connected to and calibrated together with other data gathering tools like an eye tracker, wristband, and camera for recording facial gestures.

Fifth graders were asked to think-aloud after solving the problems/responding to the items. Children's verbal reports of thinking aloud were recorded by a Sony Audio Recorder, and the problem-solving process was recorded with a GoPro video camera throughout the entire session. While Empatica E4 Wristband was used for measuring affective processes, eye-tracker was used for cognitive processes, voice recorder was used for think-aloud processes, and Wacom bamboo slate was used for handwritten problem-solving steps.

Think-aloud process protocol (see. Appendix R) was designed for 5th grade students in which students ID, name code, study date and time, and report analysis data were written. Then, I qualitatively coded 1) whether students answer was true or false, 2) item difficulty level (1 = easy, 2 = moderate, 3= difficulty), 3) the description of think-aloud voice record. For instance, I described how students' *cognitive strategy* subskills correspond to eye-tracking system data; *self-checking* subskills correspond to eye-tracking system data; how many times did a student look back and forth to the items? How many times did s/he focus on the item? Do they have the tendency of looking back and forth while responding? 4) analysis of solution steps through hand-written data: by looking at the solution steps, did student use any the cognitive strategy? Did s/he use her/his self-checking skill? 5) Qualitative description of similarities and differences between actions and discourses. 6) Regarding emotions: what does the E4 report show? (add audio, visual field note), Any sense of effort or feeling of worry? The researcher put EDA and BVP screenshot here among E4 data. 7) Finally, Area of Interest and time on task added into the protocol so that I can understand the time student started to respond the item and s/he turned to me and speak aloud.

All details of each student were collected from data collection tools and recorded into the protocol sheet. Individual deep data as recorded for 32 students.

3.3.5. Data collection procedures of the research phases

There are four phases explained for the data collection procedures of *multimodal mixed methods concurrent dominant status* design.

a. Data collection procedures of document analysis: Examination of authentic teacher-made items

Middle school mathematics teachers in Turkey are responsible to collect two core examinations from their students in one semester. Therefore, each middle school mathematics teacher (n = 10) who volunteered to participate in the study submitted to me two sets of teacher-made examinations, a total of 21 authentic teacher-made examination items were collected. The total of the mathematics item pools consisted of 380 items. The authentic teacher-made examination items were given by the teachers as a hardcopy to me and I made an online copy before conducting the analysis.

The authentic teacher-made exams were subjected to document analysis (Patton, 2002) to identify teacher-made examination items into 5 major themes. So, document analysis on the examination items was yielded into 5 major themes: 1) Item type, 2) Learning unit of the Turkish Middle School Mathematics Curriculum (MoNE, 2018), 3) Learning outcomes from the curriculum (MoNE, 2018), 4) The Revised Bloom’s Taxonomy (Bloom et al., 1956; Anderson & Krathwohl, 2001); 4.1.) Level of Knowledge and 4.2) Cognitive Process Dimension, 5) the Trends in International Mathematics and Science Study [TIMSS] Cognitive Domain and subdomains.

Each item (N = 380) was listed into a codebook in Word.docx document and became ready for the document analysis process (see Table 3.6 below for codebook). The mathematics learning outcomes in the MoNE mathematics curriculum were represented as M.5.1.1.2, for instance, in which M [mathematics curriculum]. 5 [5th grade]. 1 [1st unit]. 1[subunit]. 2[rank of the learning outcome].

Table 3. 6. Codebook and Analysis Process for Each Examination Item

Item Number	Item Type	Learning Unit	Mathematics Learning Outcomes in the Curriculum	Revised Bloom's Taxonomy Knowledge Dimension	Revised Bloom's Taxonomy Cognitive Process Dimension	TIMSS Framework Cognitive Dimension
36) Write down the place value of the underlined numbers below. A) 7 9̲58 704: B) 30̲ 582 315:	OE	M.5.1 Numbers And Operations	M.5.1.1.2 Students will be able to specify the divisions and digits of up to nine-digit natural numbers and the value of the digits.	Factual	Remember	Knowing Recall
37) Write the first 4 steps of the pattern of numbers starting from 6 and adding 7.	OE	M.5.1 Numbers And Operations	M.5.1.1.3 Students will be able to form the required steps of the number and figure patterns given the rule.	Conceptual	Understand	Knowing Classify
38) Çiğdem and Gözde have a total of 600 lira. Since Çiğdem's money is twice of Gözde's money, how many lira does Gözde have?	OE	M.5.1 Numbers And Operations	M.5.1.2.12. Students will be able to solve problems involving four operations.	Procedural	Apply	Knowing Compute
39) 43 781 + 51 217= ?	OE	M.5.1 Numbers And Operations	M.5.1.2.1 Students will be able to add up to five-digit natural numbers.	Procedural	Apply	Knowing Compute

The items before and after policy change were examined, the findings showed the teachers in School 1 shared 31 items related with pre-change and 86 items related with after change. The teachers in School 2 shared any item related with pre-change whereas 87 items related with after change. The teachers in School 3 shared any item related with pre-change, however, they shared 45 items related with after change. The teachers in School 4 shared 38 items related with pre-change and 79 items related with after change. The teacher in School 5 shared only 14 items related to after the policy change. All in all, 69 items shared by the teachers were related with prior policy change whereas 311 items were related after the policy change (see Figure 3.12).

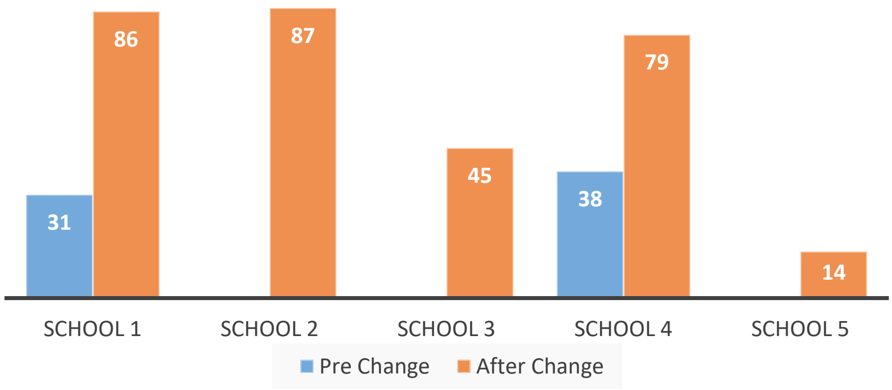


Figure 3. 12. Bar Graph of Items by Pre and After Policy Change

b. Data collection procedures of quantitative survey: Investigation of teachers' teaching method and measurement-evaluation strategy preferences

Upon granting the permissions by the Human Subjects Ethics Committee at the university (see Appendix E) and subsequently by Provincial Directorate of National Education in İstanbul (see Appendix H), the process for piloting began with the first meeting with the school managers. The volunteer participants for the pilot were approached and an interview schedule was arranged with them.

By way of explanation, authentic teacher-made mathematics examination items have been analyzed and interpreted. After this stage, I contacted the mathematics teachers from selected participant schools in Sarıyer district. Within the framework of an Interview Protocol, face to face meetings were started. The semi-structured interview protocol was prepared by me and checked by my advisor in terms of the aim of the current research, scope of research questions, and validity. The interviews have been conducted with the teachers outside the school hours, when there were no obstacles, and high motivation took place in the interview environment. 14 teachers were interviewed in total. In the process, 3 of them (from a public school) explained that they did not have enough time for individual interviews. So, they were asked to conduct focus group interviews. Unfortunately, when I arrived at the school, they said that they would like to give their answers **in a written form**, not verbally. The interviews of the teachers who gave permission to be recorded were transcribed. A codebook was prepared on the responses of the teachers who completed the Interview Protocol (see Appendix L). Because a **codebook** summarizes key information about the variables in a research project (Creswell, 2012). Encodings of the transcripts were completed through a codebook.

Writing the items for TMMESP-Questionnaire. The TMMESP-Q has totally consisted of 35 items, which were rated on a 5-point Likert Type scale from "*strongly agree*" to "*strongly disagree*" for first part in TM and "*always*" to "never" for second part in MES. The sample items for TM were "I have noticed the decrease in the number of learning outcomes" [item 3], "I prefer using group teaching methods (e.g. cooperative learning, think-pair-share etc.)" [item 10], and "I prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.)" [item 15]. The sample items for MES were "I prepare examinations that include a mixture of multiple-choice and short-answer items" [item 5], "I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought)" [item 13], "I determine the number of in-class examinations to be administered together with the students" [item 15]. The questionnaire contains 20 items in the TM section and 15 items in the MES section of the TMMESP-Q and ratings on each item were requested on a 5-point Likert type scale (1= strongly disagree, 5=strongly agree in TM; 1= never, 5=always in MES). The highest score obtained from the scale would be 175 while the lowest score would be 35. The items and the variable names are presented in Table 3.7 below.

Table 3. 7. Items of TMMESP-Q

Variable name	Description of the variable	Adaptations from the literature
	Teaching Method Preferences:	
1	I think there is no change in the purpose of the curriculum.	(Ornstein & Hunkins, 2004)
2	I think there is no change in the philosophy of the curriculum.	(Ornstein & Hunkins, 2004; Ozmon & Craver, 2008)
3	I have noticed the decrease in the number of learning outcomes.	(Ornstein & Hunkins, 2004)
4	I think the content of the subject has been enriched.	(Oliva, 2009; Ornstein & Hunkins, 2004)
5	Using concrete materials (e.g., mathematical objects) during classroom teaching helps me a lot.	(Baroody, 2017)
6	I prefer doing activities that provide opportunities for student creativity.	(Özmantar, Bingölbali, Demir, Sağlam & Keser, 2009)
7	I think there is no change in the content of the Mathematics resource books of the Ministry of National Education.	(Yüksel, 2000)
8	I prefer using the teacher's handbook.	(Demirel, 1992; 2012)
9	I change my in-class teaching method to make my students active.	(Phillipson, Riel & Leger, 2018)
10	I prefer using group teaching methods (e.g., cooperative learning, think-pair-share etc.).	(Mayer & Alexander, 2011)
11	I design lessons that enable my students to learn by exploring mathematics effectively.	(Bruner, 1961, 1996; Abrahamson & Kapur, 2018)
12	I only use direct instruction.	(Moore,1986; Westerhof, 1992)
13	Before the lesson, I check the students' readiness.	(Özer & Anıl, 2011)
14	I try to use educational technologies when teaching in-class.	(Bos, 2009)

Table 3.7. (continued)

Variable name	Description of the variable	Adaptations from the literature
15	I prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.).	(Von Glasersfeld, 1995).
16	I feel the need to use different questioning techniques (e.g. Why? How? etc.)	(Michaels, Connor, Hall, & Resnick, 2010)
17	I give examples from daily life while teaching a topic.	(Kitchen, 2016; NCTM, 2014)
18	I prefer designing a learning environment that makes students think about the topic they work on.	(Von Glasersfeld, 1995)
19	I use instructional techniques that require students to take responsibility for their learning (e.g. demonstration, question-answer, brainstorming, discussion).	(Mayer & Alexander, 2011)
20	I encourage students to do research.	(Bruner, 1961; Clabaugh, 2010)
	Measurement and Evaluation Strategy Preferences:	
1	I make changes in the measurement and evaluation process compared to the previous implementations.	(Alkharusi, Kazem & Al-Musawai, 2011)
2	I administer examinations based on downloaded online sources (e.g. forums, websites etc.).	
3	I use measurement tools that include multiple-choice items	(Kanatlı, 2008)
4	I use formative assessment to measure course learning outcomes.	(DeLuca, Valiquette, Coombs, LaPointe-McEwan & Luhanga, 2006; Marzano, 2006)
5	I prepare examinations that include a mixture of multiple-choice and short-answer items.	(Çakan, 2004; Kilmen & Çıkrıkçı-Demirtaşlı, 2009)
6	I ask problem-solving items related to real life problems.	(Shepard et al., 2005)
7	I use portfolio that will enable the students to show their performances at the end of the term.	(Barootchi & Keshavarz, 2002; Özbaşı, 2008)
8	I apply quizzes.	(Çakan, 2004)
9	I use open-ended items in my in-class examinations.	(Birgili, 2014)
10	I use the question-answer technique in my teaching.	(Özbaşı, 2008).
11	I give students choice to choose which item types they want to be included in their examinations.	(Gelbal & Kelecioğlu, 2001; Zhang & Burry-Stock, 2003).

Table 3.7. (continued)

Variable name	Description of the variable	Adaptations from the literature
12	I prefer item types that require students to use procedural skills in the examinations.	(Reiser, 2004)
13	I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought).	(Kilmen & Çıkrıkçı-Demirtaşlı, 2009)
14	I use open-ended and multiple-choice items together in my in-class examinations.	(Özbaşı, 2008)
15	I determine the number of in-class examinations to be administered together with the students.	(Acar-Erdol & Yıldızlı, 2018)

The questionnaire was based on two parts which measure mathematics teacher's preferences on TM and MES respectively. The writing the items for TMMESP-Q theoretically grounded in the literature and the inferences from the pilot interviews according to cultural specific context. It was developed based on expert opinion and then the literature was used for dimensioning. The translation of Turkish version of the TMMESP-Q items were depicted in Appendix M whilst its translation process was tabulated in Appendix N.

c. Data collection procedures of multimodal phase: Neuroeducation

Data collection procedures of neuroeducation process related to Phase 3: Reflection of students' metacognition and affective processes to different types of questions and Phase 4: Evaluation of students' reactions and responses to different types of questions with the use of eye-tracker and biometric sensors are explained in the following statements are discussed in this part.

In order to investigate research questions "How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types? and "What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?" I collaborated with an assistant professor (her co-advisor) at the Brain Dynamics Laboratory in the university during the long Covid-19 pandemic period. After the pilot preparation phase, 10 mathematics items were prepared and selected for fifth grade students similar to TIMSS, PISA, MoNE national examinations, and mathematics teachers' in-class authentic examinations. This question pool was divided into 2 groups as multiple-choice and open-ended items. After expert opinion of the departments of ME, CI and MFE, the Think Aloud Process protocol was designed. Nevertheless, due to Covid-19 conditions, permission was obtained with delay by MoNE. I had to wait around 6 months for the permission. The laboratory environment was specially prepared for the students' well-being. 32 5^{th} grade students were invited with their families and sometimes transported with a special school bus. They voluntarily participated in the experimental process, evaluated for their performance in which metacognitive and affective process measurements were conducted in the laboratory that lasted for around 40 minutes. These studies lasted 2 months (from January 27, 2021 to March 26, 2021). While the students were responding to the items, a thinking-aloud process was carried out. In-depth analysis reports were written for about 4 months, as there were different biometric tools, interviews, and detailed data from my field notes. A co-coder was also used to enable the intercoder reliability of the analysis. Each data analysis process, exchange of ideas and expert opinions were held with the advisor and the co-advisor. The data were analyzed holistically and the results were written.

Interrater Agreement of Phase 3 and 4. I studied with a co-coder in the analysis process of fifth grade students' metacognitive skills and affective responses revealed from several tools such as a thinking-aloud process and eye-tracker and biometric sensors. The co-coder is an expert (senior year) from the department of psychology.

The co-coder and I coded each transcription from the Think Aloud process protocol individually, one by one, into the codebook excel.xls google drive worksheet. The meaning of each code was prepared into a codebook for Neuroeducation Process (see. Appendix S). This was called a data aggregation process.

As O'Connor and Joffe (2020) asserted, intercoder reliability of a coding framework in qualitative analysis should be evaluated to yield for trustworthiness of the analysis process. It is highly recommended as "a good practice" for education and educational research even though some professors claim that it is an unnecessary step which may act against the nature of the goals of qualitative analysis. In this study, I as a researcher needed multiple researchers to interpret highly deep data collected from fifth grade students. Therefore, in this step a co-coder studied with me in a collaborative manner. I approached the data, which is a very unique case and strategy in Turkey, to interpret it in a similar way. I read the data grounded in the Think-Aloud Protocol for Students one by one and code it into each verbatim. Then, I reread them again in the iteration of the data analysis process and map each code into a Data Aggregation Process (an excel google document) and colored each code to enrich its understandability. The research evidence highlights that using an expert for inter-coder promotes the transparency of the coding process. I used several iterations for the inter-coder steps. I hope that it improves the systematicity, communicability of the process. At least two days a week, we (co-coder and I) came across and met in a meeting to study data, promote reflexivity and dialogues between us, and convince each other on disagreements. The dialogues between co-coders were:

The detail of agreement and disagreement sessions showed me that there were 10 items and each were responded to and answered by 32 students. Totally 1178 codes were revealed (see. Table 3.8 for disagreements between inter-coders).

Table 3. 8. Disagreements related to Items

Items	Sts ID and # of disagreements	Total # of disagreements
Item1	ID8 = 1	10
	ID12 = 2	
	ID16 = 2	
	ID18 = 5	
Item 2	ID10 = 5	15
	ID11 = 5	
	ID31 = 5	
Item 3	ID12 = 4	7
	ID28 = 3	
Item 4	ID13 = 8	21
	ID16 = 2	
	ID18 = 4	
	ID24 = 7	
Item 5	ID6 = 1	13
	ID11 = 3	
	ID14 = 6	
	ID21 = 3	
Item 6	ID9 = 1	10
	ID15 = 5	
	ID19 = 4	
Item 7	ID12 = 6	12
	ID17 = 1	
	ID20 = 2	
	ID28 = 3	
Item 8	ID15 = 2	9
	ID16 = 2	
	ID29 = 5	
Item 9	ID23 = 1	4
	ID32 = 3	
Item 10	ID 30 = 4	4

Note. # = number, Sts = students

In line with the abovementioned disagreement codes, I coded 1178 codes into the data aggregation process and 10% of them were coded by a co-coder. Finally, we disagreed on 105 codes. While in the pilot study the intercoder reliability was

90% from pilot study and research and tool calibration process with young adults; it was 85% with fifth grade students from the first iteration and 91.09% from the final analysis.

Item Difficulty Analysis on Students Responses

Item analysis was also conducted to analyze the fifth-grade students' responses to mathematics examination items with the intention of evaluating the examination quality applied in the Brain Dynamics Laboratory. It is a significant process to maintain testing effectiveness and fairness. In this part, the item difficulty was computed to the examination items to interpret on whether they were too easy or too hard. Its index ranges from 0 (the lowest value) to 1.00 (the highest value). Easy objects are indicated by higher difficulty indices. A question with an item difficulty level of.75 has been properly answered by 75% of the test takers. An item with a difficulty level of.35 was correctly answered by 35% of test takers. The formula below was used to determine the item difficulty. (Crocker & Algina, 1986).

Difficulty = (# who answered an item correctly / Total # tested) X 100

When the items were analyzed in terms of difficulty via jMetrik^{TM} psycho-measurement system (Table 3.9.), the results revealed that Item 1 (Procedural, Applying) had an item difficulty level of .84; the item answered correctly by 84% of the students. Item 2 (Conceptual, Applying) had an item difficulty level of .88; the item answered correctly by 88% of the students. Item 3 (Procedural, Evaluating) had an item difficulty level of .22; the item answered correctly by 22% of the students. Item 4 (Conceptual, Analyzing) had an item difficulty level of .53; the item answered correctly by 53% of the students. Item 5 (Conceptual, Applying) had an item difficulty level of .31; the item answered correctly by 31% of the students. Item 6 (Procedural, Applying) had an item difficulty level of .38; the item answered correctly by 38% of the students. Item 7 (Conceptual, Analyzing) had an item difficulty level of .25; the item answered correctly by 25% of the students. Item 8 (Factual, Applying) had an item difficulty level of

.59; the item answered correctly by 59% of the students. Item 9 (Procedural, Applying) had an item difficulty level of .78; the item answered correctly by 78% of the students. Item 10 (Procedural, Applying) had an item difficulty level of .75; the item answered correctly by 75% of the students. According to this item difficulty analysis, a criterion of less than .30, between .40 and .60, more than .80 (McCowan & McCowan, 1999), the item 1 as having .84, item 2 as having .88 item difficulty was found to be easy. Item 9 as having .78 and item 10 as having .75 item difficulty was found to be relatively easy. Moreover, item 4 having .53 and item 8 as having .59 item difficulty was found to be normal, not too difficult. On the other hand, item 5 as having .31 and item 6 as having .39 item difficulty was found to be nearly difficulty whereas item 3 as having .23 and item 7 as having .25 was found to be very difficult items.

Furthermore, the correlation between items were additionally examined to see the association between them. Bivariate point-biserial correlation was conducted and the findings indicated that item 1 was correlated with item 2 ($r = .36$) which was a positive medium association, with item 10 ($r = .75$) which was a positive strong association. Item 2 was correlated with item 8 ($r = .40$) which was a positive medium association. Item 3 was correlated with item 10 ($r = -.39$) which was a negative medium association. Item 6 was correlated with item 10 ($r = .45$) which was a positive moderate association. Item 7 was correlated with item 8 ($r = .54$) which was a positive large association.

The items can be interpreted in terms of item discrimination that item 1 is a good item, item 2 is a good item, **item 3 is a poor** item, item 4 is a good item, item 5 is a good item, item 6 is a fair item, item 7 is a good item, item 8 is a good item, item 9 is a fair item, item 10 is a good item as in line with the indexes of “good” if the index is above .30; “fair” if it is between .10 and.30; and “poor” if it is below .10 (Nunnally, 1967).

Table 3. 9. Statistics for Item Difficulty and Item Discrimination Index

Items	Item Difficulty Easy, Medium, Hard	Item Discrimination Good, Fair, Poor
Item 1	0.8438	0.4874
Item 2	0.8750	0.3435
Item 3	0.2188	-0.0615
Item 4	0.5313	0.4299
Item 5	0.3125	0.3229
Item 6	0.3750	0.2902
Item 7	0.2500	0.4167
Item 8	0.5312	0.4668
Item 9	0.7813	0.1021
Item 10	0.7500	0.4782

The item difficulty was measured as Easy, Medium and Hard; and item discrimination was measured as Good, Fair and Poor. For instance, item 3 was found to be difficult when item 2 was easy. On the other hand, while item 2 was a good item, item 3 was poor to have an ability of differentiating among students on the basis of how well they know the items being tested.

The test statistics for ten examination items prepared for fifth grade students reliability results were given in Table 3.10.

Table 3. 10. Test Statistics for Ten Items prepared for Fifth Grade Students

Test Level Statistics	*M*	*SD*	Median	Skewness	Kurtosis	KR21
Items	5.47	2.16	6.00	-0.22	-0.36	0.52

The test level statistics showed that the mean of students over ten items was 5.47 with standard deviation of 2.16. The reliability of the test measured with binary variables by Kuder–Richardson 21 (KR-21) obtained 0.52. Having rely on approximate inter-item consistency, 0.52 indicated moderate relationship between the items.

3.3.6. Data analysis

This part included the data analysis process according to each research questions.

a. Data analysis of document analysis: Examination of authentic teacher-made items

The data collected at the initial stage of the study were analyzed qualitatively. The 380 authentic teacher-made mathematics examination items were examined to scrutinize the extent of the recently enacted mathematics curriculum [for 2017-2018 education semesters] compatible with the proposed assessment procedures within the curriculum to prepare middle-school students for the learning outcomes. Hence, the items were subject to document analysis.

Document analysis on the 380 mathematics examination items were categorized by following the stages; 1) item types (fill in the blanks, true/false, multiple choice, open-ended (constructed OE)), 2) learning unit in the intended mathematics curriculum (i.e., M5.1. Numbers and Operations, M 5.2. Geometry and Measurement), 3) learning outcomes in the intended mathematics curriculum (i.e., M.5.1.2.11. Students will be able to find the result of parentheses that contain up to two types of operations.), 4) The Revised Bloom's Taxonomy was used to classify the items in terms of knowledge dimension (*Factual Knowledge*, *Conceptual Knowledge*, *Procedural Knowledge*, and *Metacognitive Knowledge*) and cognitive process dimension (*Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*), 5) the Trends in International Mathematics and Science Study [TIMSS] Assessment Framework Cognitive Domain (*Knowing*, *Applying*, *Reasoning*) and 6) TIMSS Cognitive Domain subcategories (i.e., *Knowing: Recall, Compute; Applying: Determine, Represent/Model; Reasoning: Analyze, Draw Conclusions, Justify*). After coding each item one by one in a codebook (see Table 3.8 above), each code was moved to a spreadsheet to be subject to descriptive analysis. Further, the frequencies and percentages were calculated using IBM Statistical Package for Social Sciences 26.0 for Mac (SPSS, 2012) to describe the main features of the data.

Following the theoretical frameworks of the Revised Bloom's Taxonomy and TIMSS Framework taxonomies, the experts independently coded each item in a blank codebook. After that, the expert and I set a meeting face to face and performed a reliability check of the classification. Interrater agreement process was based on Miles and Huberman's (1994) formula: Reliability = $\frac{Consensus}{Consensus+ Disagreement}$. The first expert and I discussed each item and their coding. When they disagreed, they discussed the reason and returned to the framework to reread the categories and examples. There were 17 categorized cells in which there was a disagreement. The disagreement related with the Revised Bloom's Taxonomy (n = 12) was due to the level of knowledge (n = 3) and cognitive process dimension (n = 9). For instance, in a geometry question, the item was asked to select true features from an obtuse-angled triangle while the expert was coding it as "factual", I coded it as "conceptual." Besides, on the question which asked the 10th step of a given pattern (M.5.1.1.3. Students will be able to find/create the desired steps of the given rule of number and shape patterns), the expert coded this learning outcome as "analyzing" while I coded as "applying." Moreover, the disagreement related to the coding of items in terms of TIMSS framework (n = 5) was due to the unrelated function between the operational description of *Applying* from Revised Bloom's Taxonomy and *Applying* category from TIMSS Framework because from the definition, *Applying* category from TIMSS Framework also included compute or computational application skills such that *Knowing* from TIMSS Framework was interrelated with of *Applying* from Revised Bloom's Taxonomy. The TIMSS Framework categories were found to measure higher-level skills in nature. Hence, Miles and Huberman's (1994) reliability was calculated as .80 between the first expert and I at the end.

The second expert and I also discussed each item and their coding. When we disagreed, we discussed the reason and returned to the framework to reread the categories and examples. There were 14 categorized cells in which there was a disagreement. The disagreement related with the Revised Bloom's Taxonomy (n = 10) was due to the level of knowledge (n = 4) and cognitive process dimension (n = 6). For example, in the numbers and operations learning unit, an item was asked to determine the ninth digit number from given premises in the item stem. The second expert coded this item as "factual" but I coded it as "procedural" (M.5.1.1.2. Students will be able to indicate the divisions, digits and digit values of natural numbers up to nine digits). Also, the question asked to find an irrelevant decimal representation from given fractions and their decimal representations (M.5.1.6.3. Students will be able to compare multiples of fractions specified by decimal and percent representations) was coded as "applying" by the expert when I coded it as "understanding". Moreover, the disagreement related to the coding of items in terms of TIMSS framework (n = 4) was due to similarly the unrelated function between the operational description of *Applying* from Revised Bloom's Taxonomy and *Applying* category from TIMSS Framework. The disagreement between the second expert and I was in the "Knowing" and "Applying" category as stated by the first expert. For instance, *Knowing* category consists of recalling, recognizing, classifying/ordering, computing (i.e., carrying out algorithmic procedures for + - * : or a combination with whole numbers, fractions, decimals, and integers. carrying out basic algebraic procedures), retrieving (retrieving information from graphs, tables, text or other sources), measuring whereas *Applying* consists of determining, representing/modeling, implementing (i.e., implementing strategies and operations to solve problems relating familiar mathematical concepts and procedures). Hence, Miles and Huberman's (1994) reliability was calculated as .84 between the second expert and I at the end.

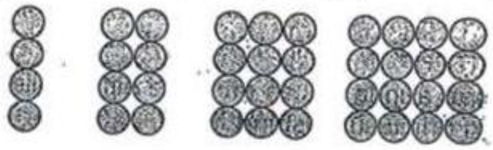
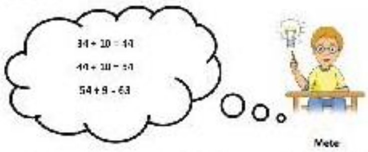
In addition, I showed the learning outcomes to the mathematics teachers who volunteered to participate in the study after aligning them with the items and received their approval or feedback as the first practitioners in this process.

Finally, I analyzed 380 examination items through 3 iterations. The first was after reading the theory of taxonomy and framework. The second was after reading additional resources and specimen items. Thirdly, after working with two experts (first an associate professor and the second a doctoral candidate) and reaching an agreement on the categories of items. I checked my own previous analysis for the last time and finalized the items after the third iteration. Table 3.11, Table 3.12 and Table 3.13 shows how sample items analyzed according to dimensions.

Table 3. 11. Sample Items from the Revised Bloom's Taxonomy Knowledge Dimension

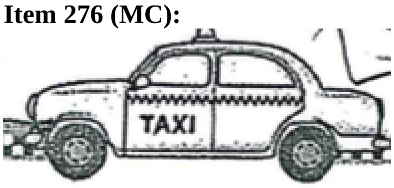
<table>
<tr><th>Types</th><th>Item # (item type)</th></tr>
<tr><td>Factual Knowledge</td><td>Item 19 (Fill in the blanks):
Write the natural numbers of which pronunciation is given below in the numerical form.
*ninety-eight million two hundred and thirty-six:
*three hundred twelve million one hundred twenty-two thousand seventeen:
*seven hundred million six hundred five thousand three hundred:
*nineteen million twenty-two:
*three hundred six million five hundred ninety thousand three hundred nine:</td></tr>
<tr><td>Conceptual Knowledge</td><td>Item 46 (OE):
Table: Number of seats and prices by firms<table><tr><th></th><th>Firm A</th><th>Firm B</th></tr><tr><td>Number of Seats</td><td>32</td><td>45</td></tr><tr><td>Ticket Price (Lira)</td><td>25</td><td>18</td></tr></table>According to the table above, if companies A and B organize a full voyage, which bus company will have more income?
Item 269 (MC):
The first four steps of a shape pattern are given below.</td></tr>
</table>

Table 3.11. (continued)

Types	Item # (item type)
Procedural Knowledge	 1st pattern 2nd pattern 3rd pattern 4th pattern Let's find the number of circles in step 5 of this pattern.
Metacognitive Knowledge	**Item 378 (OE):**  Determine what mathematical operation Mete used by examining the solution path. The mathematical operation used by Mete: Solve this operation with your own strategy:

The mathematics examination items collected from the mathematics teaches in this study were examined related to not only knowledge dimension of the Revised Bloom's Taxonomy but also its cognitive process dimension.

Table 3. 12. Sample Items From the Revised Bloom's Taxonomy Cognitive Process Dimension

Categories	Item # (item type)
Remember	**Item 57 (MC):** How do you write the expression "7 squared" in an exponential way? A) 7^3 B) 7^2 C) 2^3 D) 3^2
Understand	**Item 264 (MC):** Which of the following is the natural number that is resolved as an 3x100 000 + 9x1000 + 1x100 + 2x10? A) 390 201 B) 309012 C) 390 102 D) 309120
Apply	**Item 276 (MC):**  A taximeter landing fee is 4 TL, and 3 TL is added to the landing fee for each next kilometer. According to this, how much TL should Eymen, who travels 22 km, pay for this taxi? A) 26 B) 34 C) 55 D) 70
Analyze	**Item 373 (OE):** What values can the short side length take for a rectangular garden of which all side lengths are in cm and its area is 90 cm^2?
Evaluate	**Item 371 (OE):** Please look at the example in part a and do the other operations related to it. a) 257 x 9 = 2313 257 x 10 = 2570 [2570 - 2313 = 257] b) 85 x 99 = 85 x 100 = [ ] c) 42 x 10 = 42 x 11 = [ ] *Examine the multiplication and subtraction you have done above. How can you relate the result of the subtraction to the factors in multiplication?
Create	Not Applicable

Supplementary to the Revised Bloom Taxonomy, the authentic teacher-made examination items were further inspected related to TIMSS Framework as an international well-accepted categories.

Table 3. 13. Sample Items from the TIMSS Framework Categories

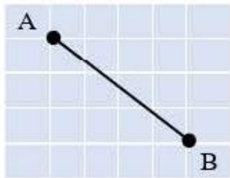
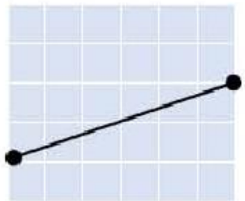
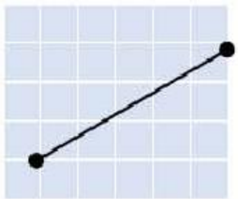
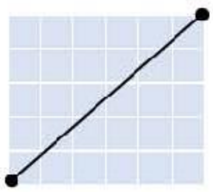
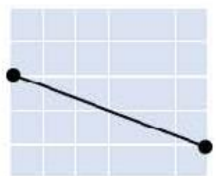
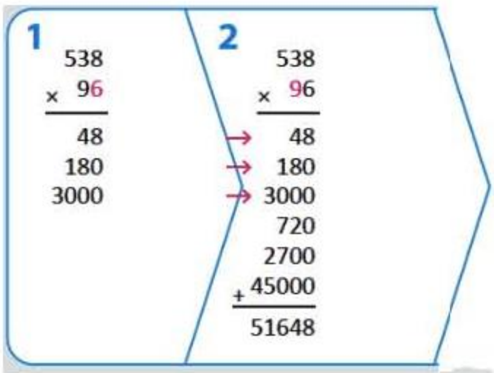
Categories	Item # (item type)
Knowing: Recognize	**Item 323 (MC):** In the grid below, [AB] is given.  So which of the following line segments is equivalent to [AB]? A)  B)  C)  D) 

Table 3.13. (continued)

Categories	Item # (item type)
Applying: Determine	**Item 302 (MC):** How many different natural numbers can be written instead of A to make the order of $\frac{1}{6} < \frac{A}{30} < \frac{23}{15}$ correct? A) 40 B) 41 C) 42 D) 43
Reasoning: Justify	**Item 368** (OE) Lydia solved the process of 538 x 96 with the method shown below. • Is Lydia's solution correct? • What method did Lydia use? Please explain.



b. Data analysis of quantitative survey: Investigation of teachers' teaching method and measurement-evaluation strategy preferences

The data obtained after the survey phase was first checked (see Appendix O), and the answers responded by each teacher to the questionnaire were first transferred to excel.xls and then to the SPSS statistical package program. I, initially, checked whether there is missing data. I analyzed data descriptively and I removed outliers by considering the opinions of the consultants. 6 responses which looked like outliers were discarded from 350 participant groups, and 344 data were stored for analysis. The data were statistically analyzed in terms of descriptive and inferential aspects following the research questions.

Having analyzed the theoretical structure of the Teaching Methods [TM] and Measurement-Evaluation Strategy [MES] Preference Questionnaire (TMMESP-Q) from the in-depth literature review, the items were also checked by my advisor and thesis committee members. Items under each dimension were tested and analyzed by using IBM Statistical Package for Social Sciences (SPSS) Version 26 for Mac.

c. Data analysis of multimodal phase: Neuroeducation

Pilot Study. Multimodal Phase: Neuroeducation. At the beginning of the Neuroeducation study, I stated above that I primarily studied and researched on young adults (i.e. university students). I observed how the multiple-choice and open-ended items to be asked in the pilot of this study work, whether they are understood correctly and where I should pay attention during experiment. I realized unclear questions. For example, in Item 7, they were asked to draw the whole of a figure whose fraction of some piece was given into the shaded area. Young adults were not ready either, as they did not encounter such questions in their middle school and high school years. They hardly ever responded. The visual figure has been redrawn in this item. The expressions in the question roots of some items were corrected grammatically. A pedagogue was also invited into the session. I discussed how and what to pay attention to while experimenting and observing the students. In addition, while I was working with young adults, the pedagogue observed me and gave feedback on how I had behaved, how I had asked questions, my mimics and my approach to the adult participants. While working with young adults, I had the opportunity to observe their reactions to the items. The most important point that distinguishes young adults from middle school students was that young adults were able to use less metacognitive subskills and could not reflect their inner speech as well as students. In particular, they were unable to use their re-expression subskills. They had little self-awareness about it. They were more comfortable with multiple-choice items since they were a group of young adults who had educated through approach which asked them to answer multiple-choice items. Usually, their inclination was to reach a solution in one way, but they did not go back and check their responses. Thinking processes were not as flexible as students. Besides, they were able to express their affective processes as they were.

Main Study. Multimodal Phase: Neuroeducation. In this part of the study, I explained 1) how I introduced into a neuroeducation process and then elaborated 2) introduction to the pilot study, 3) preparation for the main study and the process, 4) data collation process from the think aloud protocol and cognitive-affective measurement tools, 5) data aggregation process, and 6) finalization process. (see the summary of neuroeducation process in Figure 3.13.)

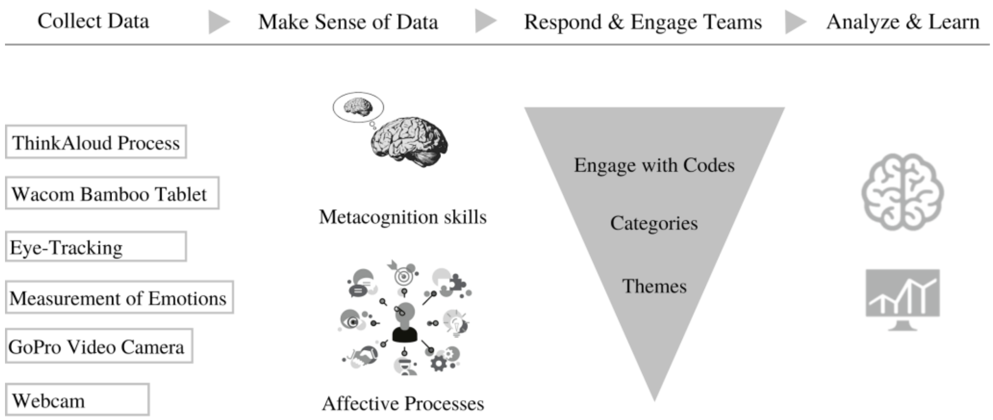


Figure 3. 13. Summary of a Neuroeducation Research Process

Step 1: Introduction to a Neuroeducation Research Process

Neuroeducation is an applied field. It is a developing field that combines practitioners, such educators, with researchers in the fields of neuroscience, educational psychology, and educational technology to study the relationships between learning, development, and dynamic brain processes. (Donoghue & Horvath, 2022). It addresses the link between brain function and pedagogy. It is an exciting opportunity to bridge the gap between researchers studying in the laboratory and educators in the real world. It is an interdisciplinary process that allows us to determine inferential results from human behaviors, their inner world, the brain, and even emotional reactions to the teaching and learning process. It covers both basic and applied research processes.

The aims of *Multimodal Phase: Neuroeducation* part of the current study are 1) to investigate how middle school students reflect their metacognitive skills and affective process when solving restricted open-ended items and 2) what students' reactions and responses are to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR). Moreover, the sessions were recorded with a 4K camera for investigating further behavioral analysis. With this aim, this part of the research study was designed as an interdisciplinary approach and required measuring and evaluating students' cognitive and metacognitive skills and affective processes. The process began with a conceptual and contextual study of neuroeducation, cognitive science, the Think Aloud process, and measurement tools such as eye-tracking and galvanic skin response tools, etc. Following a literature review of these topics, the next warm-up phase involved field observations together with a co-advisor who managed the process. The focus in particular was on the initial study and literature review on the use of cognitive and affective tools and programs, sometimes met with the co-advisor, in the Brain Dynamics Laboratory at the university, where the main experimental study was carried out. During weekly meetings, we designed the research process in conjunction with the experimental processes.

In order to become accustomed to this new area and not to feel like a fish out of water, I spent three months reading around the subject to familiarize myself with the environment, subject content, and goals. The biggest tool for me to adapt to the experimental environment was that my co-advisor had my personnel i.d. card entered into the system so I could enter and exit the room. This allowed me to get used to the laboratory, which I had not yet had the opportunity to use. I experimented with the tools and learned to use them without apprehension, even solving problems where they arose. As I got used to the environment, I began adjusting the layout for pilot studies. I learned from scientific articles on current research the importance of calibrating instruments. I became so used to being there, I even carried out departmental work from the laboratory at times. I believe that adapting to and familiarizing oneself with the environment is an important start to the research process, thus taking precautions to avoid threats to internal validity for the scientific research process (e.g., location, instrumentation: see Fraenkel et al., 2014 p.167 for more discussion).

Step 2: Introduction to the Pilot Study of Neuroeducation

Educational neuroscience as an applied field constituted the 3rd and 4th phases of this study. The pilot study was conducted with 14 young adults before starting the main part of the study. 14 young adults consisted of those from different departments of the university (e.g., the department of education, law, psychology) who volunteered to participate in. Four of them were male and ten of them were female. Their Grand Point Average (GPD) in the university was between 3.10 and 4.00 over 4.00. I learned a lot about the research environment, the multimodal tools to be used, the software embedded in the tools, calibration process before the application and what problems I might encounter in the pilot study.

I work as a research assistant in the department of mathematics and science education, MEF University, İstanbul, Turkey. As owner of the study, at every step, I exchanged ideas with my co-advisor, and we mutually agreed to proceed to perform initial experiments. The most critical stage was the pilot study and learned that the first pilot study should be carried out with young adults. For this purpose, while I was reading current literature, I worked on adjusting and calibrating the environment for young adults, how to solve any problems that might arise, and carried out weekly laboratory practices in the Brain Dynamics Laboratory. The first subject for the pilot was my co-advisor. As a researcher, I studied and performed how I should begin the research process, the initial preparation, how to introduce the research participants to the lab environment and the study itself, how long the process would take, how many items and on what subject would be asked, and how I would convey any further required instructions to the participants.

What held the research process up the most was the ethics committee approval that I had to obtain from the MoNE in İstanbul, Turkey. After four separate applications to the MoNE ethics committee, the Middle East Technical University ethics committee document was approved on March 16, 2020. I immediately sent the ethics committee documents to the MoNE. However, my documents were returned due to the excessive number of mathematics items that were intended to be asked to the fifth grade students in the experimental process, as well as an undefined inability to fully understand a part of the study, and also, due to the request to publicly share the students' and their parents' personal information, and some other items in the MoNE Ethics Committee application file. I adjusted the requested items and reposted the forms to the MoNE. However, when I received similar responses two more times during the Covid-19 pandemic, I wrote an official letter to the MoNE explaining my reasons with reference to the relevant policy articles and issue numbers, as the participants' private information requested by the ethics committee cannot be shared. Finally, the ethics committee approved my application (see Appendix J). Unfortunately, this excessive period of waiting delayed my thesis plans by nine months as I was unable to communicate with the school administrators and teachers without the ethics committee's permission. When my application was approved, I began preparations for the pilot study. Fifteen young adults from different departments voluntarily participated in the pilot study. The students differed in terms of gender and departmental distribution.

The most surprising observation from the pilot study was that the young adults were not cognitively as flexible as the middle school students while they were solving the mathematics items and their answers to some of the items (prepared for 5^{th} graders) incorrectly. They listened to the instructions I had given before the pilot study began and followed them up to the final item. I realized that the most important reason for carrying out initial experiments with adults is for the novice researcher to adapt and get used to the research process. Studying with both young adult and younger students helped me to make comparative analyses of answers to the items during the transition period. The adult participants generally tended to solve mathematics items using methods they had learned from their previous teachers. They tended to use metacognitive skills less than younger students in their Think Aloud processes. While they used some cognitive skills at a moderate level, they were more hesitant to grasp, make sense of, and use their self-control skills. The pilot study period lasted for one month between 11.01.2020 and 11.02.2020).

Step 3: Preparation for the Main Study of Neuroeducation and the Process

Following the pilot study, the Think Aloud process, asking questions, observing students' behaviors and taking field notes, answering their questions, calibrating and using the laboratory tools, and once the MoNE ethics committee had been obtained, four students voluntarily participated in the study. I studied and practiced 40-minute processes and problem-solving in case of unexpected problems. At this stage, I tried to gain access to 5th grade students and their parents in private and public middle schools in Istanbul using the snowball sampling method. First, I sent an e-mail to the schools containing the METU and MoNE ethics committee documents, a short text providing information, an advertisement comprising a research poster, and a copy of the consent forms. One-to-one contact was also made with the schools that voluntarily offered to support the study. Those who requested additional information did not hesitate to contact me and I answered their questions.

In order to investigate RQ3 and RQ4 (see page 36 for details), I collaborated with my co-advisor at the Brain Dynamics Laboratory during the Covid-19 pandemic. After the pilot preparation phase, 10 mathematics items similar to TIMSS, PISA, and MoNE national examinations and authentic mathematics teachers' in-class examinations were prepared and selected for fifth grade students. This item pool was divided into two groups: multiple-choice and open-ended items. After consulting expert opinion from the departments of Measurement and Evaluation (ME), Curriculum and Instruction (CI), and Mathematics and Science Education (MSE), the Think Aloud Process protocol was designed. However, due to Covid-19 conditions, permission from MoNE was again delayed and I had to wait almost another six months. The laboratory environment was specially prepared for the students' well-being. Thirty-two 5th grade students were invited with their families and sometimes transported by private school bus. They voluntarily participated in the experimental process, and their performance measured with metacognitive and affective processes conducted in the laboratory and lasting for around 40 minutes were evaluated. These studies lasted two months from January 27, 2021 to March 26, 2021. While the students responded to the items, a think-aloud process was carried out. Over four months, in-depth analysis reports were written incorporating different biometric tools, interviews, and detailed data from my field notes. A co-coder was also used to ensure intercoder reliability of the analysis. Each data analysis process, exchange of ideas, and expert opinions were recorded by the advisor and co-advisor. The data was analyzed holistically and results were obtained.

Step 4: Data Collection Process from the Think Aloud Protocol and Cognitive-Affective Measurement Tools

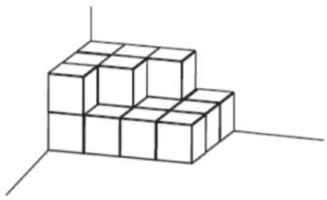
The audio recordings of the students who used the think-aloud process were transcribed verbatim. Individual files were opened for each student in Google Drive where data from the eye-tracking system, audio recordings, and GSR Empatica 4 emotional measurements were stored. It took one hour to save the data unit for each individual at this stage. In addition, each data unit was checked by the co-coder in order to avoid any mistake.

The first thing I did at this stage was to prepare a template with my co-advisor for analyzing the data. It was not easy to record the data from each data unit in the template. The steps used while analyzing the titles and data in the attached template and carrying out the mixed method analysis (i.e., qualitative and quantitative) are listed below (Table 3.14 and Table 3.15)

Table 3. 14. Thinkaloud Process Protocol Analysis for an OE item

An Open-ended Item

Soru 2.



Odanın bir köşesine şekildeki gibi kutular konmuştur. Her kutu aynı büyüklüktedir. Odanın köşesinde kaç tane kutu vardır?

1. Doğru Cevap: 18
   Öğrenci Cevabı: (Doğru ise 1 puan, Yanlış ise 0 puan)
2. Zorlanma seviyesi: (1/2/3)
3. Bilişsel strateji ve öz kontrol becerilerini nasıl kullanmış?
    a. Bilişsel strateji davranışı vs eye-tracking sistemi verisi
    b. Öz-kontrol davranışı vs eye-tracking sistemi verisi
4. Sorulara kaç kere dönüş yaptı? Kaç kere odaklandı? Dönüp dönüp bakma ve çözme davranışı var mı?
5. Çözüm yolu

* Eylemleri ile söylemleri bu soruda benzer mi?

Duygular:

EDA verileri görseli

BVP verileri görseli

Göz İlgi Alanı [Area of Interest]:

Soru üzerinde bireysel geçirdiği zaman [Time on task]: soruya başladığı dk ve bana dönüp konuştuğu dk (…… arası)

Accordingly, the thinkaloud process protocol analysis for multiple-choice items were exemplified in the following Table 3.15.

Table 3. 15. Thinkaloud Process Protocol Analysis for a MC item

A Multiple-choice Item

Soru 3.

Oya hesap makinesi ile MP3 çalar, kulaklık ve hoparlörün fiyatını toplamıştır. Elde ettiği sonuç 248'dir.

Oya'nın yanıtı yanlıştır. Oya aşağıdaki hatalardan birini yapmıştır. Oya'nın yaptığı hata aşağıdakilerden hangisidir?

A) Fiyatlardan birini iki kere toplamıştır.
B) Üç fiyattan birini eklemeyi unutmuştur.
C) Fiyatlardan birinin son basamağındaki rakamı yazmamıştır.
D) Fiyatlardan birini toplamak yerine çıkarmıştır.

1. Doğru Cevap: C
   Öğrenci Cevabı: (Doğru ise 1 puan, Yanlış ise 0 puan)
2. Zorlanma seviyesi: (1/2/3)
3. Bilişsel strateji ve öz kontrol becerilerini nasıl kullanmış?
   a. Bilişsel strateji davranışı vs eye-tracking sistemi verisi
   b. Öz-kontrol davranışı vs eye-tracking sistemi verisi
4. Sorulara kaç kere dönüş yaptı? Kaç kere odaklandı? Dönüp dönüp bakma ve çözme davranışı var mı?
5. Çözüm yolu

* Eylemleri ile söylemleri bu soruda benzer mi?

Duygular:

EDA verileri görseli

BVP verileri görseli

Göz İlgi Alanı [Area of Interest]:

Soru üzerinde bireysel geçirdiği zaman [Time on task]: soruya başladığı dk ve bana dönüp konuştuğu dk (…… arası)

Step 5: Data Aggregation Process

Data aggregation is used to support statistical analysis for collated research data and to summarize the data (Dixon & Cunningham, 2009). The collation, curation, and presentation of data are the main steps of data aggregation (see. Figure 3.17). In other words, aggregation of data is the process of compiling numerical or non-numerical data from various sources and/or on different measures, variables, or people into data summaries or summary reports, usually for the purpose of public reporting or statistical analysis. This involves looking at trends, comparing data points, or revealing information and insights that would not be apparent if the data elements were viewed separately. While the majority of an aggregate education dataset is numeric, such as the average amount of money spent per student in a state, graduate and dropout rates, average standardized-test scores for a school or district, or the average amount of funding spent per student in a state, non-numeric data is both available and prevalent, for example, a poll may be taken of a school district's instructors, students, and parents opinions on a topic and the results and comments might be "aggregated" into a report indicating what the surveyed individuals believe and feel about the problem as a whole. Data aggregation is used to conduct statistical analysis and provide a summary of collected research data (Dixon & Cunningham, 2009; Dunstone & Yager, 2009; Leuffen et al., 2013).

In the main study, data of neuroeducation (see. Figure 3.14, Figure 3.15, and Figure 3.16 for data aggregation) was collected from 32 different 10-year-old 5th grade children using the think-aloud process. Information was obtained regarding gender, school district, school size, the number of teachers in the school, and the number of schools in the region. The aim of the study was to ascertain the metacognitive subskills and affective processes used by the students while solving 10 mathematics items. Interviews were transcribed and physiological data was collected using various devices. Two different templates were prepared. Interviews were coded individually for each child using these templates.

In order to expanding upon data aggregation, I transcribed the responses of the students to the think-aloud process protocol one by one. I reread each sentence and each one was coded into which precoded title of metacognitive sub-skill or affective process it corresponded to in excel.xls. format. For example, if a middle school student verbally stated that s/he was rereading, “YES”, if he stated that he did not, “NO”, and even though the student did not give a clear answer even though we asked the question, it was coded as “NOT APPLICABLE” (a.k.a. NA). In the think-aloud process, if a student felt a negative affective process towards any item type and expressed it verbally, before this data was coded as “YES” in the codebook, its mood was confirmed from the EDA data in the software recorded by the smart watch. If I determined that the student really felt worry, this was coded as “YES.” During the think-aloud process, a student was asked what s/he noticed first at the beginning of a question and where s/he started solving the item. The student's response was also confirmed by the data from the eye tracking tool. Since the eye tracking device determines exactly where the student is looking, if this point of view contradicts the student's verbal sentence, the point indicated by the eye tracking device was taken as the correct data and recorded. I also recorded the step-by-step item solution processes via the Wacom Bamboo Tablet, which held the handwriting of the students, in the codebook by giving a code to each step. To summarize, I checked accuracy of the verbally spoken data coming from the students' think-aloud process by cross-checking the quantitative data held within the software and biometric sensors. This is how the qualitative and quantitative data were aggregated and coded in the codebook.

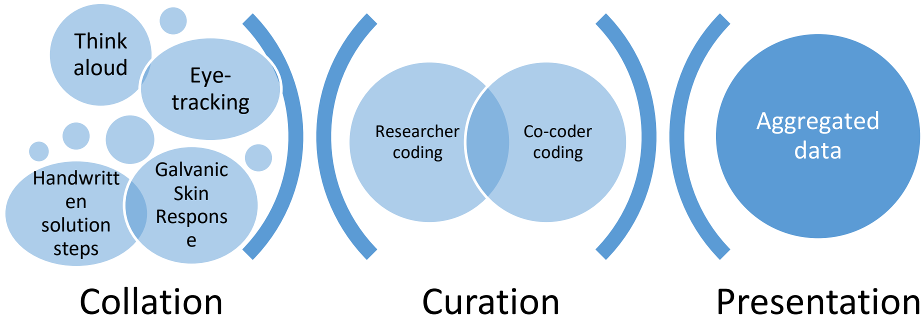


Figure 3. 14. Collapse, define, and data aggregation

The coding was re-coded into another blank excel.xls worksheet so that it could be quantified and analyzed in SPSS version 26 for Mac (see Figure 3.15). Independent samples T-test, Binomial logistic Regression, Chi-Square analyzes were used in the SPSS.

Partci part	Item Type	T (1) / F (0)	Difficulty Level (1-2-3)	Solution Process #	Cognitive Strategy (CS) 1- Organization of knowledge	CS2- Rereading	CS3- More than one strategy	CS4- Different forms/ Re-expression	Looking back #	Self-Checking 1- Check answer	SC2- Process Control	SC3- Find error
P1	MC	1	1	1	Question root	YES	NO	NA	3	NA	YES	NA
P2	OE	1	1	1	Subtraction	NA	NO	MAY BE	1	NA	NA	NA
P3	OE	1	1	1	Subtraction	NA	NO	NA	1	NO	NA	NA
P4	OE	1	1	1	Numbers, Subtraction	NA	NO	NA	1 (numbers)	NO	NO	NA
P4	OE	1	1	1	numbers,subtraction	YES	NO	NO	1 (numbers,question root)	NO	NO	?
P1	OE	1	1	1:multiplication+addition	Question root	NA	NA	NA	3	NO	MAY BE	NA
P2	MC	1	1	1:multiplication+addition	Shape and boxes	NA	NO	NA	2 (Qroot)	YES: alternatives	NA	NA
P3	MC	0	2	1:multiplication+addition	Question root	NO	NO	NO	1 (Qroot)	YES: by drawing	NA	YES
P4	MC	1	2	1:multiplication+addition	Shape, Qroot, counting	YES	NO	NA	2(Qroot, Shape)	NA	NA	NA
P4	MC	1	2	1:multiplication+addition	shape,qroot,counting	YES	NO	NA	5(Qroot,Shape)	NA	NA	NA
P1	MC	1	2	3:alternatives, eliminate, change thinking	Question root, alternatives, 248	YES	YES	NA	6(4timesQroot)	YES	NA	NA
P2	OE	1	2	2: addition, subtraction, comparison	Question root, 248, products, guessing	NA	YES	NA	2	NA	YES	YES
P3	OE	1	2	2: addition, discovery	Question root, guessing	YES	YES	NA	1	NA	YES	YES
P4	OE	0	2?3?	2: addition, binary addition	Shape, Qroot, numbers	YES	YES	NA	3	YES	YES	NA
P4	OE	0	2	2:addition,binary addition	shape,qroot,numbers	NA	no	NA	4	yes	yes	NA
P1	OE	1	3	2: drawing, hypothenuse calculation	ement, Algebra, transformation, Shape -->	YES	YES	YES: different word usage	NA	NO	YES	YES
P2	MC	1	1	2: operation, representation	QRoot, transformation, alternatives	NA	YES	NA	3 (shape)	YES	NA	NA
P3	MC	1	3	3: drawing, hypothenuse calculation, operation	Q root, transformation, hypothenus	YES	YES	NA	3 (Qroot and squares)	YES	YES	YES
P4	MC	1	1	2: Unit transformation + ratio	Shape, alternatives, Qroot	YES	YES	NA	3	NA	NA	NA
P4	MC	1	1	2:unit transformation, ratio	shape, alternatives,groot	yes	yes	na	4	NA	NA	NA
P1	MC	1	1	2: bold root, imagination of perspective	NA	NA	YES	NA	NA	NA	NA	NA
P2	OE	0	1	2: imagination, features of the house	Question root, interpretation	YES	YES	NO	NO	NO	NO	NO
P3	OE	0	1	1: Q root --> shape	Question root, perspective	NA	NO	NO	NO	NO	YES	NA
P4	OE	0	2	1: Q root --> shape	Question root, perspective, interpretation	YES	NO	YES	NO	NO	YES: drawing	NA
P4	OE	0	2	Q root --> shape	question root,perspective,interpretation	yes	no	yes	2	NO	yes	na
P1	OE	0	1	1: multiplication Q root --> shape length	Scaling, unit exchange	YES: units	NO	NA	5(2timesunits)	YES	YES: focus on units	NA

Figure 3. 15. Collapse, Define, and Data Aggregation

Participant	Gender	Time on Task	TotalTime	School_District	School_N	School_Size	St_N	Teacher_N	Iten_Type	Bloom_Knowledge	Bloom_CognitiveProcess	Metacognitive_alg		Metacog_res	True_False	Difficulty_Level	EDA	SolutionProcess_N	CS_Qroot	CS_S
ID1_RA	1	1,00	3,12	1	44	300	20	45	2	3	3	1	0	1	1	1	1	1	1	2
ID4_BMB	1	1,05	1,35	2	50	277	10	13	1	3	3	NA	0	NA	1	1	2	1	1	0
ID2_DY	1	0,57	1,28	2	50	277	10	13	1	3	3	NA	0	NA	1	1	1	1	0	1
ID3_EŞ	1	0,50	1,49	2	50	277	10	13	1	3	3	1	0	1	1	1	1	1	1	2
ID5_EKK	2	1,00	1,03	3	159	1285	21	62	2	3	3	NA	0	NA	1	1	1	1	1	2
ID6_YEE	2	1,00	1,26	2	50	277	10	13	1	3	3	1	0	1	1	1	NA	1	1	2
ID7_KÇ	2	2,00	3,01	2	50	277	10	13	2	3	3	1	0	1	1	1	2	1	0	1
ID8_YBC	2	1,00	1,47	2	50	277	10	13	1	3	3	1	0	1	1	1	2	1	0	1
ID9_EM	2	0,00	1,06	2	50	277	10	13	2	3	3	1	0	1	1	1	1	1	1	2
ID10_ME	2	0,01	1,06	2	50	277	10	13	1	3	3	NA	0	NA	1	1	1	1	0	0
ID11_MGM	1	0,00	2,39	2	50	277	10	13	2	3	3	1	0	1	1	1	1	1	1	2
ID12_BB	2	1,00	1,54	2	50		24		2	3	3	1	0	1	1	1	NA	1	1	2
ID13_BÖ	1	1,00	2,15	1	44	1080	36	58	1	3	3	NA	0	NA	0	1	NA	1	0	0
ID14_AKS	2	0,00	1,21	2	50		24		2	3	3	NA	0	NA	1	1	1	1	0	0
ID15_JT	1	1,00	1,39	2	50	277	10	13	1	3	3	1	0	1	1	1	NA	1	0	0
ID16_AS	2	0,10	1,55	1	44				2	3	3	NA	0	NA	1	1	1	1	1	2
ID17_NM	1	1,00	1,47	3	159	566	70	26	2	3	3	NA	0	NA	1	1	NA	1	1	2
ID18_MK	1	2,00	2,41	3	159	566	70	26	1	3	3	1	0	1	1	1	1	1	0	1
ID19_YÖ	1	2,00	2,48	3	159	566	70	26	2	3	3	1	0	1	1	2	1	1	1	2
ID20_İÖ	1	1,00	1,38	1	44	1080	36	58	2	3	3	1	0	1	0	1	NA	1	1	2
ID21_UB	2	2,00	4,01	3	159	1377	30	66	1	3	3	1	0	2	0	1	NA	2	1	0
ID22_AK*	1	4,00	6,13	2	50	296	26	20	1	3	3	2	0	3	0	2	1	3	1	0
ID23_EG	1	1,00	1,16	1	44	300	20	45	2	3	3	NA	0	NA	1	1	1	1	1	0
ID24_VO	2	1,00	2,32	3	159	566	70	26	1	3	3	1	0	2	1	2	1	1	1	2
ID25_CBY	2	1,00	2,04	3	159	566	70	26	1	3	3	NA	0	NA	1	1	2	1	1	2
ID26_ÖFY	2	0,00	1,44	3	159	566	70	26	2	3	3	NA	0	NA	1	1	1	1	0	1
ID27_YBÜ	2	0,00	0,56	3	159	856	25	44	2	3	3	NA	0	NA	1	1	1	1	0	1

Figure 3. 16. Data Transformation for Quantitative Analysis

Step 6: Finalization Process of Neuroeducation

As indicated in aforementioned steps, data aggregation is the process in which research data is brought together by considering triangulation and compiled in a summary form. It is typically used prior to the performance of statistical analysis. To investigate research question 3, “How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types?, Is there a significant difference between the amount of reflection of students’ cognitive strategy skill levels on their responses to multiple-choice and open-ended items?, Is there a significant difference between the levels of students’ self-checking skill in their responses to multiple-choice and open-ended items?”, and research question 4, “What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)?” I obtained the expert opinion from my advisor and co-advisor in the preparation of each step and conducted statistical analysis.

I ensured that the data was collated, aggregated, and integrated in a meaningful way at each stage. I got the expert opinion and feedback from my advisor and co-advisor during weekly 2-hour meetings. Afterwards, I ran the statistical analyses by checking the analyses that would seek answers to the research questions from the core statistics books (e.g., Field, 2013; Tabachnick & Fidell, 2013). As I elaborated on the results of the study, I realized a large percentage of the data I collected and analyzed in the neuroeducation research process was categorical, in other words, it had a nominal scale of measurement; Chi-Square Test for Independence and Binomial Logistic Regression was selected to conduct an inferential analysis of the neuroeducation data (Tabachnick & Fidell, 2013). The neuroeducation process begun with think-aloud process and ended with data aggregation finalization. The whole picture of this story that I was eagerly involved in came to the light as in Figure 3.17.

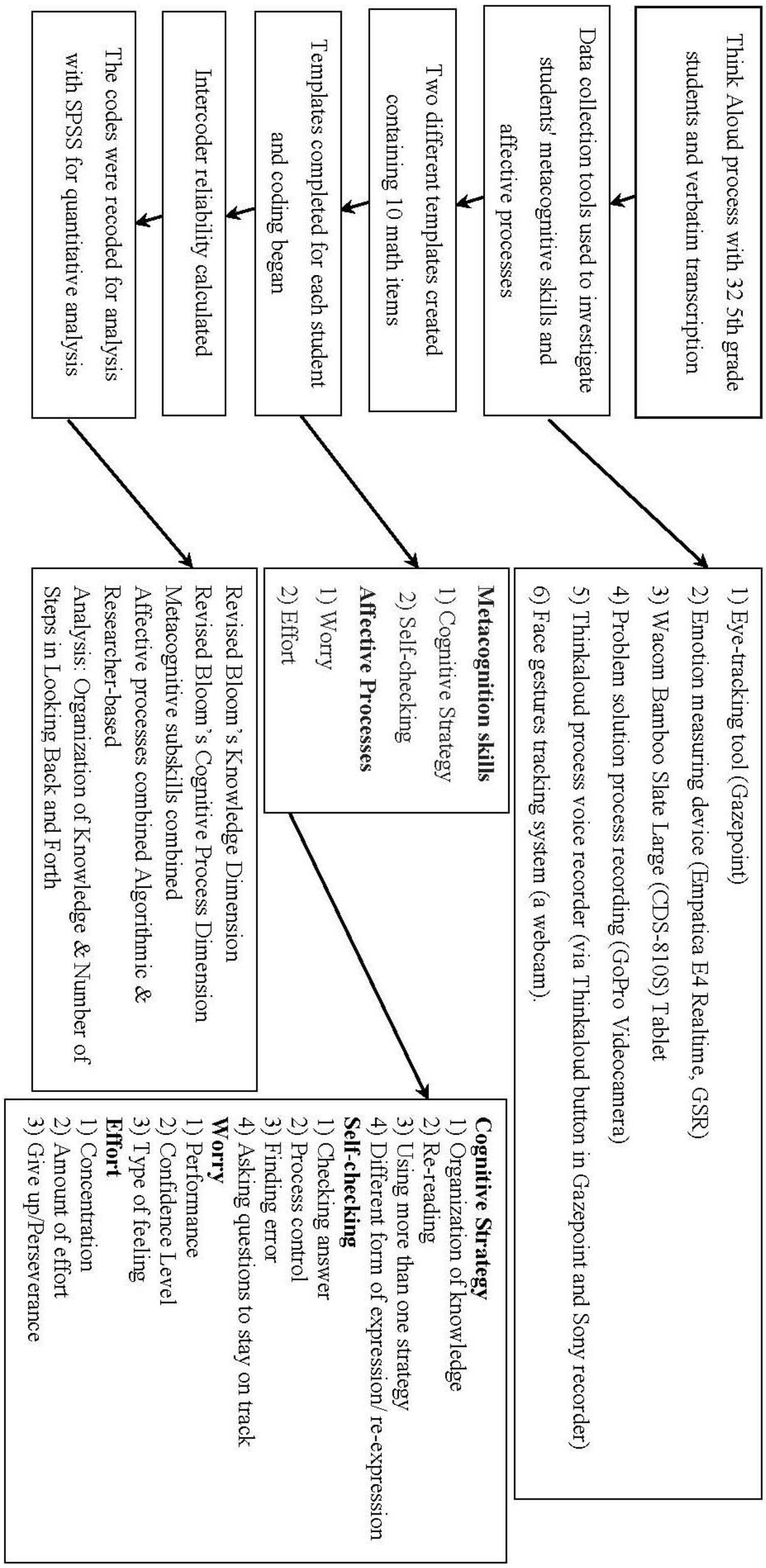


Figure 3. 17. Elements of neuroeducation process

3.3.7. Trustworthiness, Reliability and Validity of the Study Phases

Trustworthiness, in other words, the rigor of the study expresses the degree of assurance in data, explanation, and methods used to ensure the quality of a study. To explain and validate this comprehensive process, I shared the examples or proofs from the data.

Trustworthiness of document analysis. In establishing trustworthiness in qualitative parts of the study, credibility, dependability, transferability, and confirmability were followed by certain steps (Lincoln & Guba, 1985), and authenticity (Guba & Lincoln, 1994). To ensure dependability, I constructed the coding frame that captures the analytically significant descriptions of the data. I shared an empty version of the codebook structure along with the frameworks of Revised Bloom's Taxonomy and TIMSS with two experts; an associate professor from the department of mathematics education and additionally expert in Bloom's taxonomy, and a Ph.D. candidate who was graduated from secondary school mathematics teaching, writing her dissertation in the department of measurement and evaluation, and additionally expert in TIMSS framework. Each coder studied independently on the table of specification and analyzed 10% of the total examination items shared by the researcher. This process was significant due to degree of agreement among two or more independent qualitative coders. Even though there is little consensus between researchers regarding the proportion of the data set that would be shared to assist a trustworthy estimate of intercoder reliability; depending on the size of entire set, 10–25% of data units would be standard (Campbell et al., 2013). Therefore, I randomly selected 10% from a subsample of the items so as to ensure the representativeness of the entire data set.

In order to ensure credibility, I adopted the qualitative research method well established (Campbell et al., 2013; Guba & Lincoln, 1994; Lincoln & Guba, 1985; Patton, 2002). Before the first data collection, I visited the schools starting from the nearest province, its managers and mathematics teacher group who volunteered to participate in this study. The ethical approvals from the university and ministry of national education along with the aims of the study were disseminated to those schools. I had contacted the schools for an early familiarity with the culture of participating schools before the first data collection dialogues took place. I achieved those processes via consultation of appropriate documents and introductory visits to the schools themselves. Hence, a wide range of teachers (n = 10) participated in this part of the study in five schools to triangulate via data sources. The data collection sessions involve only those who are genuinely willing to take part and prepared to offer examination papers freely, so I ensured honesty in teachers when contributing data. I purposefully selected the participant teachers who are knowledgeable, willing to participate, in rapport with me to share their experiences and examinations. Teachers were reached through getting advice from the district national education directorate. I, as a researcher, am sure that these teachers gave the originals of the classroom examinations without hesitation. I always work with my supervisor (a university professor) and get frequent feedback. Qualification, my experience, my supervisors and the experts were explained in detail.

In order to ensure transferability, I adopted purposeful sampling during field search of collecting examination items in schools. The schools participating in this phase of the study were settled in Sarıyer district in İstanbul, Turkey. However, the findings of the qualitative approach are specific to a small number of particular environments (i.e., five schools) and individuals (i.e., 10 teachers). I ensured that a sufficient number of examination papers during school visits was provided from the teachers. However, the examination papers were limited to the level and amount allowed by the school management and what the teachers shared. It was also important that sufficient thick description of the research process under investigation, data collection and analysis were provided to have a proper understanding of it. Besides, I provided the province, the number of schools, the teachers taking part in the study, the number of participants involved in the study, employed data collection methods, and the period of time over which the data was collected. The results are expected to be limited to the participant schools. The results of the study were determined. It will be in their best interest when there are school managers and mathematics teachers who are interested in the results of this study. Hence, research results are transferred to the extent that they are used for their benefits.

In order to ensure confirmability, I analyzed and shared the data in line with the examination items. I reflected the findings objectively also by each examination paper and the schools. I used triangulation to eliminate the effect of investigator bias with the use of document analysis, interviews with the mathematics teachers. I used tables, figures to disseminate the results objectively, shared examples and sample items when needed to demonstrate "audit trail". In addition, I found an audience who were an expert in English language teaching and manager of a publishing house for extrinsic observation of the research process. We discussed the results and exchange of ideas. In order to ensure authenticity, I reflected on "Have people been changed by the process? To what extent did the investigation prompt action?" All examination papers were even-handedly after being collected from the middle school teachers.

Participants understood their situation and why they were asked to share examination items in more informed ways as a result of participation in the research. They knew that this is a part of scientific research. All of the teachers were aware of and had enough consciousness on the fact that participating in this study was also part of gaining experience and changing their constructions as the research participants also developed a better understanding of the alignment between the enacted curriculum and the proposed assessment procedures. By collecting their authentic teacher-made examination items, the study was an authentic representation of their experiences. The participants and I, jointly, assessed the degree of empowerment that evolved during the study.

Validity and reliability of survey phase. In establishing validity and reliability in the quantitative survey phase, I got expert opinion from the advisor and a professor of the department of educational sciences, curriculum and instruction program and a professor the department of educational sciences, measurement and evaluation program during each phase. While writing the items for TMMESP-Q, first expert opinion was ensured for its dimensions and clarity of the items for the constructs. Then, second expert opinion was ensured the constructed survey which became ready for administration to mathematics teachers in Istanbul, Turkey. The Cronbach's alpha reliability for teaching method (TM) preferences part of TMMESP-Q was .75 and the Cronbach's alpha reliability for measurement-evaluation strategy (MES) preferences part of TMMESP-Q was .73. Hence, internal consistency between items were found to be acceptable level in line with rule of thumb Cronbach's alpha $.80 > \alpha \geq .70$ (Nunally, 1978). While the data collected for main study of quantitative survey phase from mathematics teachers in Istanbul, *ecological validity* was tried to be ensured since the delivery of surveys were carried out in the teachers' own environments, namely in the middle schools.

Validity and reliability of multimodal phase. In establishing validity and reliability of multimodal phase (neuroeducation) of the current study, firstly, calibration of eye-tracking and biomarker tools was performed while piloting with fourteen adult participants. The adult participants read the mathematics open-ended and multiple-choice items. They solved and responded the researcher's think-aloud process protocol. They talked to me about the visuality and understandability of the items. The computer screen and software in the laboratory to measure multimodal data were checked beforehand. The system was adjusted by checking whether the system detected the participants' presence and follow what they are looking at in real-time. Then the same calibration was implemented for the research participants. Calibration worked effectively. Even the height of the chair was adjusted for the students, and it was provided to sit in a way that would not disturb the calibration. Whether or not the students focused on the shape into an examination item was monitored from the side screen (there were two screens in the study, one that the student saw and the other that I followed). In cases where there was a shift in the eye-movement, the student was ensured to look at the right place and sit upright. As a matter of fact, one advantage of using multimodal data was to provide validity in laboratory studies, that is, to determine a robust data; the *accuracy and validity* of qualitative data from student think-aloud process and quantitative data from eye-tracking and biomarker tools with my own observations in the research process. In every step, I got an expert view from calibration to finalization of the study. Besides, I studied with inter-coder (see page 149 for detail) to ensure *reliability* of the multimodal data. For external validity, I spent effort to invite the students from different large districts in Istanbul (e.g., Fatih, Beşiktaş, Sarıyer), I tried to keep the sample large in that research conditions, and the study was able to be completed with 32 students.

3.3.8. Limitations and delimitations

This section explains with some limitations and delimitations placed in document analysis, quantitative survey phase, and multimodal phase. Firstly, qualitative data should be collected for a deeper understanding of mathematics teachers' assessment strategies and their corresponding influence on the preparation of authentic teacher-made mathematics examination items. Phenomenological studies would definitely help enrich their authentic experiences and measurement-evaluation strategy preference study. Secondly, school type differences should be considered in this and future research, especially in the case of private schools with university-school partnerships. The analysis of mathematics teachers' authentic teacher-made items found that the quality of in-class examinations in schools whose teachers cooperate with faculty members in the faculty of education from universities take professional development courses. As the assessment is related to professional development, it is an expected result. How teachers reflect on quality in-class examinations by collaborating with academics can be investigated in more detail. Furthermore, regional differences in İstanbul, a metropolitan city, might have affected our results when looking at teacher-made items. Thirdly, mathematics teachers' perception of the curriculum change, of curriculum learning outcomes and mathematical content is admirable. As they are open to changing their teaching methods thanks to their courage in being "agents of change" and their beliefs, it should be elaborated further on why they prefer to follow traditional methods in measurement and evaluation, and why they are not sufficiently flexible. Lastly, the result of students' reactions and responses to differential effect of item types (i.e. MC and OE) and their ability to use metacognitive and affective processes highlight the association between total time on task and gender variables with the rereading skill. However, adequate and in-depth re-explanation skill data could not be obtained from the students to determine how or why they could not. The importance of using self-checking sub skills and how the students could use them if given the opportunity have been investigated in depth. Logistic regression results were based on the deep data that could be collected in a short period of time (i.e. within the time limits allowed for the 5th graders) and may not have shown as a significant variable in the logistic analysis because deep data was not applicable in some themes. Hence, the results should always be interpreted in the light of these limitations.

Despite these limitations, this study contributes to a multi-layered research design on curriculum change, teaching methods and assessment, and teachers' experience and preferences followed by students' reflections on innovative items. The study helps to clarify what kind of hidden metacognitive and affective skills can be discovered if mathematics teachers are professionally developed to make measurement and evaluation of their students with qualified authentic teacher-made items. We have also developed a better understanding of the complexity of students' problem-solving skills, their thinking paths while also managing their cognition, metacognition and affective processes with their natural behaviors. It was observed that the students who were allowed to come up with their own solutions toward mathematics items and who were encouraged to use those skills in the classroom reflected on their inner ideas and processes and believed they made progress in knowledge and skill base. For future studies, special attention should be paid to teachers' autonomy on the curriculum, instruction and assessment to enrich the knowledge base to transform implementations in those areas.

3.3.9. Researcher Experience

I share with my research experience during all research phases one by one as both mathematics teacher since 2011 and research assistant since 2014. During the document analysis phase, nearly all of the school administrations generally volunteered to support the research. However, since the examinations of pre-policy change were archived, they had to be drawn up, reviewed and shared with me by a volunteer teacher. For this reason, I could only get pre-policy change examinations from solely two schools. If more schools had archived all semester examinations online, the teachers would have quickly accessed old examinations and I could collect more data. They could not provide sufficient data on the difference between pre policy and after policy change.

During the quantitative survey phase, some mathematics teachers who did not want to participate in the interview during the quantitative survey phase showed up to me and said, "This research is of no use to me. I don't think it will benefit of my students”, “There is no project in this study and I have no income.” They used sentences such as "I did not want to help." As a researcher, I suggest arranging a meeting with other teachers without losing the motivation. School principals and teachers should better understand the importance of research and more volunteered action should be taken in this direction. You can motivate the participating teachers with small gifts or incentives for participating in the research. Although I received ethical permission from the MoNE, I rejected through the doors of some middle schools and was not allowed to meet with mathematics teachers. However, good memories remained in my mind. Some middle schools welcomed me. Since I arrived on their lunch break, they hospitably offered me to accompany the meal. I even reached the schools that were far away in the Fatih district. The manager of a public school was very happy that I chose them as a participatory school. I met all of their mathematics teachers and was able to gather data efficiency.

Moreover, as a mathematics teacher and an educational scientist, I read a lot about the topic for three months to become comfortable with the educational neuroscience, subject matter, and objectives in order to adjust to this new laboratory environment without feeling like square peg in round hole. The fact that my co-advisor had my personnel ID card loaded into the university system so I could enter and depart the room was the chance of helping me adapt to the experimental atmosphere. I had not yet had the chance to use the laboratory, so this gave me a chance to get accustomed to it. I practiced with the tools and grew confident using them, even resolving issues and problems once they emerged. I started modifying the arrangements for pilot studies as I grew accustomed to the surroundings. I discovered the significance of instrument calibration through scientific journals. I eventually became so accustomed to being there that I occasionally worked from the laboratory. I think that getting accustomed to and adjusting to one's surroundings is a crucial first step in the research process, thus one should take steps to prevent risks to internal validity for the scientific study process.

CHAPTER 4

RESULTS

In the results section, the findings from each five phases are explained. The chapter begins with the results of document analysis of authentic teacher-made items. The document analysis sectioned as examination items related to item types, learning units, learning outcomes, revised Bloom's taxonomy, TIMSS framework. The chapter continues with the results of quantitative survey including construct validity procedures of the TMMESP questionnaire and descriptive and inferential statistical findings from the questionnaire. Then, the chapter carries on the results of multimodal phase (i.e. neuroeducation). The chapter ends with a design of a deep data modeling.

4.1. Results of Document Analysis of Authentic Teacher-Made Examinations

The results of document analysis of authentic teacher-made examinations were consisted of five substeps. First, the results were examined in terms of item types; second, the results were examined in terms of learning units; third, the results were examined in terms of learning outcomes in the middle school mathematics curriculum; forth, the results were examined in terms of revised Bloom's taxonomy; fifth, the results were examined in terms of Trends in International Mathematics and Science Study [TIMSS]'s Framework. The findings related to document analysis based on teacher-made examination items (N = 380) indicated that most of the items (n = 364, 95.8%) were delivered to the researcher by public school teachers and few of the items delivered by private school teachers (n = 16, 4.2%).

4.1.1. Examination items related to item types

Authentic teacher-made items analyzed related to item types. In terms of the type of items, the findings revealed that 1.8% of items were constructed as fill in the blanks (*n* = 7), 1.3% of items as true/false (*n* = 5), 53.7% of the items as multiple-choice (*n* = 204), and 43.2% of them as restricted open-ended (*n* = 164). The descriptive statistics for item types were illustrated in Table 4.1.

Table 4. 1. Items by Item Type

Item types	*f*	%
Fill in the blanks	7	1.8
True/false	5	1.3
Multiple choice	204	53.7
Open-ended (constructed OE)	164	43.2
Total	380	100

4.1.2. Examination items related to learning units

In terms of learning units, the items were examined in relation to the intended mathematics curriculum content for middle school students. The results revealed that items were mostly prepared about Numbers and Operations (81%). On the other hand, Measurement (0.3%), Geometry and Measurement (1.1%), and Data Analysis (0.8%) were scarcely used to prepare in-class examination items. Interestingly the findings showed in Table 4.2 that teachers who were teaching in 5th grade prepare their items not only aligned with 5th-grade learning outcomes but also 3rd, 4th, 6th and 7th-grade learning outcomes with a limited emphasis. Generally, items were found to be related to the content domain of Numbers and Operations (*f* = 288, 75.8%) followed by Geometry and Measurement (*f* = 4, 1.1%), and to a lesser extent, Data Processing (*f* = 3, 0.8%) adjusted with MoNE 5th-grade mathematics curriculum. The Learning units within the Table 4.2 see below.

Table 4. 2. Items that Related to Learning Units in the Middle School Mathematics Curriculum

Learning units (i.e., content domain)	*f*	%
M.3.1. Numbers and Operations	2	0.5
M.4.1. Numbers and Operations	10	2.6
M.4.3. Measurement	1	0.3
M.5.1. Numbers and Operations	288	75.8
M.5.2. Geometry and Measurement	4	1.1
M.5.3. Data Processing	3	0.8
M.6.1. Numbers and Operations	7	1.8
M.7.1. Numbers and Operations	1	0.3
Total	380	100

4.1.3. Examination items related to learning outcomes

The examination items were analyzed in terms of 5^{th} grade mathematics curriculum learning outcomes they were related with. The results including items that were identified as Higher-order Thinking Skills (HoT) and Lower-order Thinking Skills (LoT) are revealed in Table 4.3.

Table 4. 3. Levels of Learning Outcomes

	Objectives	Categories	Themes	*f*	%
Higher-order thinking skills	M.5.2.3.2. form different shapes. M.6.1.1.4. solve and pose problems.	Constructing	Produce new or original work	5	1.4
	M.5.2.1.6. interpret whether it is parallel.	Judging	Justify a stand or decision	3	0.8
	M.4.1.2.2. compare its forecast with the result of the solution. M.4.3.1.4. solve problems. M.5.1.2.12. solve problems related to basic operations. M.5.1.3.3. compare fractions. M.5.1.4.1. do calculation and interpretation. M.5.1.4.2. solve and pose problems. M.5.1.5.3. understand the relation. M.5.1.6.3. compare quantities. M.5.2.3.1. transform and solve related problems. M.5.2.3.3. transform and solve related problems. M.5.2.4.4. solve problems. M.5.3.1.1. design research questions. M.5.3.1.3. solve problems. M.7.1.5.4. solve problems related to percentages.	Comparing, Problem-solving, Relating, Problem-posing, Questioning	Draw connections among ideas	62	17.0
Lower-order thinking skills	M.3.1.1.4. round numbers to the nearest hundred. M.4.1.1.4. round. M.4.1.6.1. find fractions and demonstrate with models. M.5.1.1.3. sequence the expected steps. M.5.1.2.1. do subtraction. M.5.1.2.10. demonstrate and calculate the value. M.5.1.2.11. find the results. M.5.1.2.2. determine the strategy and use it. M.5.1.2.4. multiply. M.5.1.2.5. divide. M.5.1.2.7. determine the strategy and use. M.5.1.2.9. find the unknowns. M.5.1.3.1. demonstrate and sequence. M.5.1.3.2. transforms.	Rounding, Sketching, Implementing, Demonstrating, Operating, Using, Sequencing, Demonstrating, Transforming, Drawing, Calculating	Use information in new situations	212	55.0

Table 4.3. (continued)

	Objectives	Categories	Themes	*f*	%
	M.5.1.3.4. represent fractions M.5.1.3.5. sequence the equivalent fractions. M.5.1.3.6 operate. M.5.1.5.5. demonstrate and sequence. M.5.1.5.6. do addition and subtraction. M.5.1.6.2. transform to each other. M.5.1.6.4. find the amount of correspondence. M.5.2.1.3. sketch a line segment. M.5.2.2.3. determine and draw. M.5.2.2.4. determine and find the unknown angles. M.5.2.4.1. calculate the area. M.6.1.1.1. calculate the value. M.6.1.5.1. sketch on the number line. M.6.1.5.2. do subtraction.				
	M.4.3.4.1. explain the relation. M.5.1.2.3. guess. M.5.1.2.6. guess the results. M.5.1.2.8. discuss. M.5.2.1.2. express by using unit and direction. M.5.2.2.2. classify.	Explaining, Guessing, Discussing, Classifying	Explain ideas or concepts	25	6.7
	M.5.1.1.1. read and write. M.5.1.1.2. state. M.5.1.5.1. determine the expression. M.5.1.5.4. read and write. M.5.1.6.1. state with the symbol of percentage (%). M.5.2.1.1. using a symbol. M.5.2.1.4. determine obtuse-angled triangle. M.5.2.2.1. determine the basic elements. M.6.1.2.4. determine prime factors.	Reading, Writing, Determining, Stating, Memorizing, Using	Basic concepts	73	19.1
	TOTAL			380	100

I connected each teacher-made items with learning outcomes founded in the MoNE mathematics curriculum instead of mathematics teachers that was the most essential part of the above table. In this stage of document analysis, I analyzed the learning outcomes connected with teacher-made examination items. I intended to investigate the level of thinking skills (i.e., higher-order thinking skills [HoTs] or lower-order thinking skills [LoTs]) (Brookhart, 2010) for each mathematics items. Pertaining to my aim, I read each learning outcomes' verb part to grasp its skills and then combine them into meaningful categories. The similar sets of techniques were used to analyze textual data and elucidate the themes according to the qualitative content analysis (Vaismoradi et al., 2016). More specifically, the categories from each learning outcomes were determined and categorized as follows: Students will be able to "M.5.1.6.3. compare the quantities" coded as "Comparing", "M.5.2.3.2. construct different shapes" as "Constructing", "M.5.3.1.1. construct research questions" as "Questioning". "M.5.2.3.1. transform and solve related problems" as "Problem-solving." Among 380 learning outcomes 17 individual outcomes (f = 70) were figured out and six categories revealed. More than one learning outcome were assigned to the same code. These are 1) constructing, 2) judging, 3) comparing, 4) problem-solving, 5) relating, 6) problem-posing, 7) questioning. Hence, these categories combined into three main themes such that (1) Produce new or original work, (2) Justify a stand or decision, (3) Draw connections among ideas emerged from the document analysis and related with HoTs (for detail of Analyzing, Evaluating, Creating level see Anderson and Krathwohl, 2001). Hence, only 70 learning outcomes (19.2%) pointed out items assessing HoTs.

Moreover, the other learning outcomes determined and categorized as follows: students will be able to "M.4.1.6.1. know fractions and demonstrate its model" coded as "demonstrating", "M.5.1.2.1. do subtraction" coded as "calculating", "M.5.1.3.2. transform to each other" coded as "transforming", "M.5.1.3.5. sequence the equivalent fractions" coded as "sequencing." Data depicted that most of the examination items were prepared to measure the same learning outcomes. Among 380 learning outcomes, 28 individual outcomes (f = 212) were matched with the related items and eleven categories revealed. More than one learning outcomes were assigned to the same category. These are 1) Rounding, 2) Sketching, 3) Implementing, 4) Demonstrating, 5) Operating, 6) Using, 7) Sequencing, 8) Demonstrating, 9) Transforming, 10) Drawing, 11) Calculating. Hence, these categories combined into one main theme such that (1) Use information in new situations emerged from the content analysis and related with LoTs (for detail of Applying level see Anderson and Krathwohl, 2001). In addition, students will be able to "M.5.1.2.3. guess", "M.5.1.2.6. guess the results" coded as "Guessing", "M.5.1.2.8. interpret" coded as "Explaining." Among 380 learning outcomes, 6 individual outcomes (*f* = 25) were figured out and four categories revealed. More than one learning outcomes were assigned to the same code. These are 1) Explaining, 2) Guessing, 3) Discussing, 4) Classifying. Hence, these categories combined into one main theme such that (1) Explain ideas or concepts emerged from the content analysis and related with LoTs (for detail of Understanding level see Anderson and Krathwohl, 2001). Lastly, Students will be able to "M.5.1.1.2. determine" and "M.5.1.5.1. determine the expression of" coded as "Determining" whereas "M.5.1.5.4. write and read" coded as "Writing" and "Reading", "M.5.1.6.1. show percentage symbol as %" coded as "Stating." Among 380 learning outcomes, 9 individual outcomes (*f* = 73) were figured out and five categories revealed. More than one learning outcome was assigned to the same code. These are 1) Reading, 2) Writing, 3) Determining, 4) Stating, 5) Memorizing.

Hence, these categories combined into one main theme such that (1) Basic concepts emerged from the content analysis and related to LoTs (for detail of Remembering level see Anderson and Krathwohl, 2001). Hence, totally, 310 learning outcomes (80.8%) pointed out items assessing LoTs. 19.2% of the objectives (*f* = 70) in 5th grade national mathematics curriculum was found to be related with **higher-order thinking skills** (HoTs) while 80.8% of them (*f* = 310) were found to be related with lower-order thinking skills (LoTs).

4.1.4. Examination items related to the revised bloom's taxonomy

In this part, I analyzed all authentic teacher-made examinations and items in total. Then, I examined each item in the examinations by the middle schools and the administered semesters respectively. Also, the analysis findings were demonstrated in the charts of Table 4.4 (see. Appendix P for cross-case school analysis).

I distributed the items to examine in terms of pre and after policy change. For this process, I splitted them into two parts as those reflecting pre-change and those reflecting after-change. The numbers were illustrated in Table 4.4 related to schools, school types, examination semesters, total numbers.

Table 4. 4. Number of Teacher-made Examinations by Prior to and After Policy Change

School Name	School Type	Pre-Change	*f*	After-Change	*f*
School 1	Public School	1. 2016-2017 1st semester 1st exam	14	1. 2017-2018 1st semester 1st exam	16
		2. 2016-2017 2nd semester 1st exam	17	2. 2017-2018 1st semester 2nd exam	18
				3. 2017-2018 1st semester 3rd exam	20
				4. 2017-2018 2nd semester 1st exam	12
				5. 2017-2018 2nd semester 2nd exam	20
	Total		31		86
School 2	Public School	-		1. 2017-2018 1st semester 1st exam	17
				2. 2017-2018 1st semester 2nd exam	18
				3. 2017-2018 1st semester 3rd exam	16

Table 4.4. (continued)

School Name	School Type	Pre-Change	*f*	After-Change	*f*
				4. 2017-2018 2nd semester 1st exam	15
				2017-2018 2nd semester 2nd exam	21
	Total		0		87
School 3	Public School	-		1. 2017-2018 1st semester 1st exam	20
				2017-2018 2nd semester 1st exam	25
	Total		0		45
School 4	Public School	1.2016-2017 1st semester 1st exam	19	1. 2017-2018 1st semester 1st exam	20
		2. 2016-2017 2nd semester 1st exam	19	2. 2017-2018 1st semester 2nd exam	19
				3. 2017-2018 1st semester 3rd exam	20
				2017-2018 2nd semester 1st exam	20
	Total		38		79
School 5	Private School	-		1. 2018-2019 1st semester 2nd exam	14
	Total		0		14
GRAND TOTAL			69		311

The results revealed in Table 4.5 that teacher-made examination items mostly relied on Procedural level of knowledge dimension (f = 228, 60%) and Applying level of cognitive process dimension (f = 217, 57.11%) in line with the revised Bloom’s taxonomy.

Table 4. 5. Distribution of Items in the Total of Teacher-made Examinations (*N* = 380)

Taxonomy	*Dimension*	*f*	*%*
	Knowledge		
Revised Bloom Taxonomy	Factual	55	14.47
	Conceptual	96	25.26
	Procedural	228	60
	Metacognitive	1	0.26
	Cognitive Process		
	Remember	44	11.58
	Understand	91	23.95
	Apply	217	57.11
	Analyze	25	6.58
	Evaluate	3	0.79
	Create	0	0
	Total		100

When the schools were examined by pre and after policy change, the findings in Table 4.6 showed that the School 1 did not reflect a significant difference between pre and after change in terms of knowledge and cognitive process dimension. For instance, in terms of knowledge dimension there is a very slight decrease in factual knowledge whereas a slight increase in conceptual and procedural knowledge. In terms of cognitive process dimension, there is a decrease in remembering and analyzing skill whereas slightly increase in applying skill. Findings from School 2 could not be compared in terms of pre and after policy change because the teachers from School 2 did not share pre-policy examination items. In terms of knowledge dimension, it was as similar as with School 1. In terms of cognitive process dimension, the items from School 2 in remembering skills were nearly two times higher than those in School 1. Findings from School 3 could not be compared in terms of pre and after policy change because the teachers from School 3 did not share pre-policy examination items. In line with knowledge dimension, there was a profound difference between the items in procedural knowledge in School 3 than School 1 and School 2. In terms of cognitive process dimension, the items from School 3 in applying skill were almost higher than those in School 1 and School 2. However, the number of items in understanding skill was lower than those in School 1 and School 2.

In addition, findings from School 4 examined that the items regarding knowledge dimension were nearly similar in terms of pre and after policy change. Yet, the items regarding cognitive process dimension showed that the number of items in remembering skills from after-change policy were two times higher than those from pre-change policy. The number of items in applying skill from after-change policy slightly decreased. Moreover, when the findings from the School 5 examined, it showed that the pre and after policy change could not be compared because the teachers from School 5 did not share pre-policy examination items. In terms of knowledge dimension, the number of items in procedural level was quite high compared to other schools. In terms of cognitive process dimension, the number of items in applying skill was almost lower than the other schools. Instead, the items in analyzing skills were significantly higher than the other schools.

Table 4. 6. Distribution of Items Related to Schools vs. Pre and After Policy Change (f (%))

Dimension	School 1		School 2		School 3		School 4		School 5	
	Pre	After	Pre	After	Pre	After	Pre	After	Pre	After
Knowledge										
Factual	6(19.35%)	14(16.28%)	0	17(19.54%)	0	6(13.33%)	3(7.89%)	9(11.39%)	0	0
Conceptual	7(22.58%)	20(23.26%)	0	29(33.33%)	0	5(11.11%)	8(21.05%)	27(34.18%)	0	0
Procedural	18(58.06%)	52(60.47%)	0	41(47.13%)	0	34(75.56%)	27(71.05%)	43(54.43%)	0	13(92.86%)
Metacognitive	0	0	0	0	0	0	0	0	0	1(7.14%)
Cognitive Process										
Remember	5(16.13%)	8(9.30%)	0	15(17.24%)	0	6(13.33%)	3(7.89%)	15(18.99%)	0	0
Understand	5(16.13%)	22(25.58%)	0	27(31.03%)	0	3(6.67%)	8(21.05%)	18(22.78%)	0	0
Apply	18(58.06%)	50(58.14%)	0	45(51.72%)	0	35(77.78%)	25(65.79%)	41(51.90%)	0	3(21.43%)
Analyze	3(9.68%)	6(6.98%)	0	0	0	1(2.22%)	2(5.26%)	5(6.33%)	0	8(57.14%)
Evaluate	0	0	0	0	0	0	0	0	0	3(21.43%)
Create	0	0	0	0	0	0	0	0	0	0
Total	**31**	**86**	**0**	**87**	**0**	**45**	**38**	**79**	**0**	**14**

Long story short, mathematics teachers in School 5 remarkably, prepared examination items related to applying, analyzing and evaluating cognitive process dimension which were distinctly different from the public schools (see. Figure 4.1, Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5 below).

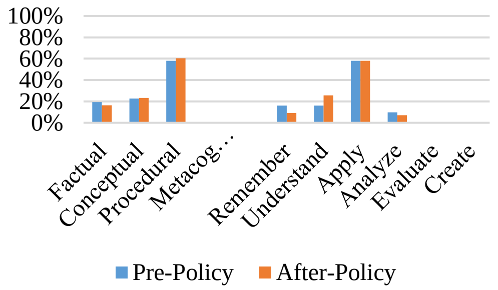


Figure 4. 1. Distribution of the Revised Bloom's Taxonomy Subdimensions in School 1 by Policy

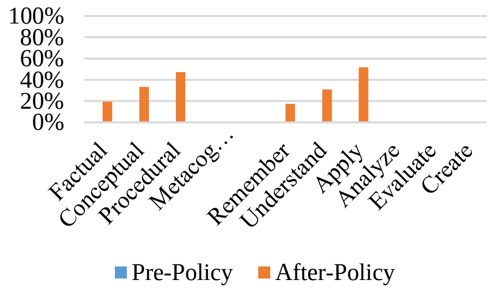


Figure 4. 2. Distribution of the Revised Bloom's Taxonomy Subdimensions in School 2 by Policy

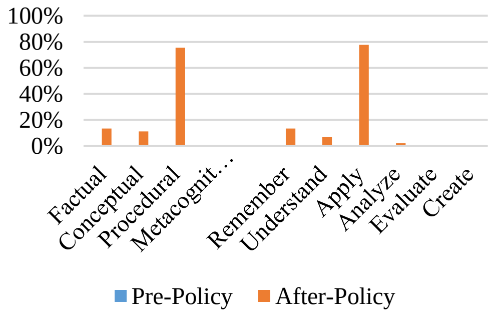


Figure 4. 3. Distribution of the Revised Bloom's Taxonomy Subdimensions in School 3 by Policy Change

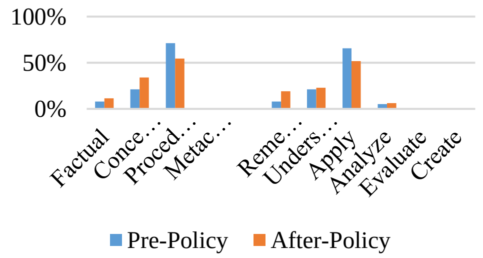


Figure 4. 4. Distribution of the Revised Bloom's Taxonomy Subdimensions in School 4 by Policy

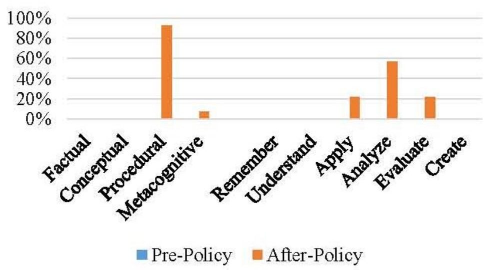


Figure 4. 5. Distribution of the Revised Bloom's Taxonomy Subdimensions in School 5 by Policy Change

4.1.5. Examination Items Related to TIMSS Framework

The examination items were explored related to TIMSS Framework. The results in Table 4.7 revealed that teacher-made examination items (*N* = 380) mostly relied on Knowing level of main domain (*f* = 331, 87%) and almost half of them at Computing level of sub domain (*f* = 164, 43%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework. (see. Figure 4.6 distribution of the TIMSS framework domains: All in one)

Table 4. 7. Distribution of Items in the Total of Teacher-made Examinations by TIMSS (N = 380)

	Dimension	*f*	*%*
	TIMSS Main Domain		
Revised Bloom Taxonomy	Knowing	331	87
	Applying	27	7
	Reasoning	22	6
	TIMSS Sub Domain		
	Recall	54	14
	Recognize	52	14
	Classify/Order	36	9
	Compute	164	43
	Retrieve	20	5
	Measure	5	1
	Determine	12	3
	Represent/Model	1	0
	Implement	17	5
	Analyze	12	3
	Integrate/Syntheses	2	1
	Evaluate	2	1

Table 4.7. (continued)

Dimension	*f*	%
Draw Conclusion	0	0
Generalization	1	0
Justify	2	1
Total	380	100

The overall analysis of examination items related to the TIMSS framework Main and Subdomain are illustrated in Figure 4.6 below.

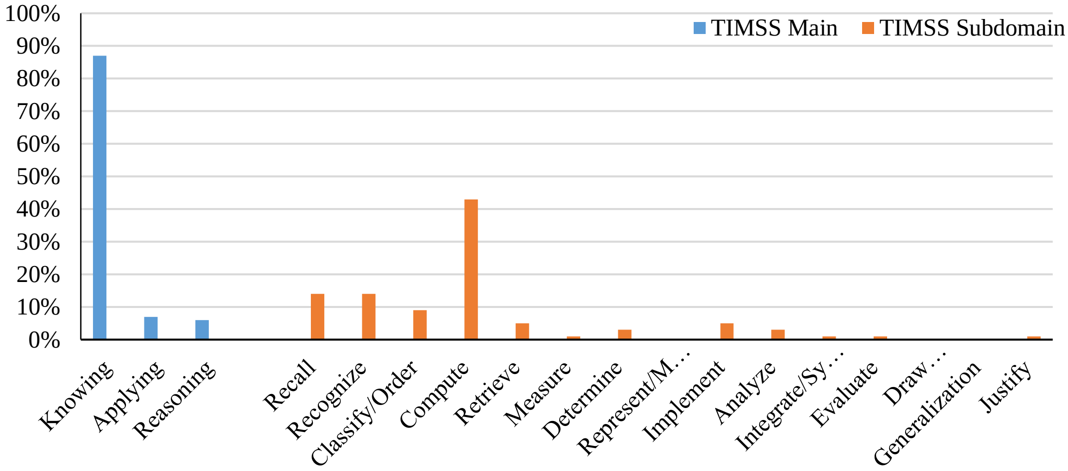


Figure 4. 6. Distribution of the TIMSS Framework Domains: All in One

4.1.6. Comparison of the findings from document analysis of phase 1

Regarding the results from 4.1.4 and 4.1.5; the level of knowledge and cognitive process dimensions of the Revised Bloom's Taxonomy, mathematics items were also tested and Table 4.6 revealed larger numbers for procedural knowledge and apply subdimensions in the level of knowledge dimension and cognitive process dimension, respectively. To be more precise, mathematics items reflected procedural (f = 228, 60%), conceptual (f = 96, 25.3%), factual (f = 55, 14.5%), and metacognitive (f = 1, 0.3%) level of knowledge; applying (f = 217, 57.1%), understanding (*f* = 91, 23.9%), remembering (*f* = 44, 11.6%), analyzing (*f* = 25, 6.6%), evaluating (*f* = 3, 0.8%). Nevertheless, any finding indicated creating level of cognitive process dimension. The results on the TIMSS framework showed in Table 4.7 that when analyzed the authentic teacher-made examination items, the number of items categorized into knowing cognitive dimension (*f* = 331, 87.1%) was higher than applying (*f* = 27, 7.1%) and reasoning (*f* = 22, 5.8%) cognitive dimensions. When looked at an international level, it was revealed that they were prepared at a lower level for fifth grade students in Turkey. When cognitive subdomains were analyzed deeply to examine which skills the items had been measuring, the findings showed us that in knowing cognitive dimension, computing skills were measured in high frequency (*f* = 164, 43.4%) than recalling (*f* = 54, 14.2%), recognizing (*f* = 52, 13.7%), classifying/ordering (*f* = 36, 9.5%), retrieving (*f* = 20, 5.3%), and measuring (*f* = 5, 1.3%) skills respectively. Hence, the categories of knowing seemed to be almost equally distributed except computation, and as expected, very low frequency of the items in measuring subdimension. This might be due to there is no significant number of learning outcomes from Measurement learning unit within the mathematics curriculum.

More specifically, in terms of the applying cognitive dimension, results indicated that 4.5% of the items reflected implementing skills (*f* = 17) and 3.2% of the items reflected determining skills (*f* = 12) whereas only 0.3% of them reflected representing or modeling skills (*f* = 1). In a related manner, with respect to the reasoning cognitive dimension, results showed that the majority of the items reflected analyzing skills (3.2%, *f* = 12), while 0.5% reflected integrating/synthesizing (*f* = 2), 0.5% evaluating (*f* = 2) and 0.5% justifying (*f* = 2) and 0.3% generalizing (*f* = 1) skills/subdimensions. Not surprisingly, none of the items are reflected by concluding sub dimensions, which require one of the higher-order thinking skills. Moreover, other related higher-order skills such as generalization or evaluation had one of the lowest frequencies among reasoning dimensions. This might be due to there being no relation between applying cognitive processes from Revised Bloom taxonomy and applying from TIMSS framework. They are not at a similar difficulty level.

Viewed together, these findings inferred that the mathematics items in authentic teachers made exams mainly focus on the routine use of procedures (e.g., automatized solution procedures). The students can often utilize procedural applications during the solution process. However, it is worth to note that some restricted open-ended items prepared by private schools had features of measuring metacognitive knowledge of students, analyzing, and evaluating, synthesizing, and justifying skills. Mathematics items need to be structured at a more advanced level during the in-class assessment so that students can reflect their differentiation, evaluation, making an inference, modeling skills during international large-scale assessments such as PISA, TIMSS etc.

4.1.7. Overall summary from document analysis: Revisited

Totally the teacher made items were delivered by *10 mathematics teachers* from 5 different public (62%) and private schools (37%) located in lower-middle and middle SES districts in Turkey. In this part, I will disseminate the results by each different school voluntarily participating in this research process.

All 380 authentic teacher-made mathematics examination items were analyzed regarding national middle school mathematics curriculum learning outcomes and subject area (i.e., mathematics units). Specifically, the findings from the items revealed that the middle school mathematics teachers tend to prepare test items in 5^{th} grade frequently based on the basic unit of Numbers and Operations, and MC and restricted OE in terms of item types. They prepared items in line with curriculum learning outcomes. Nevertheless, these learning outcomes were found to be above (i.e., 6th or 7^{th} grade) or below (i.e., 3^{rd} or 4^{th} grade) the intended 5^{th} grade level on which the items were developed. One-fifth of the learning outcomes (f = 70) in 5^{th} grade national mathematics curriculum was found to be related with *higher-order thinking skills* (HoTs) while four-fifths of them (*f* = 310) were found to be related with *lower-order thinking skills* (LoTs).

All authentic teacher-made items (*N* = 380) analysis delivered by *10 mathematics teachers* in five different schools showed that 13 examinations were prepared according to 1st semester learning outcomes whereas eight were prepared according to 2nd semester learning outcomes. In relation to pre and after policy change, in other words, 69 examination items were prepared in the pre-policy change and 311 were prepared after-policy change. From 10 mathematics teachers, 55 items in *Factual* (14.47%), 96 items in *Conceptual* (25.26%), 228 items in *Procedural* (60%), 1 item in *Metacognitive* (0.26%) level in line with Knowledge dimension of the Revised Bloom Taxonomy; 44 items in *Remembering* (11.58%), 91 items in *Understanding* (23.97%), 217 items in *Applying* (57.11%), 25 items in Analyzing (6.58%), 3 items in Evaluating (0.79%), and no item in Creating level in line with Cognitive process dimension of the Revised Bloom Taxonomy. To sum, there were totally 21 examinations and 380 individual items; the results revealed that teacher-made examination items mostly relied on *Procedural* level of knowledge dimension (*f* = 228, 60%) and *Applying* level of cognitive process dimension (*f* = 217, 57.11%) in line with the Revised Bloom's Taxonomy.

Authentic teacher-made items analysis delivered by *two mathematics teachers* in School 1 showed that four examinations were prepared according to 1st semester learning outcomes whereas three were prepared according to 2nd semester learning outcomes. There were a total of seven examinations and 117 individual items. The results revealed that teacher-made examination items in School 1 mostly relied on Procedural level of knowledge dimension (*f* = 70, 60%) and Applying level of cognitive process dimension (*f* = 68, 58%) in line with the Revised Bloom's Taxonomy. Moreover, School 1 Middle School (*n* = 117) mostly relied on Knowing level of main domain (*f* = 101, 86%) and Computing level of sub domain (*f* = 50, 43%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework.

Authentic teacher-made items analysis delivered by *three mathematics teachers* in School 2 showed that three examinations were prepared according to 1st semester learning outcomes whereas two were prepared according to 2nd semester learning outcomes. There were totally five examinations and 87 individual items. The results revealed that teacher-made examination items in School 2 mostly relied on Procedural level of knowledge dimension (*f* = 40, 46%) and Applying level of cognitive process dimension (*f* = 44, 51%) in line with the Revised Bloom's Taxonomy. Moreover, School 2 Middle School (*n* = 87) mostly relied on Knowing level of main domain (*n* = 82, 94%) and Computing level of sub domain (*f* = 35, 40%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework.

Authentic teacher-made items analysis delivered by *two mathematics teachers* in School 3 showed that one examination was prepared according to 1st semester learning outcomes whereas the other one was prepared according to 2nd semester learning outcomes. There were totally two examinations and 45 individual items. The results revealed that teacher-made examination items in School 3 mostly relied on Procedural level of knowledge dimension (*f* = 34, 76%) and Applying level of cognitive process dimension (*f* = 35, 78%) in line with the Revised Bloom's Taxonomy. Moreover, School 3 Middle School (*n* = 45) mostly relied on Knowing level of main domain (*n* = 43, 96%) and Computing level of sub domain (*f* = 24, 54%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework.

Authentic teacher-made items analysis delivered by *three mathematics teachers* in School 4 showed that four examinations were prepared according to 1st semester learning outcomes whereas two were prepared according to 2nd semester learning outcomes. There were totally six examinations and 117 individual items. The results revealed that teacher-made examination items in School 4 mostly relied on Procedural level of knowledge dimension (*f* = 70, 60%) and Applying level of cognitive process dimension (*f* = 66, 56%) in line with the Revised Bloom's Taxonomy. In addition, School 4 Middle School (*n* =

117) mostly relied on Knowing level of main domain (*n* = 104, 89%) and Computing level of sub domain (*f* = 54, 46%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework.

Authentic teacher-made items analysis delivered by *one mathematics teacher* in School 5 showed that one examination was prepared according to 1st semester learning outcomes. There was a total of one examination and 14 individual items. The results revealed that teacher-made examination items in School 5 mostly relied on Procedural level of knowledge dimension (*f* = 13, 93%) and Analyzing level of cognitive process dimension (*f* = 8, 57%) in line with the Revised Bloom's Taxonomy. Furthermore, School 5 Middle School (*f* = 14) mostly relied on Reasoning level of main domain (*f* = 11, 79%) and Analyzing level of sub domain (*f* = 4, 29%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework.

The findings revealed that teachers tend to use traditional objective testing mostly. Regarding the level of knowledge and cognitive process dimensions of the Revised Bloom's Taxonomy, most of the mathematics items reflected Procedural (*f* = 228, 60%), a forth Conceptual (*f* = 96, 25.3%), and some Factual (*f* = 55, 14.5%), and rarely Metacognitive (*f* = 1, 0.3%) level of knowledge dimension. Regarding cognitive process dimension, half of the teachers prepared Applying (*f* = 217, 57.1%), about a fifth Understanding (*f* = 91, 23.9%), 11 percent Remembering (*f* = 44), and a some Analyzing (*f* = 25, 6.6%), and few Evaluating (*f* = 3, 0.8%) levels were preferred respectively. No teacher used Creating the level of the cognitive process dimension. The complementary findings revealed that teacher-made examination items (*f* = 380) mostly relied on Knowing level of main domain (*f* = 331, 87%) and Computing level of sub domain (*f* = 164, 43%) in line with the Trends in International Mathematics and Science Study [TIMSS]'s Framework. To sum, findings revealed curriculum change did not assure full renewal of teacher practices. Also, teacher-made items meet international standards at a very basic level.

4.2. Results of Quantitative Survey Phase: Investigation of Teachers' Teaching Method and Measurement-Evaluation Strategy Preferences

This part yielded construct validity procedures of the quantitative survey for each scale of TMMESP-Q, and descriptive and inferentials statistics from TMMESP-Q.

4.2.1. Construct validity procedures of the survey

All boxplots and z-scores were examined. I determined the outliers and deleted 9 outliers for Teaching Method (TM) items; five had extreme z-scores and four had at least five missing values among 20 items (see Table 4.8, 4.9 and 4.10 for the statistical parameters).

Table 4. 8. Mean and Standard Deviations for Teaching Method Items (*N* = 294)

Variable	*M*	*SD*
TM1	3.22	1.13
TM2	3.30	1.13
TM3	4.05	1.08
TM4	2.85	1.19
TM5	4.39	.93
TM6	4.38	.81
TM7	3.56	1.23
TM8	3.15	1.23
TM9	4.48	.72
TM10	4.19	.85
TM11	4.29	.67
TM12	2.20	1.10
TM13	4.25	.82
TM14	4.33	.80
TM15	4.39	.68
TM16	4.59	.57
TM17	4.62	.56
TM18	4.41	.64
TM19	4.30	.77
TM20	4.30	.73

In addition, univariate normality skewness and kurtosis values as well as Kolmogorov-Smirnov and Shapiro-Wilk significance values were calculated.

The results showed that all of the items were within the limit (-3, +3) for Kurtosis value except item 5, item 6 and item 9. Skewness and Kurtosis values are totally satisfied. The values imply the normality for all but item 5, item 6 and item 9.

Table 4. 9. Skewness and Kurtosis Values for Teaching Method Items

Items	*Skewness*	*Kurtosis*
TM1	-.27	-1.04
TM2	-.27	-1.01
TM3	-1.07	.30
TM4	.29	-.94
TM5	-1.82	**3.13**
TM6	-1.60	**3.10**
TM7	-.50	-.89
TM8	-.02	-1.11
TM9	-1.57	**3.01**
TM10	-.96	.39
TM11	-.69	.56
TM12	.79	-.33
TM13	-1.39	2.54
TM14	-1.40	2.36
TM15	-.94	.78
TM16	-1.08	-.83
TM17	-1.40	2.23
TM18	-.87	.75
TM19	-.93	.43
TM20	-.74	-.06

For other normality checks, all the values for Kolmogorov-Smirnov, Shapiro-Wilk were also explored. All of them were smaller than $p = .00 < .05$. So, they are significant, which means that normality is not satisfied.

Table 4. 10. Kolmogorov-Smirnov & Shapiro-Wilk Significance Values for Teaching Method Items

Items	*Kolmogorov-Smirnov*	*Shapiro-Wilk*
TM1	.27	.87
TM2	.26	.88
TM3	.25	.80
TM4	.24	.90
TM5	.34	.68
TM6	.31	.72
TM7	.25	.87
TM8	.19	.90
TM9	.35	.70
TM10	.25	.79
TM11	.27	.77
TM12	.30	.83
TM13	.27	.75
TM14	.28	.74
TM15	.30	.75
TM16	.39	.67
TM17	.40	.64
TM18	.31	.74
TM19	.28	.78
TM20	.28	.78

Finally, histogram, and normal Q-Q plots were checked to ensure univariate normality. Histogram for Item 1, Item 2, Item 4, Item7, Item 8, Item 12 depicted to be normal. Yet remaining items such as Item 3, Item 5, Item 6, Item9, Item 10, Item 11, Item 13, Item 14, Item 15, Item 16, Item 17, Item 18, Item 19, Item 20 depicted negatively skewed distribution whereas Item 12 positively skewed distribution. Similarly, the Q-Q plot for Item 1, Item 12 illustrated linearity whereas Item 3, Item 17, for instance, illustrated the non-linearity in terms of univariate normality. Boxplot figures also depicted the outliers that still need to be dealt with (see Figure 4.7 – 4.14).

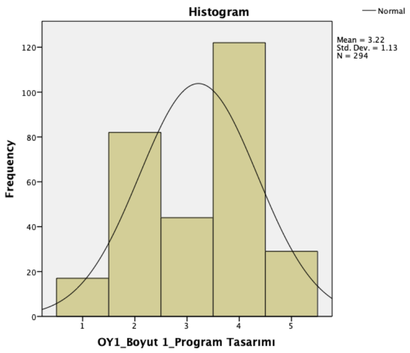


Figure 4. 7. Histogram for Item 1

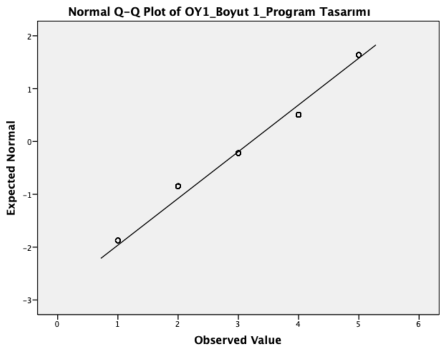


Figure 4. 8. Q-Q Plot for Item 1

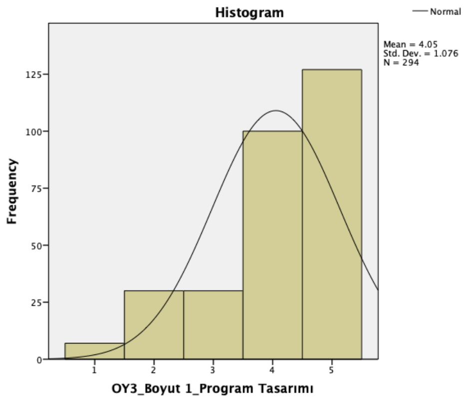


Figure 4. 9. Histogram for Item 3

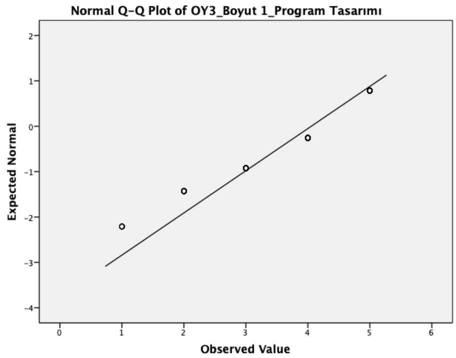


Figure 4. 10. Q-Q Plot for Item 3

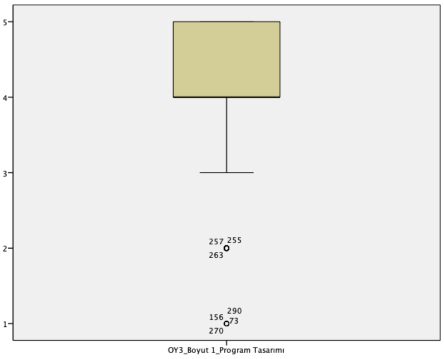


Figure 4. 11. Boxplot for Item 3

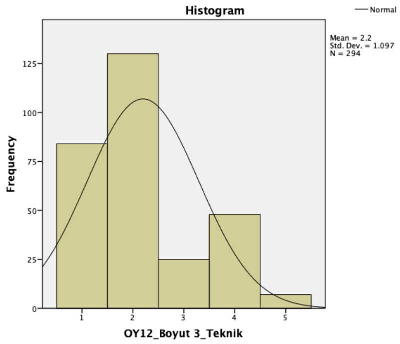
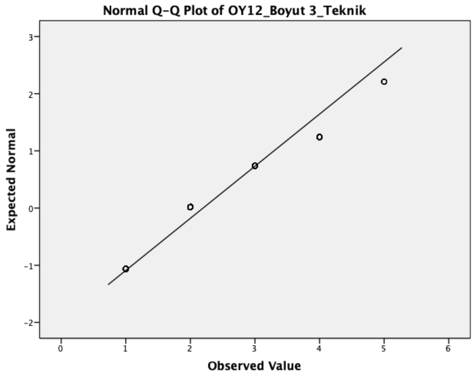


Figure 4. 12. Histogram for Item 12



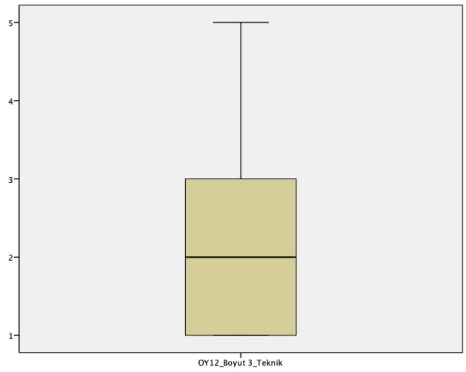
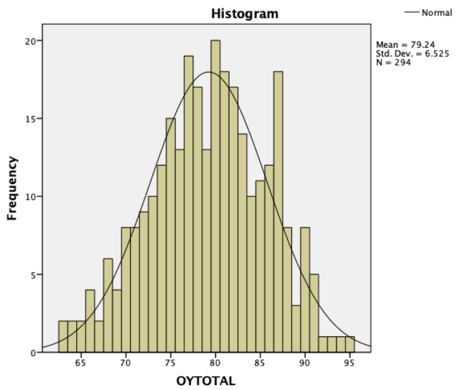


Figure 4. 13. Q-Q Plot for Item 12

Figure 4. 14. Box Plot for Item 12

Furthermore, when the overall data examined in relation with the univariate normality, histogram for all depicted to be normal. Q-Q plot for total items satisfied with the linearity in terms of univariate normality. Boxplot figure also showed no outliers (see Figure 4.15 – 4.17).



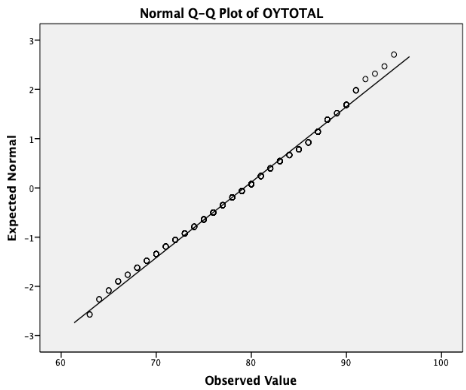


Figure 4. 15. Histogram for Total

Figure 4. 16. Q-Q Plot for Total

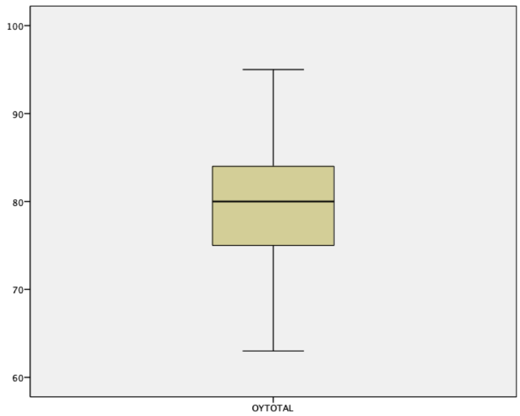


Figure 4. 17. Boxplot for Total

Consequently, the first theme of TM scale of TMMESP-Q was theoretically named as “Curriculum Design” including six items (item 1, item 2, item 3, item 4, item 7, item 8). Items related to measuring teachers’ general in-class applications/instructional choices were considered under the “General Instruction” dimension. The second theme was constructed to include four items (item 6, item 9, item 11, item 13). On the other hand, the third theme was named as “Instructional Technique” comprising five items (item 5, item 10, item 12, item 14, item 16, item 17). In this theme, items related to teachers’ amount of tendency relevant to use instructional techniques were taken into consideration. Lastly, theme 4 was called “Constructivism” including theoretically four items (item 15, item 18, item 19, item 20). Items related to teachers’ constructivist application skills are concerned in this dimension (see Table 4.11 for descriptives).

Table 4. 11. Mean and Standard Deviations for Measurement-Evaluation Strategy Items (*N* = 327)

Variable	*M*	*SD*
MES1	3.47	.96
MES2	2.56	1.23
MES3	3.46	1.02
MES4	3.46	.97
MES5	3.71	1.10
MES6	3.98	.85
MES7	2.80	1.23
MES8	3.66	1.08
MES9	3.94	.96
MES10	4.27	.66
MES11	2.67	1.17
MES12	3.77	.81
MES13	3.12	1.12
MES14	4.03	1.06
MES15	2.08	1.26

In addition, univariate normality skewness and kurtosis values as well as Kolmogorov-Smirnov and Shapiro-Wilk significance values were calculated. The results showed that Skewness and Kurtosis values are totally satisfied that they were between the limits of (-3, +3). The values show the normality (see Table 4.12).

Table 4. 12. Skewness and Kurtosis Values for Measurement-Evaluation Strategy Items

Items	*Skewness*	*Kurtosis*
MES1	-.20	-.10
MES2	.22	-1.09
MES3	-.54	-.10
MES4	-.25	-.27
MES5	-.71	-.17
MES6	-.53	-.18
MES7	.17	-.90
MES8	-.41	-.59
MES9	-.84	.42
MES10	-.55	.11
MES11	.29	-.70
MES12	-.45	.22
MES13	.00	-.74
MES14	-1.19	.92
MES15	.91	-.38

For other normality checks, all the values for Kolmogorov-Smirnov, Shapiro-Wilk were also explored. All of them are smaller than p =.00 < .05. So, they are significant which means that normality is not satisfied (see Table 4.13 for normality check).

Table 4. 13. Kolmogorov-Smirnov and Shapiro-Wilk Significance Values for Measurement-Evaluation Strategy Items

Items	*Kolmogorov-Smirnov*	*Shapiro-Wilk*
MES1	.23	.89
MES 2	.18	.89
MES 3	.24	.89
MES 4	.20	.90
MES 5	.25	.87
MES 6	.25	.85
MES 7	.17	.91
MES 8	.20	.89
MES 9	.26	.85
MES 10	.28	.78
MES 11	.18	.91
MES 12	.28	.86
MES 13	.17	.92
MES 14	.27	.80
MES 15	.27	.80

Finally, histogram, and normal Q-Q plots were checked to ensure univariate normality. Histogram for MES1, MES2, MES3, MES4, MES5, MES6, MES7, MES8, MES10, MES11, MES12, MES13 depicted to be normal. Yet remaining items such as MES9 and MES14 depicted negatively skewed distribution whereas MES 15 depicted positively skewed distribution. Similarly, the Q-Q plot for MES1 illustrated linearity whereas MES10, MES14, for instance, illustrated the non-linearity in terms of univariate normality. Boxplot figures also depicted the outliers that need to be dealt with (see Figure 4.18 – 4.25).

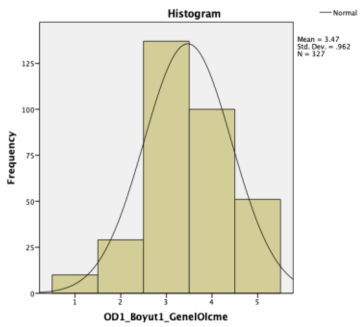


Figure 4. 18. Histogram for MES 1

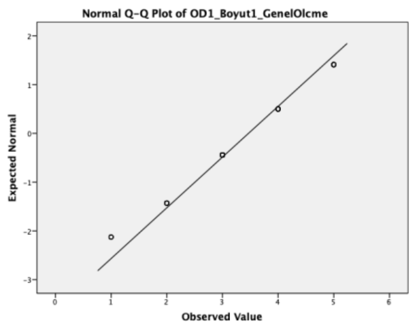


Figure 4. 19. Q-Q Plot for MES 1

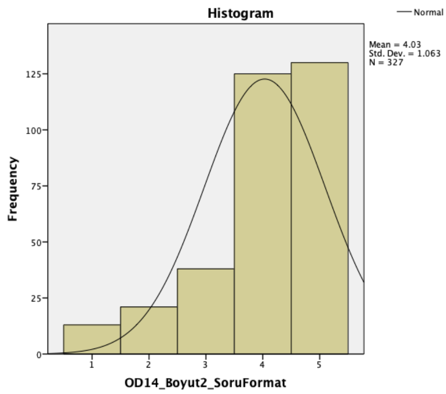


Figure 4. 20. Histogram for MES 14

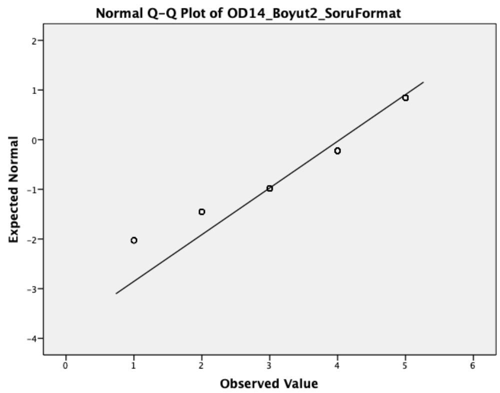


Figure 4. 21. Q-Q Plot for MES 14

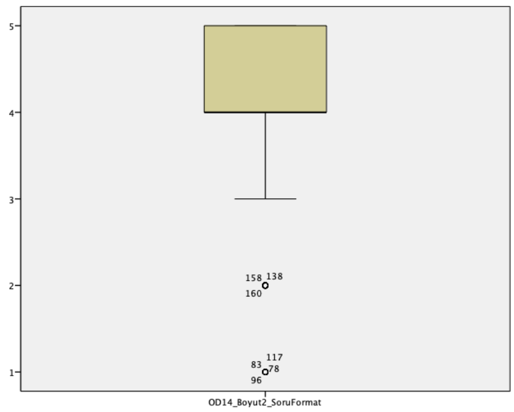


Figure 4. 22. Boxplot for MES 14

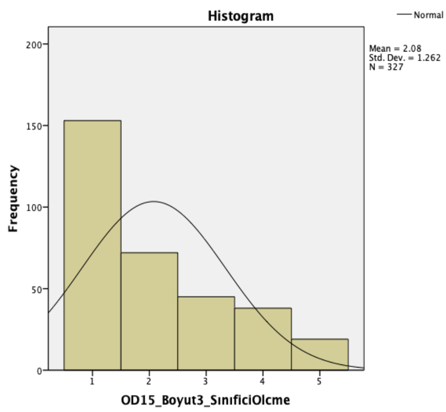
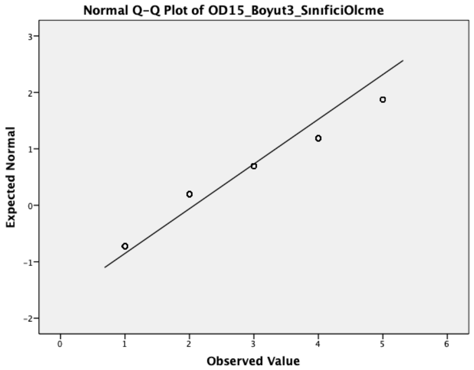


Figure 4. 23. Histogram for MES 15



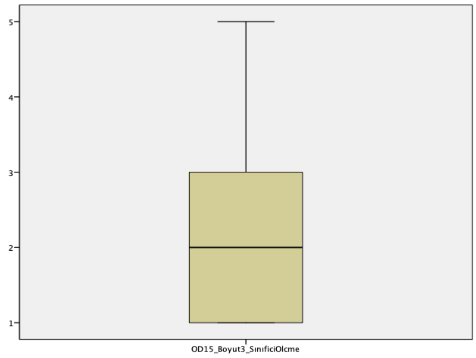
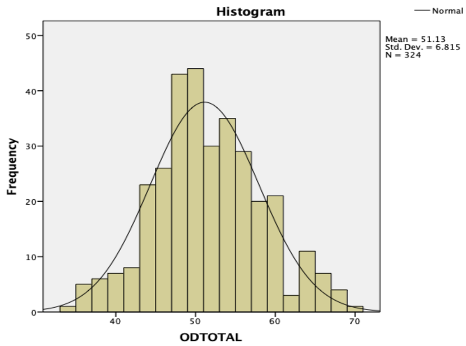


Figure 4. 24. Q-Q Plot for MES 15 **Figure 4. 25.** Boxplot for MES 15

Furthermore, when the overall data examined in relation with the univariate normality, histogram for all depicted to be normal. Q-Q plot for total items satisfied with the linearity in terms of univariate normality. Boxplot figure also showed no outliers (see Figure 4.26 – 4.28).



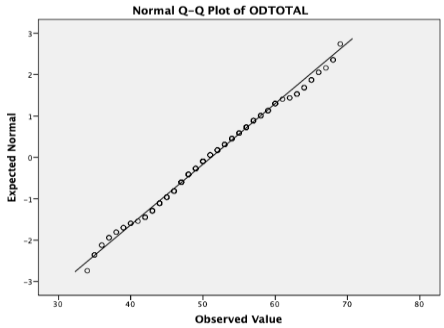


Figure 4. 26. Histogram for Total **Figure 4. 27.** Q-Q plot for Total

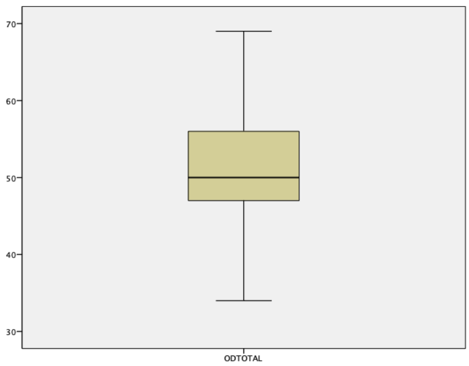


Figure 4. 28. Box Plot for Total

The first theme of MES scale of TMMESP-Q was theoretically named as "General Measurement-Evaluation Process" including five items (item 1, item 2, item 4, item 11, item 12). Items related to measuring teachers' views on question formats were considered under the "Views with Question Formats" dimension. The second theme was theoretically constructed to include six items (item 3, item 5, item 7, item 8, item 9, item 14). On the other hand, the third theme was named as "In-class Teaching Measurement-Evaluation Techniques" theoretically comprising four items (item 6, item 10, item 13, item 15).

However, in order to analyze the TMMES-Q descriptively and inferentially, for the aim of this current thesis, I used a theoretical model. I suggest applying further examination of the dimensions in the future goals.

4.2.2. Results of main study of survey phase

The findings of the descriptive and inferential statistics from TMMESP-Q presented in this section.

a. Findings of the descriptive statistics from TMMESP-Q

In this part, item-level statistics for the main data of 35 questions in TMMESP-Q was given in the Table Q.1 (TM Scale) and Table Q.2 (MES Scale) in Appendix Q.

When the TMMESP-Q was analyzed on the TM part (see Figures 4.29-4.32), the findings revealed that after the curriculum renovation the mathematics teachers reflected their preference frequently on the items about curriculum design as follows: "I think there is no change in the purpose of the curriculum" as agree (41.60%), "I think there is no change in the philosophy of the curriculum" as agree (39.50%), "I have noticed the decrease in the number of learning outcomes" as strongly agree (42.20%), "I think the content of the subject has been enriched" as disagree (37.50%), "I think there is no change in the content of the Mathematics resource books of the Ministry of National Education" as agree (34%), and “I prefer using the teacher's handbook” as agree (26.70%). It can be reflected from overall preferences that within the dimension of curriculum design, the teachers agree that there is no change in the purpose of the curriculum; they agree that there is no change in the philosophy of the curriculum; they agree that they have noticed the decrease in the number of learning outcomes; on the contrary, they disagree that the content of the subject has been enriched; they agree that there is no change in the content of the Mathematics resource books of the Ministry of National Education; they agree that they prefer using the teacher's handbook.

Moreover, the teachers reflected their preference frequently on the items about general instruction as follows: “I prefer doing activities that provide opportunities for student creativity” as strongly agree (50.30%), “I change my in-class teaching method to make my students active” as strongly agree (57.30%), “I design lessons that enable my students to learn by exploring mathematics effectively” as agree (51.20%), “Before the lesson, I check the students’ readiness” as agree (46.20%). It can be reflected from overall preferences that within the dimension of general instruction, the teachers agree that they prefer doing activities that provide opportunities for student creativity; they change their in-class teaching method to make their students active; they design lessons that enable their students to learn by exploring mathematics effectively; Before the lesson, they check the students’ readiness.

Furthermore, they reflected their preference frequently on the items about instructional technique as follows: “Using concrete materials (e.g. mathematical objects) during classroom teaching helps me a lot” as strongly agree (57.30%), “I prefer using group teaching methods (e.g., cooperative learning, think-pair-share etc.)” as agree (41.30%), “I only use direct instruction” as disagree (42.70%), “I try to use educational technologies when teaching in-class” as strongly agree (48%), “I feel the need to use different questioning techniques (e.g., Why?, How? etc.)” as strongly agree (61.60%), and “I give examples from daily life while teaching a topic" as strongly agree (66%). It can be derived from overall preferences that within the dimension of instructional technique, the teachers agree that using concrete materials (e.g., mathematical objects) during classroom teaching helps them a lot; they prefer using group teaching methods (e.g., cooperative learning, think-pair-share etc.); on the other hand, they disagree that they only use direct instruction; they agree that they try to use educational technologies when teaching in-class; they feel the need to use different questioning techniques (e.g., Why?, How? etc.); and they give examples from daily life while teaching a topic.

Finally, they reflected their preference frequently on the items about constructivism as follows: "I prefer using the constructivist approach techniques when teaching (e.g., research, interpret and analyze information, improve the thinking process etc.)" as strongly agree (49.70%), "I prefer designing a learning environment that makes students think about the topic they work on" as strongly agree (48%), "I use instructional techniques that require students to take responsibility for their learning (e.g., demonstration, question-answer, brainstorming, discussion)" as strongly agree (46.80%), and "I encourage students to do research" as strongly agree (43.6%). It can be inferred from overall preferences that within the dimension of constructivism, the teachers agree that they prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.); they prefer designing a learning environment that makes students think about the topic they work on; they use instructional techniques that require students to take responsibility for their learning (e.g. demonstration, question-answer, brainstorming, discussion); and they encourage students to do research. The statistical parameters for TMMESP-Q were illustrated in Table 4.18 (see. Appendix Q).

The descriptives for TMMESPQ related to highest adverbs of frequency are shown in Table 4.14 and 4.15.

Table 4. 14. Item-Level Descriptives for 20 TM Items from TMMESP-Q

Scale Items for TM[a]	*M*	*SD*	*f*[b]	%
1. I think there is no change in the purpose of the curriculum.	3.25	1.13	180	52.4
2. I think there is no change in the philosophy of the curriculum.	3.29	1.13	180	52.3
3. I have noticed the decrease in the number of learning outcomes.	4.06	1.07	260	75.6
4. I think the content of the subject has been enriched.	2.84	1.18	164	47.7
5. Using concrete materials (e.g. mathematical objects) during classroom teaching helps me a lot.	4.39	.91	298	86.7
6. I prefer doing activities that provide opportunities for student creativity.	4.36	.82	297	86.3
7. I think there is no change in the content of the Mathematics resource books of the Ministry of National Education.	3.59	1.22	210	61
8. I prefer using the teacher's handbook.	3.17	1.22	148	43
9. I change my in-class teaching method to make my students active.	4.48	.71	313	91
10. I prefer using group teaching methods (e.g. cooperative learning, think-pair-share etc.).	4.20	.85	283	82.3

Table 4. 14. (continued)

Scale Items for TM[a]	*M*	*SD*	*f*[b]	%
11. I design lessons that enable my students to learn by exploring mathematics effectively.	4.28	.66	304	88.4
12. I only use direct instruction.	2.10	1.10	243	70.6
13. Before the lesson, I check the students' readiness.	4.27	.80	301	87.5
14. I try to use educational technologies when teaching in-class.	4.33	.79	301	87.5
15. I prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.).	4.41	.68	311	90.4
16. I feel the need to use different questioning techniques (e.g. Why? How? etc.)	4.60	.56	326	94.7
17. I give examples from daily life while teaching a topic.	4.64	.56	328	95.4
18. I prefer designing a learning environment that makes students think about the topic they work on.	4.41	.65	316	91.9
19. I use instructional techniques that require students to take responsibility for their learning (e.g. demonstration, question-answer, brainstorming, discussion).	4.32	.76	296	86
20. I encourage students to do research.	4.29	.73	298	86.6

[a]The 5-point response options were as follows: strongly disagree, disagree, undecided, agree, strongly agree.
[b]Highest adverbs of frequency were calculated to show teachers' tendency of TM preferences.

Moreover, item level descriptives for MES items from TMMESP-Q are shown in Table 4.15.

Table 4. 15. Item-Level Descriptives for 15 MES Items from TMMESP-Q

Scale Items for MES[a]	*M*	*SD*	*f*[b]	%
1. I make changes in the measurement and evaluation process compared to the previous implementations.	3.46	.95	158	45.9
2. I administer examinations based on downloaded online sources (e.g. forums, websites etc.).	2.55	1.22	172	50
3. I use measurement tools that include multiple-choice items.	3.44	1.03	181	52.6
4. I use formative assessment to measure course learning outcomes.	3.45	.97	167	48.6
5. I prepare examinations that include a mixture of multiple-choice and short-answer items.	3.70	1.09	217	63.1
6. I ask problem-solving items related to real life problems.	3.96	.85	250	72.7
7. I use portfolio that will enable the students to show their performances at the end of the term.	2.82	1.21	145	42.2
8. I apply quizzes.	3.67	1.07	198	57.6
9. I use open-ended items in my in-class examinations.	3.97	.93	253	73.5
10. I use the question-answer technique in my teaching.	4.27	.68	304	88.4
11. I give students choice to choose which item types they want to be included in their examinations.	2.70	1.18	159	46.2
12. I prefer item types that require students to use procedural skills in the examinations.	3.76	.82	228	66.3
13. I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought).	3.14	1.12	130	37.8
14. I use open-ended and multiple-choice items together in my in-class examinations.	4.04	1.05	268	77.9
15. I determine the number of in-class examinations to be administered together with the students.	2.09	1.27	236	68.6

[a]The 5-point response options were as follows: never, seldom, sometimes, often, always
[b]Highest two adverbs of frequency were calculated to show teachers' tendency of MES preferences.

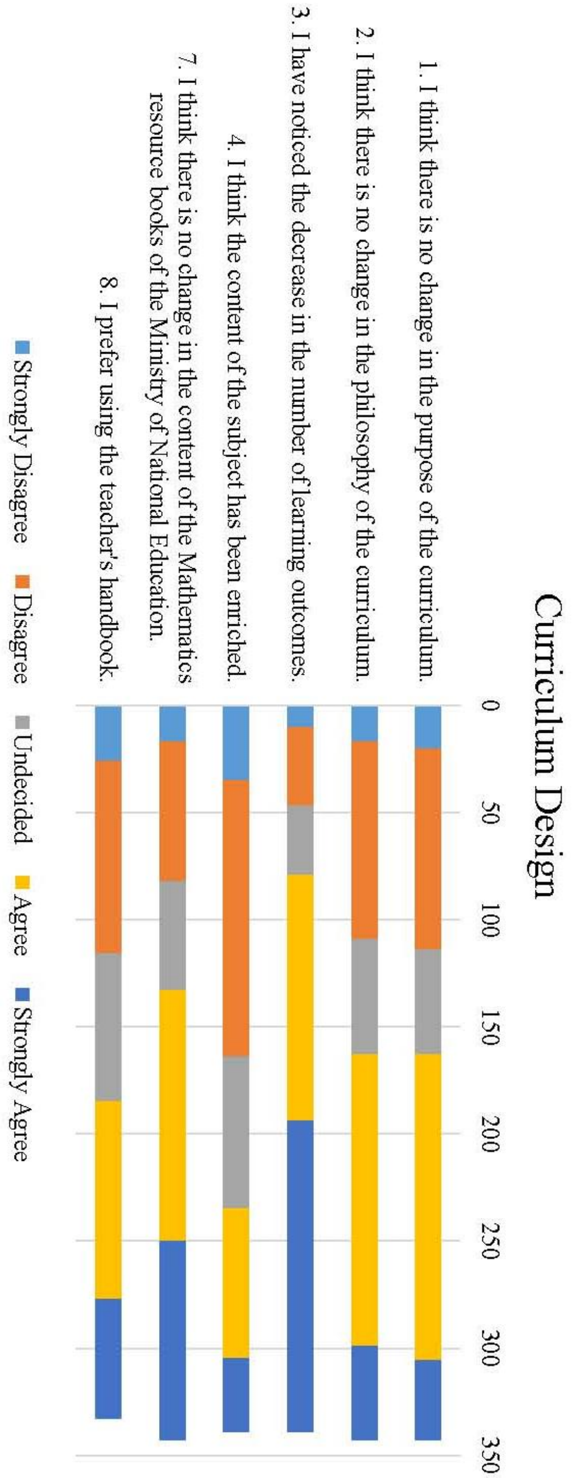


Figure 4. 29. Teachers' Preferences about Curriculum Design Dimension related to TMMESP-Q

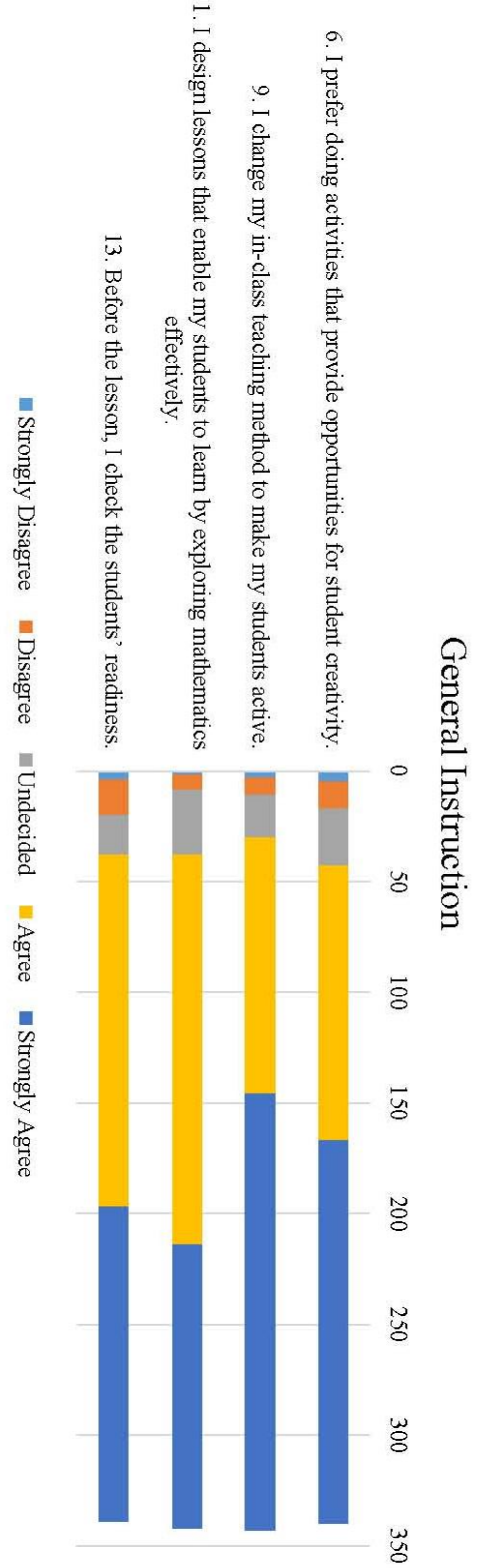


Figure 4. 30. Teachers' Preferences about General Instruction Dimension related to TMMESP-Q

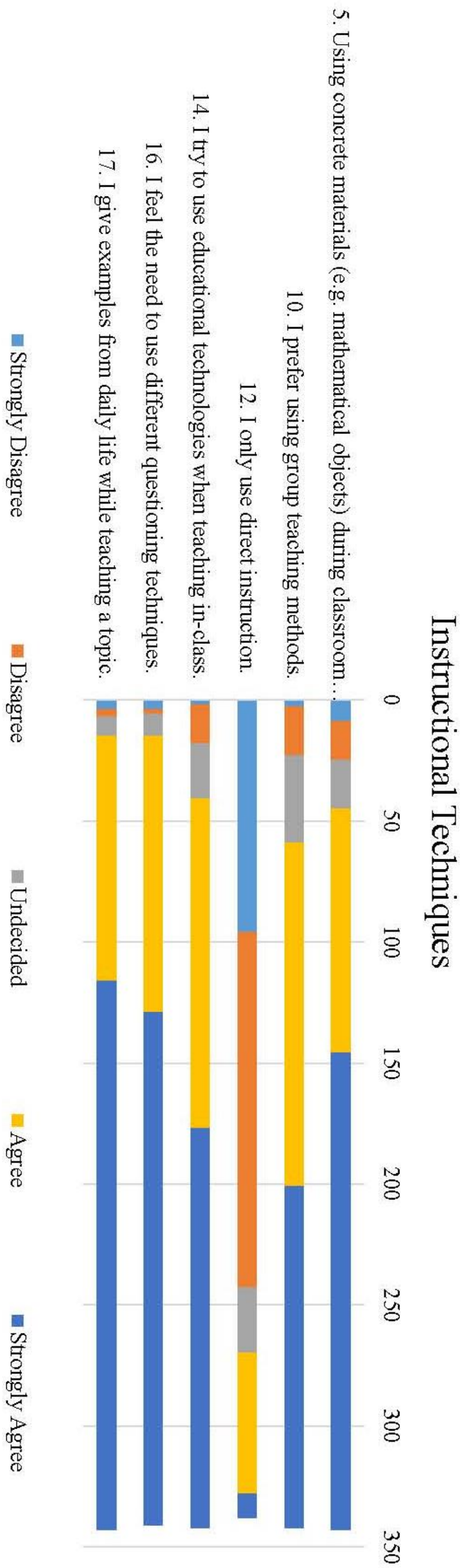


Figure 4. 31. Teachers' Preferences about General Instruction Dimension related to TMMESP-Q

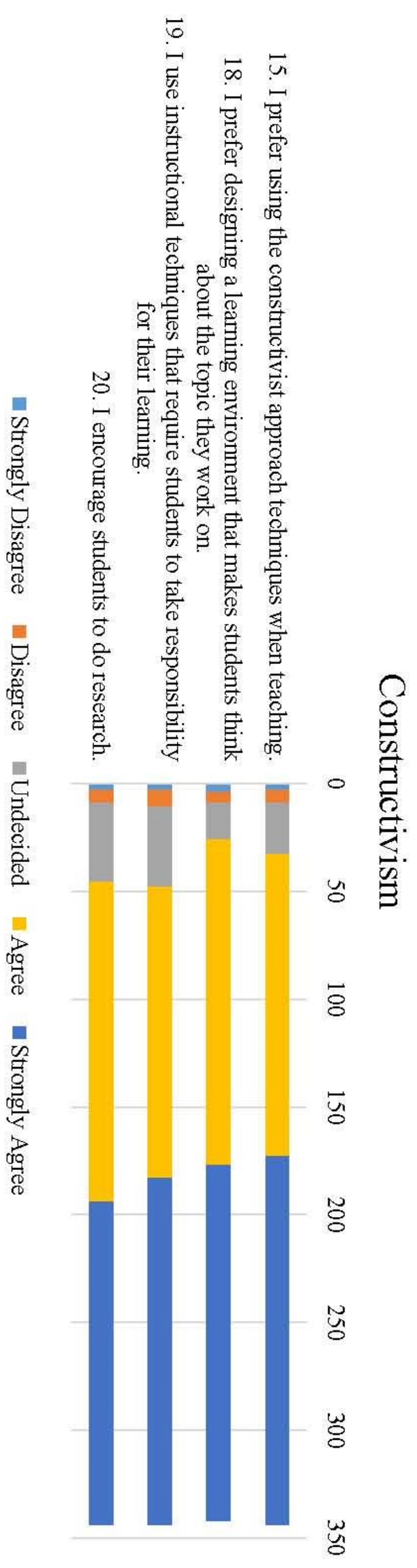


Figure 4. 32. Teachers' Preferences about Constructivism related to TMMESP-Q

Finally, most of them reflected their preferences on the items of in-class teaching assessment "I ask problem-solving items related to real life problems" as often (44.80%), "I use the question-answer technique in my teaching" as often (50.30%), "I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought)" as sometimes (32.60%) but often (24.40%) and always (13.40%), and "I determine the number of in-class examinations to be administered together with the students" as never (46.50%). It can be reflected from overall preferences that within the dimension of in-class teaching assessment, 52.6% of the teachers prefer to use measurement tools that include multiple-choice items, 63.10% of them prefer to prepare examinations that include a mixture of multiple-choice and short-answer items, 53.80% of them [sometimes and seldom] prefer to use portfolio that will enable the students to show their performances at the end of the term, 57.60% of them prefer to apply quizzes, and 73.50% of them prefer to use open-ended items in my in-class examinations.

When the TMMESP-Q was analyzed on the MES part (see. Figures 4.33-4.35), the findings revealed that after the curriculum renovation the mathematics teachers reflected their preference frequently on the items as follows: "I make changes in the measurement and evaluation process compared to the previous implementations" as sometimes (41.90%), "I administer examinations based on downloaded online sources (e.g. forums, websites etc.)" as never (26.20%) and seldom (23.80%), "I use formative assessment to measure course learning outcomes" as sometimes (36%) and often (34.60%), "I give students choice to choose which item types they want to be included in their examinations" as seldom (28.80%) and sometimes (28.80%), "I prefer item types that require students to use procedural skills in the examinations" as often (49.40%). It can be reflected from overall preferences that within the dimension of general measurement-evaluation process, the teachers' tendency is sometimes to make changes in the measurement and evaluation process compared to the previous implementations, never or seldomly administer examinations based on downloaded online sources (e.g. forums, websites etc.), sometimes to use formative assessment to measure course learning outcomes, often to prefer item types that require students to use procedural skills in the examinations.

Moreover, most of the teachers reflected their preferences on the items of question/assessment formats "I use measurement tools that include multiple-choice items" as often (40.10%), "I prepare examinations that include a mixture of multiple-choice and short-answer items" as often (37.80%), "I use portfolio that will enable the students to show their performances at the end of the term" as sometimes (27.90%) but seldom (25.90%), "I apply quizzes" as often (31.1%) and "I use open-ended items in my in-class examinations" as often (42.40%). It can be reflected from overall preferences that within the dimension of assessment formats, 52.6% of the teachers prefer to use measurement tools that include multiple-choice items, 63.10% of them prefer to prepare examinations that include a mixture of multiple-choice and short-answer items, 53.80% of them seldom or sometimes prefer to use portfolio that will enable the students to show their performances at the end of the term, 57.60% of them prefer to apply quizzes, and 73.50% of them prefer to use open-ended items in my in-class examinations.

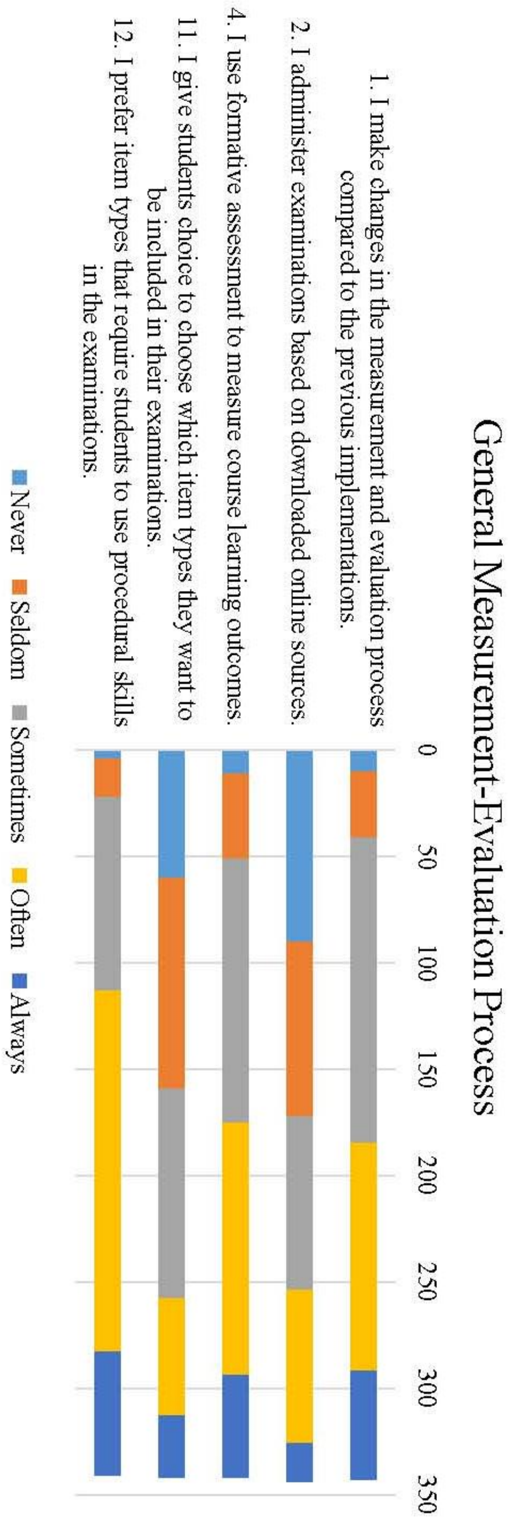


Figure 4. 33. Teachers' Preferences about General Measurement-Evaluation Process related to TMMESP-Q

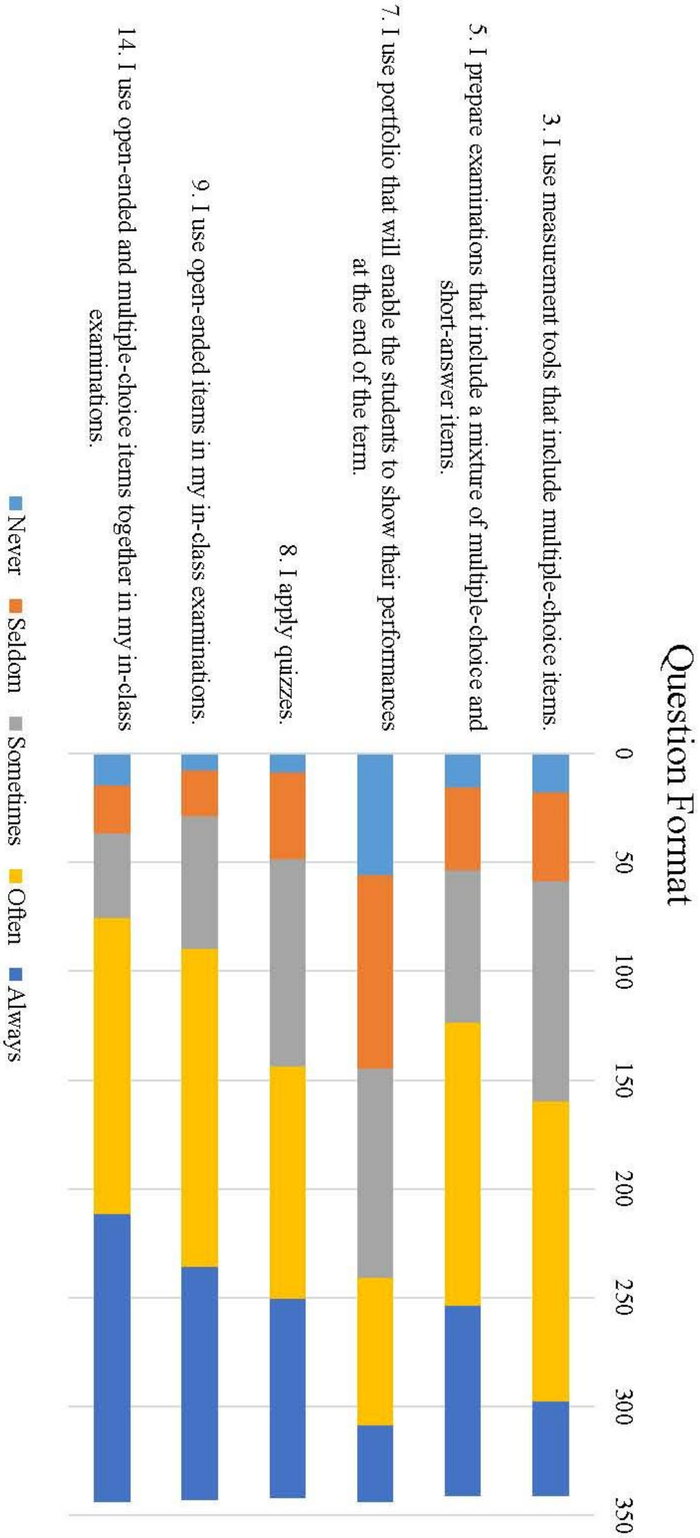


Figure 4. 34. Teachers' Preferences about Question Format related to TMMESP-Q

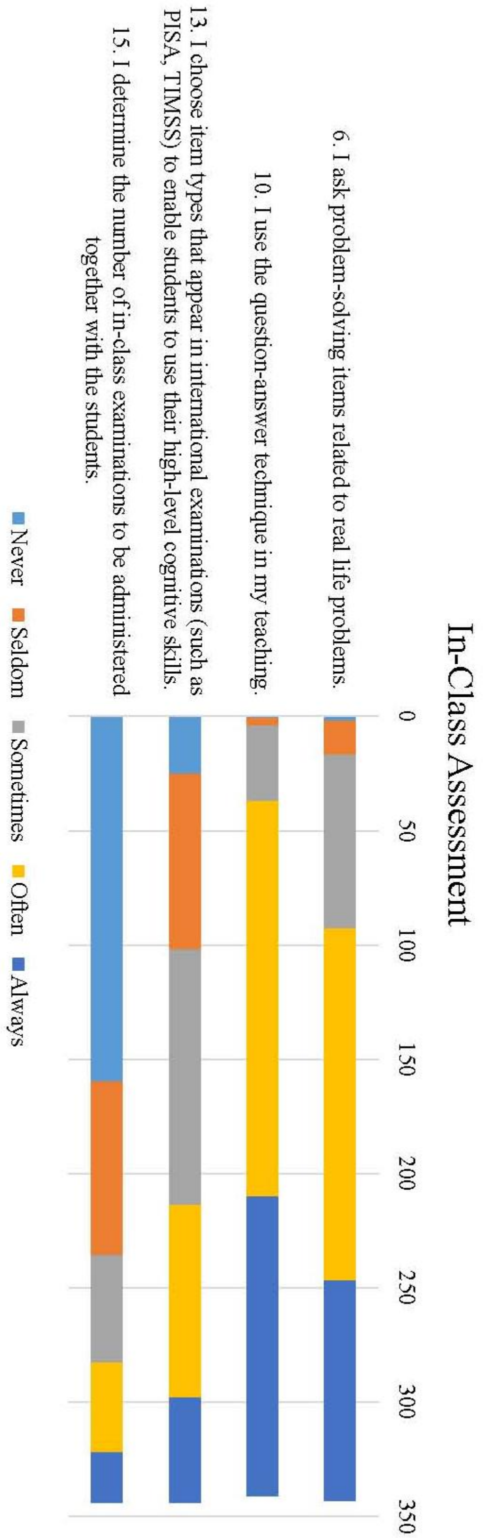


Figure 4. 35. Teachers' Preferences about In-class Assessment related to TMMESP-Q

b. Results of TMMESP questionnaire

Mathematics teachers teaching method preferences and measurement-evaluation strategy preferences were measured and evaluated by TMMESP questionnaire. The findings related to TMMESP-Q were also yielded to statistical and inferential aspects. The results were also evaluated with the inferential statistics because related literature has revealed that some variables such as teachers' gender, school type, seniority year (i.e., professional year), and educational level (i.e., graduation program) were related with the teachers' perceptions or preferences (see Table 4.16 for descriptive statistics).

Table 4. 16. Descriptive Statistics of Teachers Gender, School Type, Seniority Year, Educational Level

Characteristics	TM_{Total}[a]		MES_{Total}[b]		TOTAL[c]		
Gender	*f*	*M(SD)*	*f*	*M(SD)*	*f*	*%*	*Cumulative %*
Female	203	80.04 (6.27)	224	51.08 (6.70)	246	70.3	70.3
Male	93	77.47 (6.71)	99	51.27 (7.13)	103	29.4	99.7
School type							
Public	176	78.63 (6.78)	206	51.42 (7.04)	217	62	62
Private	109	80.14 (5.86)	117	50.77 (6.26)	132	37.7	99.7
Seniority Year							
Year < 15		78.11 (8.19)		50.55 (7.93)			
Year >= 15		80.27 (6.84)		51.15 (6.68)			
Educational Level							
Mathematics education	203	78.71 (6.65)	220	51.49 (7.15)	233	66.6	67.0
Arts and sciences	93	80.47 (5.96)	102	50.51 (5.90)	115	32.9	100.0

The hypotheses were as follows:

H_0: $\mu_A - \mu_B = 0$ or $\mu_A = \mu_B$ meaning that there is no significant difference between teachers' teaching method and measurement-evaluation strategy preferences in total or subscales in terms of their gender, seniority year (i.e., professional year), educational level, school type.

H_1: $\mu_A - \mu_B \neq 0$ or $\mu_A \neq \mu_B$ meaning that there is a significant difference between teachers' teaching method and measurement-evaluation strategy preferences in total or subscales in terms of their gender, seniority year (i.e., professional year), educational level, school type.

First of all, assumptions of t-test statistics have been checked. Random sampling, independent observation, normality assumptions were checked and found to be non violated. By following homogeneity assumption through Levene's Test result, significance value is .98> .05. It depicts that we reject H_o and homogeneity of variance assumption has not been violated. Then equal variances assumed line was followed to read t value.

An independent-samples t-test was used to compare the teachers' teaching method preferences in case of their gender (Table 4.17). The results indicated that there was a statistically significant difference in teaching method preferences total scores between female teachers (M = 79.36, SD = 7.98) and male teachers (M = 77.07, SD =7.71); t (297) = 2.32, p < .05, r^2 = .02, two-tailed test. Female teachers' adoption after curriculum change was found to be higher than male teachers. By reading mean difference, Cohen's d = (Mean Difference/SD) = (2.82)/ $\sqrt{15.69}$= 0.71. According to Cohen's Standard's (1988) such as 0 < d < .20 = small effect; .20 < d < .80 = medium effect and d > .80 = large effect, it is a medium effect due to the fact that 0.71 > .20. In addition, Eta squared $\eta^2 = t^2/(t^2+df) = (2.32)^2/ (2.32^2+297) = 0.02$. According to Cohen's Standard's (1988) such as .01 = small effect, .06 = moderate effect and .15 = large effect; it is a small to moderate effect due to the fact that .02 > .01. Also, it can be interpreted that 2% of the variance in teaching method preferences is explained by gender as male or female. Interestingly, the results indicated we fail to reject H_o; there was not a statistically significant difference in measurement-evaluation strategy preferences total scores between male teachers (*M* = 51.34, *SD* = 7.70) and female teachers (*M* = 50.87, *SD* = 6.90); *t* (322) = -.55, $p > .05$.

Table 4. 17. Teaching Method Preferences of Teachers by Gender

	Female		Male					
	M	*SD*	*M*	*SD*	*t*(297)	*p*	Cohen's *d*	Eta squared r^2
Teaching method preferences total scores	79.36	7.98	77.07	7.71	2.32	.02	0.71	.02

Another independent-samples t-test was used to compare the teachers' teaching method preferences in case of their seniority years (Table 4.18). The results indicated that there was a statistically significant difference in teaching method preferences total scores between teachers whose seniority year is smaller than 15 (*M* = 78.11, *SD* = 8.19) and those whose seniority year is bigger and equal to 15 (*M* = 80.27, *SD* = 6.84); *t* (296) = 2.05, $p < .05$, $r^2 = .01$, two-tailed test. The teachers who has experienced in the profession more or equal to 15 years were found to have higher teaching method preferences total score after curriculum change than those who has experienced in the profession smaller than 15 years. It can be inferred that the teachers who has at least 15 years in the profession tended to prefer different constructivist teaching methods after curriculum change and may become change agents in terms of teaching method preferences. By reading mean difference, Cohen's d = (Mean Difference/SD) = (2.16)/ √15,03= 0.56. According to Cohen's Standard's (1988) such as 0 < d < .20 = small effect; .20 < d < .80 = medium effect and d > .80 = large effect, it is a medium effect due to the fact that 0.56 > .20. In addition, Eta squared $\eta^2 = t^2/(t^2+df) = (2.05)^2/ (2.05^2+296) = 0.01$. According to Cohen's Standard's (1988) such as .01 = small effect, .06 = moderate effect and .15 = large effect; it is a small effect due to the fact that .01 = .01. Also, it can be interpreted that 1% of the variance in teaching method preferences is explained by seniority year. Interestingly, the results indicated we fail to reject H_o; there was not a statistically significant difference in measurement-evaluation strategy preferences total scores between the teachers whose seniority year is bigger and equal to 15 (*M* = 50.55, *SD* = 7.93) and those whose seniority year is smaller than 15 (*M* = 51.15, *SD* =6.86); *t* (321) = -.66, *p* > .05.

Table 4. 18. Teaching Method Preferences of Teachers by Seniority Year

	Year <15		Year >=15					
	M	*SD*	*M*	*SD*	*t*(296)	*p*	Cohen's *d*	Eta squared r^2
Teaching method preferences total scores	78.11	8.19	80.27	6.84	2.05	.04	0.56	.01

On the other hand, there was no statistically significant difference in teaching method preferences total scores in terms of educational level and school type. In other words, there was no statistically significant difference in teaching method preferences total scores between the teachers who has been working in public schools (*M* = 78.34, *SD* = 7.69) and those who has been working in private schools (*M* = 79.14, *SD* = 8.35); *t* (298) = -.85, *p* > .05. There was no was a statistically significant difference in teaching method preferences total scores between the teachers who graduated from faculty of education (*M* = 78.17, *SD* = 7.93) and those who graduated from faculty of arts and sciences (*M* = 79.77, *SD* = 7.88); *t* (297) = -1.619, *p* > .05. Similarly, there was no statistically significant difference in measurement-evaluation strategy preferences total scores between the teachers who has been working in public schools (*M* = 51.32, *SD* = 7.43) and those who has been working in private schools (*M* = 50.48, *SD* = 6.60); *t* (323) = 1.204, *p* > .05. There was no was a statistically significant difference in measurement-evaluation strategy preferences total scores between the teachers who graduated from faculty of education (*M* = 51.27, *SD* = 7.35) and those who graduated from faculty of arts and sciences (*M* = 50.60, *SD* = 6.58); *t* (221) = .82, *p* > .05.

Another independent-samples t-test was used to compare the subscales scores in case of school type (Table 4.19). The results indicated that there was a statistically significant difference in curriculum design scores between the teachers working in public schools (*M* = 20.65, *SD* = 3.65) and the teachers working in private schools (*M* = 19.29, *SD* = 3.40); t (320) = 3.40, $p < .05$, r^2 = .01, two-tailed test. By reading mean difference, Cohen's d = (Mean Difference/SD) = (1.36)/ √6,78= 0.52. It is a medium effect. Eta squared η^2= $t^2/(t^2+df)$ = $(3.40)^2/\ (3.40^2+320)$ = 0.03 meaning 3% of the variance in curriculum design scores is explained by school type. Also, there was a statistically significant difference in in-class assessment preferences scores between the teachers working in public schools (*M* = 13.22, *SD* = 2.66) and the teachers working in private schools (*M* = 13.87, *SD* = 2.00); t (337) = -2.30, $p <$ .05, r^2 = .02, two-tailed test. By reading mean difference, Cohen's d = (Mean Difference/SD) = (-.65)/ √4,66= 0.30. It is a medium effect. Eta squared η^2= $t^2/(t^2+df)$ = $(-2.39)^2/\ (-2.39^2+337)$ = 0.02 meaning 2% of the variance in in-class assessment preference scores is explained by school type.

Table 4. 19. Descriptive Statistics of Subscales of TMMESP-Q Scores by School Type

	Public		Private					
	M	*SD*	*M*	*SD*	*t*(320)	*p*	Cohen's *d*	Eta squared r^2
Curriculum Design	20.65	3.65	19.29	3.13	3.40	.001	0.52	.03
General Instruction	16.92	2.29	17.71	2.40	-2.99	.00	0.36	.03
Constructivism	16.97	2.57	17.83	2.49	-3.03	.00	0.38	.03
General Measurement -Evaluation	16.37	3.03	15.18	2.64	3.76	.00	0.50	.04
In-Class Assessment	13.22	2.66	13.87	2.00	-2.39	.003	0.30	.02

4.3. Results of Multimodal Phase: Neuroeducation

In the final phase of the current study, I aimed to examine the following research questions: “How do middle school students reflect their metacognitive skills (cognitive strategy and self-checking) and affective process (effort and worry) levels of their responses to different item types? Is there a significant difference between the amount of reflection of students’ metacognitive skill levels on their responses to multiple-choice and open-ended items? 3.1) Is there a significant difference between the amount of reflection of students’ cognitive strategy skill levels on their responses to multiple-choice and open-ended items?, 3.2) Is there a significant difference between the amount of reflection of students’ self-checking skill levels on their responses to multiple-choice and open-ended items?” and “4) What are students' reactions and responses to different types of questions with respect to the requirement (active use) of different cognitive strategies with the use of eye-tracker and biometric sensors including galvanic skin response (GSR) and heart rate (HR)? 4.1) Is there a significant difference between the amount of reflection of students’ affective process levels on their responses to multiple-choice and open-ended items?, 4.2) Is there a significant difference between the amount of reflection of students’ worry levels on their responses to multiple-choice and open-ended items?, 4.3) Is there a significant difference between the amount of reflection of students’ effort levels on their responses to multiple-choice and open-ended items? ,4.4) Do total time on task and gaze shifts have an impact on predicting rereading or not while responding to items?”

4.3.1. Descriptive and inferential results for students’ metacognitive skills and affective processes on their responses to multiple-choice and open-ended items

The descriptive findings in Table 4.20 showed that the frequencies of students’ cognitive strategy skill levels on their responses to multiple-choice items were slightly lower in number (*f* = 129, 26%) than those who were not able to reflect (*f* = 166, 34%). In addition, the frequencies of students' self-checking skill levels on their responses to multiple-choice items were 3 times more (*f* = 190, 29%) than those who were not able to reflect (*f* = 67, 10%). In other words, one third of the fifth grade students' ability to use their self-checking skills was slightly higher (*f* = 190, 29%) than those to use their cognitive strategy skills (*f* = 129, 26%) on their responses to multiple-choice.

Table 4. 20. Frequency Distribution in terms of Metacognitive Skills by Multiple-choice Items

Metacognitive skills	f_{YES}	%	f_{NO}	%
Cognitive strategy	129	26	166	34
Self-checking	190	29	67	10
TOTAL	319	55	233	44

In other words, it can be inferred that the students who did not use cognitive strategy skills may tend to use heuristic solution processes. If their confidence level was low toward the solution, they tended to use self-checking skills.

The descriptive findings in Table 4.21 showed that the frequencies of fifth grade students' cognitive strategy skill levels on their responses to open-ended items were slightly higher (*f* = 162, 33%) than those who were not able to reflect (*f* = 148, 30%). In addition, the frequencies of the students' self-checking skill levels on their responses to open-ended items were more than 3 times (*f* = 172, 26%) than those who were not able to reflect (*f* = 53, 8%).

Table 4. 21. Frequency Distribution in terms of Metacognitive Skills by Open-ended Items

Metacognitive subskills	f_{YES}	%	f_{NO}	%
Cognitive strategy	162	33	148	30
Self-checking	172	26	53	8
TOTAL	334	59	201	38

In other words, more than one third indicated their tendency to use cognitive strategy skills (*f* = 162, 33%) while three tenth indicated their tendency to use their self-checking skills (*f* = 172, 26%) on their responses to open-ended. It can be inferred that the students tended to use their cognitive strategy skills 1.3 times more than self-checking skills while responding to open-ended items.

The descriptive findings in Table 4.22 shown below show that the frequencies of the students' cognitive strategy skill levels on their responses to open-ended items were slightly higher (*f* = 162, 33%) than those on their responses to multiple-choice items (*f* = 129, 26%). Moreover, the frequencies of students' self-checking skill levels on their responses to multiple-choice items were slightly higher (*f* = 190, 29%) than those on their responses to open-ended items (*f* = 172, 26%).

Table 4. 22. Frequency Distribution in terms of Metacognitive Skills by Item Formats

		Multiple-choice		Open-ended	
Metacognitive subskills		*f*	%	*f*	%
Cognitive strategy	YES	129	26	162	33
	NO	166	34	148	30
	TOTAL	295	60	310	63
Self-checking	YES	190	29	172	26
	NO	67	10	53	8
	TOTAL	257	39	225	34
GRAND TOTAL	YES	319	55	334	59
	NO	233	44	201	38
	TOTAL	552	99	535	97

In other words, nearly one third indicated their tendency for cognitive strategy while more than one third did not on their responses to multiple-choice items. More than one third indicated their tendency for cognitive strategy while one third did not on their responses to open-ended items. While more than one third indicated their tendency for cognitive strategy on their responses to open-ended items (*f* = 162, 33%), lower than one third did not on their responses to multiple-choice items (*f* = 129, 26%). On the other hand, one third replied their tendency for self-checking on their responses to multiple-choice (*f* = 190, 29%) while a bit lower than one third replied their tendency for self-checking on their responses to open-ended (*f* = 172, 26%).

The descriptive findings from grand total in Table 4.21 above showed that the frequencies of students' metacognitive skills on their responses to multiple-choice items were similar (*f* = 319, 28%) to open-ended items (*f* = 334, 29%).

When the dependent variable is assessed at a nominal level, the Chi-square statistic is a non-parametric (distribution free) technique used to analyze group differences. Like all non-parametric statistics, the Chi-square is reliable regardless of how the data are distributed. In particular, it does not require homoscedasticity in the data or equality of variances between the research groups. (McHugh, 2013). In the current study, due to the variables being coded at nominal level, to see the difference and relationship between categorical variables, the Chi-Square Test for Independence was included for further inferential analysis.

Chi-Square Test for Independence was used to see the relationship between metacognitive subskills and multiple-choice item types. Independent observation and size of expected frequencies was met and assumptions were not violated. The Chi-Square statistic results indicated there was a significant relationship between metacognitive subskills and the solution of multiple-choice items, $X^2(1, N = 32) = 258.84$, $p < .05$. In multiple-choice items, while the students' tendency to use cognitive strategy skill decreases, the probability of their tendency to use self-checking skills increases. Similarly, the Chi-Square statistic results indicated there was a significant relationship between metacognitive sub skills and item types within the solution of open-ended items, $X^2(1, N = 32) = 212.05$, $p < .05$. In open-ended items, while the students' tendency to use cognitive strategy skill increases, the probability of their tendency to use self-checking skills also increases.

Moreover, to investigate the differential effect of item types on metacognitive skills, the Chi-Square Test for Independence was again examined to see the difference between metacognitive skills by item types. There was a significant difference between students' metacognitive skills usability by item types, $X^2(1, N = 32) = 179.45, p < .05$. The students who were responding open-ended were able to use metacognitive skills such as cognitive strategy more than those who were responding multiple-choice.

Table 4. 23. Frequency Distribution in terms of Affective Processes

Affective Processes	*f*	%
Positive feelings (e.g., Excitement, good, easy, happiness, comfort, confident, humor)	126	39.13
Negative feelings (e.g., worry, anxiety, stress, difficulty, sadness, fear, afraid)	85	26.40
TOTAL	322	100

Note: Other frequencies were nötr and not applicable.

The descriptive findings in Table 4.23 showed that fifth grade students' affective processes were independent variables. It included two levels in the study. They aroused positive feelings such as feeling happiness, comfort and negative feelings such as sad, regret. The frequencies of students' positive feelings on their responses to items were nearly 1.5 times more ($f = 126$, 39.13%) than the frequencies of negative feelings on their responses to items ($f = 84$, 26.40%). In other words, while two fifths indicated they aroused positive feelings toward the items ($f = 126$, 39.13%), nearly one third indicated they aroused negative feelings toward the items. To sum up, the findings revealed that the students aroused positive feelings toward different item formats.

Table 4. 24. Frequency Distribution in terms of Affective Processes by Item Types

	Multiple-choice		Open-ended	
Affective Processes	*f*	%	*f*	%
Positive feelings (e.g., excitement, good, easy, happiness, comfort, confident, fun)	72	44.44	58	35.37
Negative feelings (e.g., worry, anxiety, stress, difficulty, sad, fear, afraid)	35	21.60	45	27.44
TOTAL	162	100	159	100

Note: Other frequencies were nötr feeling and not applicable.

The descriptive findings in Table 4.24 showed the arousal of students' positive feelings toward multiple-choice items were two times more (f = 72, 44.44%) than the arousal of negative feelings toward those items (f = 35, 21.60%). Similarly, the arousal of students' positive feelings on open-ended items were nearly 1.3 times more (f = 58, 35.37%) than the arousal of negative feelings toward these items (f = 45, 27.44%).

Nearly two fifths aroused positive feelings toward open-ended items (f = 58, 35.37%) whereas nearly half aroused positive feelings toward multiple-choice items (f = 72, 44.44%). On the contrary, one fifth aroused negative feelings toward multiple-choice (f = 35, 21.60%) whereas three tenth aroused negative feelings toward open-ended items (f = 45, 27.44%).

To sum up, the descriptive findings revealed that the students aroused positive feelings toward multiple-choice items and negative feelings toward open-ended items. In addition to descriptive analysis, inferential analysis was conducted to explore whether there was a significant relationship between item types and affective processes, the Chi-Square statistics were held, and the results indicated there was a marginally significant relationship between affective processes and item types, $X^2(1, N = 32) = 54.92$, $p < .05$. While the probability of the students' arousal positively toward multiple-choice items increases, the probability of their arousal negatively toward open-ended items increases.

Table 4. 25. Frequency Distribution of Students' Area of Interest (AoI) Categories by Items

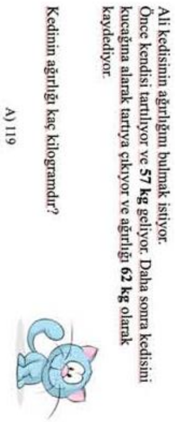
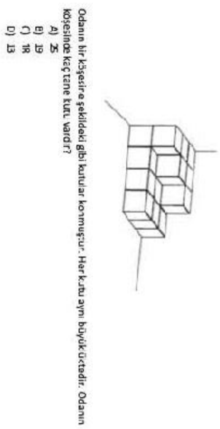
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Item1 	Procedural	Applying		
Numbers			21	65.63
Question root			5	15.63
Item 2 	Conceptual	Analyzing		

Table 4.25. (continued)

Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Boxes			16	50
Middle of boxes			8	25
Mixed areas (e.g., top&bottom, top&leftwall, top & alternatives)			3	9.38
 Aşağıdaki şekle göre Oya hesap makinesi ile MP3 çalar, kulaklık ve hoparlörün fiyatını toplamıştır. Elde ettiği sonuç 248'dir. Oya'nın yanıtı yanlıştır. Oya aşağıdaki hatalardan birini yapmıştır. Oya'nın yaptığı hata aşağıdakilerden hangisidir? A) Fiyatlardan birini iki kere toplamıştır. B) Üç fiyattan birini eklemeyi unutmuştur. C) Fiyatlardan birinin son basamağındaki rakamı yazmamıştır. D) Fiyatlardan birini toplamak yerine çıkarmıştır. **Item 3**	Procedural	Evaluating		
The price table			11	33.33
Question root			3	9.09

Table 4.25. (continued)

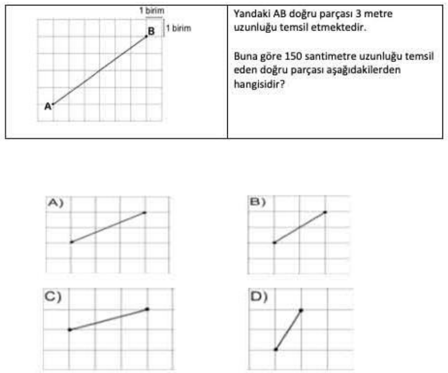
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Price table and 248			3	9.09
Item 4	Conceptual	Analyzing		
				
Between question root and main shape			7	21.88
Main shape			6	18.75
Not Applicable (i.e., NA)			6	18.75

Table 4.25. (continued)

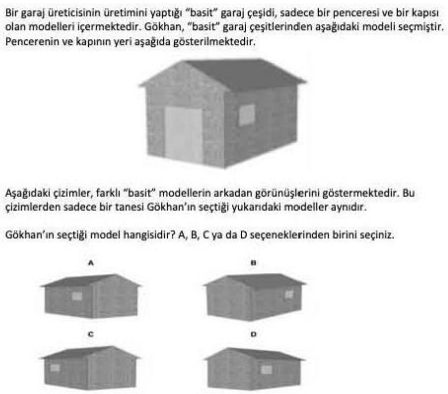
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Between question root, main shape and alternatives			4	12.5
A-B line			3	9.38
Question root			3	9.38
Item 5  Bir garaj üreticisinin üretimini yaptığı "basit" garaj çeşidi, sadece bir penceresi ve bir kapısı olan modelleri içermektedir. Gökhan, "basit" garaj çeşitlerinden aşağıdaki modeli seçmiştir. Pencerenin ve kapının yeri aşağıda gösterilmektedir. Aşağıdaki çizimler, farklı "basit" modellerin arkadan görünüşlerini göstermektedir. Bu çizimlerden sadece bir tanesi Gökhan'ın seçtiği yukarıdaki modeller aynıdır. Gökhan'ın seçtiği model hangisidir? A, B, C ya da D seçeneklerinden birini seçiniz. A B C D	Conceptual	Applying		

Table 4.25. (continued)

Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
The house			5	15.63
The house and question root			5	15.63
The house and alternatives			5	15.63
Not Applicable (i.e., NA)			5	15.63
Question root			3	9.38
Windows			3	9.38
Door			3	9.38

Table 4.25. (continued)

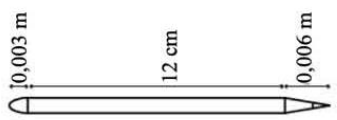
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Item 6 Aşağıdaki şekilde yer alan kurşun kalemin uzunluğu kaç santimetredir?  A) 21 B) 18,3 C) 15,6 D) 12,9	Procedural	Applying		
Numbers			12	37.50
Decimals			10	31.25
Between decimals and alternatives			5	15.63
Not Applicable (i.e. NA)			4	12.50

Table 4.25. (continued)

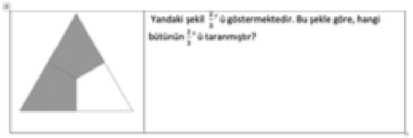
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Item 7	Conceptual	Analyzing		
				
Fraction of 2/3			8	24.24
Question root and alternatives			6	18.18
Question root and main shape			4	12.12
Alternative C			4	12.12
Question root			3	9.09
Alternative A-C			3	9.09

Table 4.25. (continued)

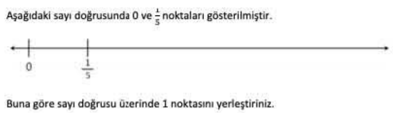
Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Not Applicable (i.e. NA)			3	9.09
Item 8 	Factual	Applying		
Fraction of 1/5			7	18.92
Not Applicable (i.e. NA)			7	18.92
Line between 0 and 1/5			6	16.22
Alternative D			6	16.22
Alternatives			3	8.11
Point 1 on the number line			3	8.11

Table 4.25. (continued)

Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Item 9 Aşağıdaki grafikte bir şirketin dört aylık kâr miktarı gösterilmiştir. Şirketin dört aylık toplam kârı kaç liradır?  A) 110 000 B) 120 000 C) 130 000 D) 140 000	Procedural	Applying		
Graph			17	53.13
Numbers on y axis			11	34.38
Item 10 Bir kasada 35 elma vardır. Bir çuvalda ise kasadakinden 124 fazla elma vardır. Buna göre 833 çuvaldaki toplam elma sayısı aşağıdaki işlemlerden hangisi ile bulunabilir? A) 35 x (833 – 124) B) 833 x (124 – 35) C) 833 x (35 – 124) D) 124 x (833-35) 	Procedural	Applying		

Table 4.25. (continued)

Items	*Revised Bloom's Knowledge Domain*	*Revised Bloom's Cognitive Domain*	*f*	%
Numbers in question root			12	36.36
Question root			8	24.24
Not Applicable (i.e. NA)			5	15.15
Between question root and alternatives			5	15.15
Alternative A-C			3	9.09

Note: The categories in which frequencies were more than 2 were kept.

According to the Areas of Interest (Table 4.25), also referred to as an AoI in an eye-tracking system, the results showed that students' AoI while responding to Item 1 (Procedural, Applying) was numbers (f = 21, 65.63%), Item 2 (Conceptual, Analyzing) was the visual shape (f = 24, 70%), Item 3 (Procedural, Evaluating) was the price table including numbers (f = 11, 33.33%), Item 4 (Conceptual, Analyzing) was the area between question root or the visual shape and only visual shape (f = 13, 40.63%), Item 5 (Conceptual, Applying) was the area between question root and the visual shape or only visual shape and alternatives (f = 15, 46.89%), Item 6 (Procedural, Applying) was numbers or decimals (f = 22, 68.75%), Item 7 (Conceptual, Analyzing) was the fraction of ⅔ (f = 8, 24.24%), Item 8 (Factual, Applying) was faction of ⅕ (f = 7, 18.92%), Item 9 (Procedural, Applying) was the graph as the visual shape (f = 17, 53.13%) and Item 10 (Procedural, Applying) was numbers (f = 12, 36.36%).

Hence, the students' eye-tracking metrics on the ten items in terms of AoI depicted that the students focused on the areas including "numbers" to which they would use it in the solution process while responding to items.

The research hypothesis relying on the research question 4 was that "the likelihood that 5th grade middle school students' rereading skill is related to their gender, total time on task (i.e., time spent by the student) and gaze shifts (i.e., number of looking back and forth into the item). In this process, the outcome variable was students' rereading skill and predictors were gender having nominal scale of measurement, total time on task and gaze shifts, each of which had ratio scale of measurement. The outcome variable was categorical and dichotomous (coded as 1: Yes, 0: No) so that it was binomial whereas the predictor variables were categorical and dichotomous for gender but numerical and continuous scores for total time on task and gaze shifts.

The research question 4.4 was "Do total time on task and gaze shifts have an impact on predicting rereading or not while responding to items?" To answer this question, Binomial Logistic Regression (Peng, Lee & Ingersoll, 2002) was conducted and fifth students participated in this process of the study where their time on task was collected by my note taking and gaze shifts by eye-tracking tools. Results indicated that of these 32 children, gender distribution was with 15 (46.9%) female and 17 (53.1%) male. The total time on task ranged from 0.51 to 15.57 minutes, with a mean of 4.07 minutes and standard deviation of 2.24 minutes. Gaze shifts ranged from 0 to 30, with a mean of 4.18 and standard deviation of 4.35. Descriptives for rereading skills were 149 (46.6%) for using and 63 (19.7%) for not while responding to a total of ten multiple-choice and open-ended items.

The three-predictor binomial logistic model was fitted to the data to test the research hypothesis regarding the relationship between likelihood of fifth grade students' rereading skill and their total time on task, gaze shifts and gender. The results in Table 4.24 showed that:

> the predicted logit of (REREADING) = .93 + (-.27)*TOTAL TIME + (-.35)*GAZESHIFT + (.70)*GENDER.

The findings indicated that the likelihood ratio test of the full model versus null model (model with intercept only) was statistically significant, X^2 (4) = 54.25, Nagelkerke R^2 = .33. In other words, the logistic model was more effective than the null model. Using Wald's statistic, total time on task variable was significant, Wald's X^2 = 5.82 with a df = 1. Gaze shifts variable was significant, Wald's X^2 = 17.82 with a df = 1. Gender variable was significant, Wald's X^2 = 3.80 with a df = 1. According to the model, the log of the odds of a student using a rereading sub-skill of cognitive strategy was negatively related to total time on task and gaze shifts; and positively related to gender. In other words, higher number of gaze-shifts and total time on task, less likely the student tends to use rereading subskill. Each point increase on total time on task and gaze shifts was associated with a decrease in the probability of rereading by a factor of .76 and .70 respectively. Given the same total time on task and gaze shifts, girls were less likely to use rereading sub-skill than boys. The odds of a boy tended to use rereading sub-skill were 2.01 times greater than the odds for a girl. As an overall, 91% of occurrences (1) was correctly predicted while 49.2% of non-occurrences (0) was correctly predicted in this model (see Table 4.26).

Table 4. 26. Logistic Regression Analysis of 32 Students' Predictors for Rereading Subskill

Predictor	*β*	SE of *β*	Wald	*df*	*p*	Odds ratio
Constant	.93	.47	3.98	1	.046	2.53
Gender (1- girl, 2-boy)	.70	.36	3.80	1	.05	2.01
Total Time on the task (min.)	-.27	.11	5.82	1	.02	.76
Gaze shifts	-.35	.08	17.82	1	.00	.70

Note. Cox and Snell R^2 = .23. Nagelkerke R^2 = .33. All statistics reported herein were 2 decimal places in order to maintain statistical precision.

4.3.2. Summary of multimodal phase: Neuroeducation with the middle school students

The results of the current study revealed are in line with the hypotheses formulated, H_1: there was a significant difference between the amount of reflection of students' cognitive strategy skill levels on their responses in favor of open-ended items. In open-ended items, while the students' tendency to use cognitive strategy skill increases, the probability of their tendency to use self-checking skills also increases. The 5^{th} grade students tended to use cognitive strategies more frequently in responding to open-ended items than in multiple-choice items. There is a significant difference between the amount of reflection of students' self-checking skill levels on their responses in favor of open-ended items. There was a significant difference between the amount of reflection of students' worry levels on their responses in favor of open-ended items. Students aroused negative feelings toward open-ended items.

Eye-tracker metric results triangulated the data and provided validity of the findings that in terms of Area of Interest showed the students focused much on the areas including “numbers” to which they would use it in the solution process while responding to items. When the students began to answer an item, the Area of Interest (AoI) (i.e., they focus at most on the items) was the question root of that item. However, while responding to the items, their AoI were usually “figures” and “numerical information.” One of the significant cognitive strategy sub-skills was *rereading*. The students’ tendency of rereading was significantly predicted by their total time spent in the process, gaze shifts and gender.

It might be a reason related to using rereading sub-skill appropriately under cognitive strategy skill that the students may tend to focus on the items deeply and read the question root and related area of interest without diverging gaze shifts too much.

a. Middle school participants analysis from think-aloud process, problem-solving steps with hand-writing analysis, eye-tracking and GSR values

Fifth grade students (N = 32) were the participants in neuroeducation process whose 53% were male and 47% were female. The following parts synthesize their responses to questions, and analysis of their Thinkaloud process, problem-solving steps, eye-tracking and GSR values. In addition, how the analysis was conducted by the researcher and co-coder were illuminated and remarked in brackets in case of disagreements between co-coders.

Table 4. 27. Students Performance on Item I Difficulty Measures

Item 1	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	.84 (Easy)	90.6%	9.4%	-	84.4%	15.6%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)	X					
Self-checking skills (e.g., finding errors)	X					
Worry	X					
Effort	X					

Students' performance on item I related to multimodal data (Procedural, Applying). Most of the students (*n* = 27) responded the item correctly. The item difficulty analysis overlaps with the students think-aloud responses that found to be easy. Arousal levels relying on EDA values from GSR (65.6%) was also coincided with students' feelings during think-aloud. The students tend to use low metacognitive skills. While responding, they tend to use one solution step such that question root, subtraction, operation were the solution steps used to organize knowledge given in the item. Students tend to look back and forth once or two times. According to the co-coder's coding, the first participant looked back 3 times. According to my coding, the first participant looked back 4 times [disagreement between co-coders]. No information was obtained about concentration. According to the co-coder's coding, the first participant made a low effort. According to me, no information was obtained from the first participant [disagreement between co-coders]. The table 4.27 provides a visualization of item I.

Table 4. 28. Students Performance on Item II Difficulty Measures

Item 2	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
	.88 (Easy)	Easy 81.3%	Moderate 18.8%	Difficult -	True 87.5%	False 12.5 %
	Low	Middle	High			
Cognitive strategy (e.g., rereading)	X					
Self-checking skills (e.g., finding errors)	X					
Worry	X					
Effort	X					

Students' performance on item II related to multimodal data (Conceptual, Analyzing). Most of the students (n = 28) responded the item correctly. The item difficulty analysis overlaps with the students think-aloud responses that found to be easy. Arousal levels relying on EDA values from GSR (59.4%) was also coincided with students' feelings during think-aloud. The students tend to use low metacognitive skills. While responding, they tend to use mostly one solution step such that question root, shape, counting, imagination, re-reading, operation, trying to understand, using different roads to organize knowledge. Students tend to look back and forth once or three times. They responded they felt high confidence level and less worry. According to the co-coder's coding, the first participant felt comfort, according to me, there was no anxiety for the first participant [disagreement between co-coders]. Second participant was a little uncomfortable and controlled. The third participant was confident. They did not show the sign of giving up that most of them kept perseverence. The table 4.28 provides a visualization of item II.

Table 4. 29. Students Performance on item III Difficulty Measures

Item 3	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
	0.22 (Hard)	Easy 21.9%	Moderate 31.3%	Difficult 46.9%	True 21.9%	False 62.5%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)			X			
Self-checking skills (e.g., finding errors)		X				
Worry	X					
Effort			X			

Students' performance on item III related to multimodal data (Procedural, Evaluating). Most of the students (*n* = 20) responded the item incorrectly. The item difficulty analysis overlaps with the students think-aloud responses that found to be very difficult. Arousal levels relying on EDA values from GSR (43.8%) was also coincided with students' feelings during think-aloud. The students tend to use high metacognitive skills. While responding, they tend to use cognitive strategies such as question root, shape, guessing, operation, alternatives, read aloud were used to organize knowledge. Students tend to look back and forth at least four times. They responded they felt low confidence level but low worry. According to the co-coder's coding, no information was obtained from the first participant, according to me, First participant read the question again [disagreement between co-coders]. According to the co-coder's coding, the first participant looked back 9 times, according to me, the first participant looked back 11 times [disagreement between co-coders]. Participants with low self-confidence felt worried and unconfident. Participants with high self-confidence felt comfortable and calming. The participant, who was moderate confident, had a trembling voice and felt stressed. Some of them (*n* = 13) said to think giving up due to low perseverence. The table 4.29 provides a visualization of item III.

Table 4. 30. Students Performance on Item IV Difficulty Measures

Item 4	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.53 (Medium)	37.5%	-	37.5%	53.1%	31.3%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)	X					
Self-checking skills (e.g., finding errors)		X				
Worry		X				
Effort	X	X	X			

Students' performance on item IV related to multimodal data (Conceptual, Analyzing). Half of the students (*n* = 17) responded the item correctly and half did incorrectly (*n* = 10). The item difficulty analysis did not exactly overlap with the students think-aloud responses that half (*n* = 12) evaluated as easy whereas half (*n* = 12) evaluated as difficult. Arousal levels relying on EDA values from GSR (56.3%) was also coincided with students' feelings during think-aloud. The students tend to use "low" to "moderate" metacognitive skills. While responding, they tend to use cognitive strategies such as question root, transformation, subtraction, counting, re-expression, eliminating, and ratio were used to organize knowledge. Students tend to look back and forth four or five times. The students' feelings were fluctuated that both negative and positive feelings detected. Moreover, the students' effort was also varied into low, middle and high. According to the co-coder's coding, the first participant used two solution methods, according to me, the students used three solution methods [disagreement between co-coders]. Also, the co-coder's coding, no answer was received from the first participant, according to me, the first participant read the question again [disagreement between co-coders]. In addition, the co-coder's coding, the first participant did not feel any negative emotion, according to me, the first participant was comfortable [disagreement between co-coders]. Fourth student was comfortable due to alternatives whereas third one was excited and anxious. No answer was received about concentration. The table 4.30 provides a visualization of item IV.

Table 4. 31. Students Performance on Item V Difficulty Measures

Item 5	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.31 (Nearly difficult)	71.9%	21.9%	6.3%	31.3%	68.8%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)		X				
Self-checking skills (e.g., finding errors)	X					
Worry	X					
Effort	X	X				

Students' performance on item V related to multimodal data (Conceptual, Applying). Most of the students (n = 22) responded the item incorrectly. The item difficulty analysis found to be nonsimilar with the students think-aloud responses that many of them (n = 23) evaluated the item as easy. Arousal levels relying on EDA values from GSR (68.8%) was also coincided with students' feelings during think-aloud. The students tend to use "low" metacognitive skills (62.5%). Rereading subskills were used moderately whilst self-checking was done low frequency. While responding, they tend to use cognitive strategies such as such as rereading question root, imagination, analyzing shape, perspective, going over alternatives were used to organize knowledge. Nearly all of the students did not tend to look back or once. Nevertheless, those who were looking back and forth implied four or six times. Near half (43.8%) felt positive while responding. Moreover, the students' effort was found to be between "low" to "middle." According to the co-coder's coding, the first participant looked back 5 times, according to me, the first participant looked back once [disagreement between co-coders]. Some students had high self-confidence while few of them felt comfortable due to alternatives, some were neutral. The table 4.31 provides a visualization of item V.

Table 4. 32. Students Performance on Item VI Difficulty Measures

Item 6	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.38 (Nearly difficult)	68.8%	21.9%	9.4%	37.5%	62.5%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)	X					
Self-checking skills (e.g., finding errors)			X			
Worry	X					
Effort	X					

Students' performance on item VI related to multimodal data (Procedural, Applying). Most of the students (*n* = 20) responded the item incorrectly. The item difficulty analysis found to be nonsimilar with the students think-aloud responses that many of them (*n* = 22) evaluated the item as easy. Arousal levels relying on EDA values from GSR (71.9%) was also coincided with students' feelings during think-aloud. The students tend to use "moderate" metacognitive skills (62.5%). Rereading subskills were used very low frequency whilst self-checking subskills mostly preferred to be used. While responding, they tend to use cognitive strategies, for instance, elimination, transformation, guessing, rereading question root, analysis of shape, addition, and transformation were used to organize knowledge. Half of the students (50%) tended to look back and forth once, three or four times. One third showed arousal of positive feelings (31%) while other one third showed of negative feelings (34%). So, the students tended to feel low level of worry. Moreover, the students' effort was found to be "low." According to the co-coder's coding, the first participant looked back 7 times, according to me, the first participant looked back 2 times [disagreement between co-coders]. According to me, the first participant found an error, according to the co-coder's coding, no answer was received from the first participant [disagreement between co-coders]. According to the co-coder's coding, the first participant had high self-confidence, according to me, no answer was received from the first participant [disagreement between co-coders]. All in all, they tend to respond the item via one solution process. The table 4.32 provides a visualization of item VI.

Table 4. 33. Students Performance on Item VII Difficulty Measures

Item 7	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.25 (Hard)	37.5%	31.3%	31.3%	25%	62.5%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)			X			
Self-checking skills (e.g., finding errors)		X				
Worry		X				
Effort		X				

Students' performance on item VII related to multimodal data (Conceptual, Analyzing). All participants found this question challenging. Most of the students (n = 20) responded the item incorrectly. The item difficulty analysis was not found to be similar with the students think-aloud responses. Students answers during think-aloud oscillated that they found the item easy, moderate, and difficult even though item difficulty indicated as a hard item. Arousal levels relying on EDA values from GSR (59.4%) was also coincided with students' feelings during think-aloud. The students tend to use "low" (66%) to "moderate" metacognitive skills (16%). Cognitive stratgey subskills preferred to be used in high amount whilst self-checking subskills preferred in moderate amount. While responding, they tend to use cognitive strategies, for instance, rereading question root, drawing, expansion, analysis from question root to shape, simplification, re-interpretation, elimination, part whole relationship, understanding question, drawing 3 different ways, expanding fraction, were used to organize knowledge. More than half of the students (63%) tended to look back and forth once, four or seven times. One fifth showed arousal of positive feelings (22%) while others showed of negative feelings (41%). So, the students tended to arouse moderate level of worry. Moreover, the students' effort was found to be "moderate" as well due to the challenging item. According to me, the first participant did not check its answer, according to the co-coder, no answer was received from the first participant [disagreement between co-coders]. According to me, no answer was received from the first participant [disagreement between co-coders]. According to the co-coder's coding, asking questions to stay on track was coded as "yes" for the last participant, according to me, asking questions to stay on track was coded as "not applicable" for the fourth participant. According to me, the first participant had high self-confidence, however, according to the co-coder's coding, no answer was received from the first participant [disagreement between co-coders]. The table 4.33 provides a visualization of item VII.

Table 4. 34. Students Performance on Item VIII Difficulty Measures

Item 8	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.53 (Medium)	68.8%	25%	6.3%	53%	47%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)	X					
Self-checking skills (e.g., finding errors)		X				
Worry		X				
Effort	X	X	X			

Students' performance on item VIII related to multimodal data (Factual, Applying). Half of the students found this question moderately challenging. Nearly half of the students (*n* = 17) responded the item correctly. The item difficulty analysis was not found to be similar with the students think-aloud responses. Students answers during think-aloud oscillated that they found the item generally easy even though item difficulty index indicated as medium item. Arousal levels relying on EDA values from GSR (59.4%) was also coincided with students' feelings during think-aloud. 44% of them (*n* = 14) was feeling positive whereas 22% of them (*n* = 7) was feeling negative toward the solution process. The students tend to use "low" (38%) metacognitive skills in which they tend to show low cognitive strategy subskills whilst moderate self-cheking subskills. Cognitive strategies such as rereading question root, analysis of shapes, going over alternatives, drawing, drawing 1/5 piece of whole, counting, were used to organize knowledge. The effort spent by the students was found to be fluctuated from low (*n* = 12, 37.5%) to high (*n* = 8, 25%). According to the co-coder's coding, the first participant looked back 4 times, according to me, the participant looked back 3 times [disagreement between co-coders]. Other participants looked back 3 times and 6 times. According to the co-coder's coding, asking question to stay on track was coded as "not applicable" for the first participant, according to me, asking a question to stay on track was coded as "valid" for the first participant [disagreement between co-coders]. The table 4.34 provides a visualization of item VIII.

Table 4. 35. Students Performance on Item IX Difficulty Measures

Item 9	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
	0.78 (Easy)	Easy	Moderate	Difficult	True	False
		78.1%	15.6%	6.3%	78%	22%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)			X			
Self-checking skills (e.g., finding errors)		X				
Worry	X					
Effort	X					

Students' performance on item IX related to multimodal data (Procedural, Applying). Most of the students found this question very easy. Similarly, 78% of the students (*n* = 25) responded the item correctly. The item difficulty analysis was not found to be similar with the students think-aloud responses. Students answers during think-aloud declared that they found the item generally easy as well as item difficulty index indicated as an easy item. Arousal levels relying on EDA values from GSR (71.9%) was also agreed with students feelings during think-aloud. Half of them (*n* = 16) was feeling positive toward the solution process. The students tend to use "high" metacognitive skills in which they tend to show high cognitive strategy subskills and moderate self-cheking subskills. Nearly all of the students tended to look back and forth once to five times. Cognitive strategies such as questioning the shape, rereading question root, operation, sorting, addition, analysis of questin root to shape were used to organize knowledge. According to the co-coder's coding, the first participant did not re-express the question. On the other hand, according to me, no answer was received from the first participant [disagreement between co-coders]. The table 4.35 provides a visualization of item IX.

Table 4. 36. Students Performance on Item X Difficulty Measures

Item 10	Item difficulty index from item analysis	Item difficulty index from think-aloud			Students' responses	
		Easy	Moderate	Difficult	True	False
	0.75 (Easy)	75%	18.8%	6.3%	75%	25%
	Low	Middle	High			
Cognitive strategy (e.g., rereading)			X			
Self-checking skills (e.g., finding errors)		X				
Worry	X					
Effort	X	X				

Students' performance on item X related to multimodal data (Procedural, Applying). Most of the students found this question easy. Similarly, 75% of the students (*n* = 24) responded the item correctly. The item difficulty analysis was totally found to be similar with the students think-aloud responses and responses. Students answers during think-aloud declared that they found the item generally easy as well as item difficulty index indicated as an easy item. Arousal levels relying on EDA values from GSR (53%) was also agreed with students' feelings during think-aloud. Half of them (*n* = 16) was feeling positive toward the solution process. The students tend to use "moderate" metacognitive skills in which they tend to show high cognitive strategy subskills and moderate self-cheking subskills. Some of the students tended to look back and forth once to three times. Cognitive strategies such as basic operations, analyis of alternatives, using numbers, rereading question root were used to organize knowledge. According to the co-coder's coding, the first participant used two solution methods, whereas according to me, the first participant used one solution method [disagreement between co-coders]. According to me, the first participant did not re-express the question, on the other hand, according to the co-coder's coding, no answer was received from one participant [disagreement between co-coders].

Moreover, in line with the co-coder's coding, the first participant was calm and confident, however, according to me, the first participant was neutral and comfortable [disagreement between co-coders]. The table 4.36 provides a visualization of item X.

b. Sample heat map figures from eye-tracking analysis related to students achievement

The below figures (Figure 4.36 – Figure 4.49) depicted the heap maps from student participants. The students, who responded the item I correctly and incorrectly, smooth gaze movements were found to be on the question root as in Figure 4.36 and 4.37.

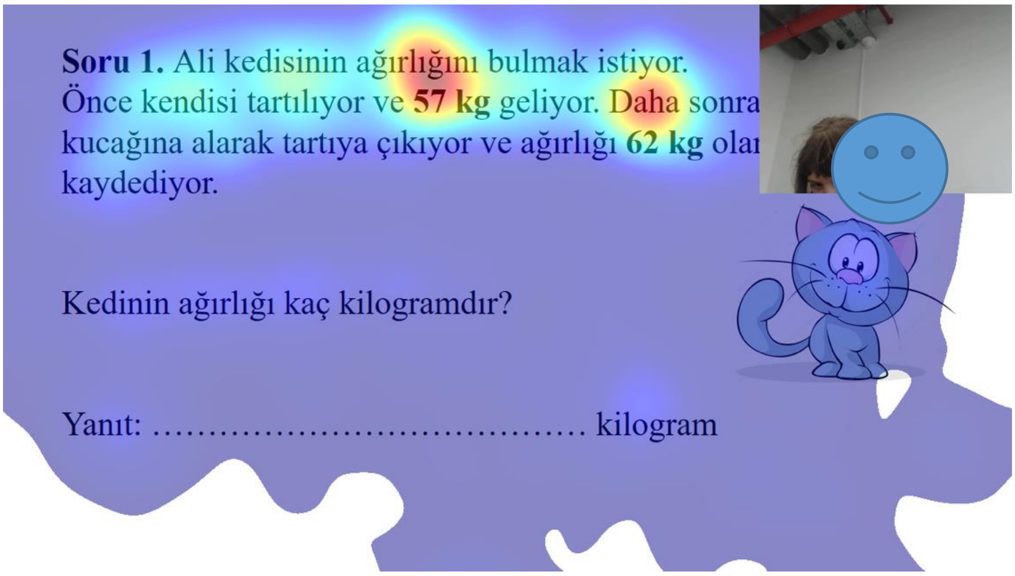


Figure 4. 36. Heatmap from a successful student at item I who gaze shifting on question root

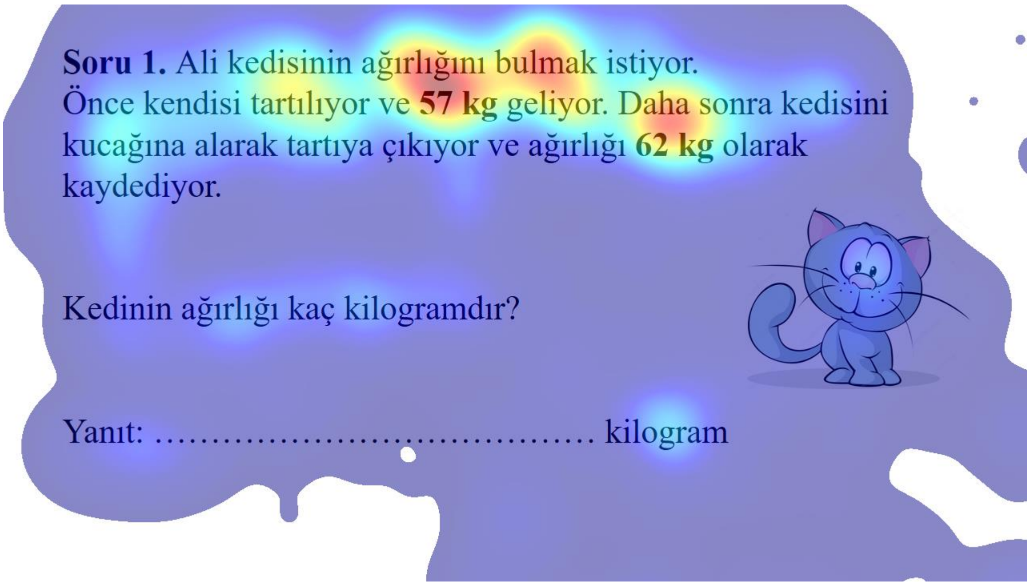


Figure 4. 37. Heatmap from an unsuccessful student at item I who gaze shifting on question root

The student's, who responded the item II correctly, smooth gaze movements were on the top of the given shape (see Figure 4.38) whereas the student's, who responded the item II incorrectly, smooth gaze movements were on the middle of the given shape (see Figure 4.39).

Figure 4. 38. Heatmap from a successful student who gaze shifting on the top of the shape

Figure 4. 39. Heatmap from an unsuccessful student who gaze shifting on the middle of the shape

The student's, who responded the item III correctly, smooth gaze movements were on the given number settle at the bottom of the screen (see Figure 4.40) whereas the student's, who responded the item III incorrectly, smooth gaze movements were on the upper middle unit of given accessories (see Figure 4.41).

Figure 4. 40. Heatmap from a successful student who gaze shifting on the given total number

Figure 4. 41. Heatmap from an unsuccessful student who gaze shifting on the upper middle unit

The student's, who responded the item X correctly, smooth gaze movements were on the question root (see Figure 4.42) whereas the student's, who responded the item X incorrectly, smooth gaze movements were on the options (see Figure 4.43).

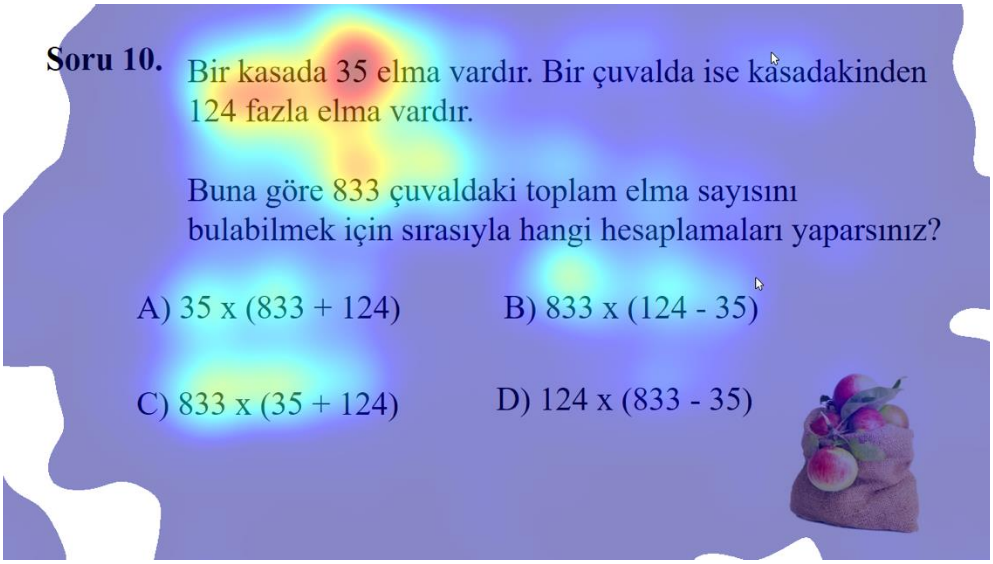


Figure 4. 42. Heatmap from a successful student who gaze shifting on the question root

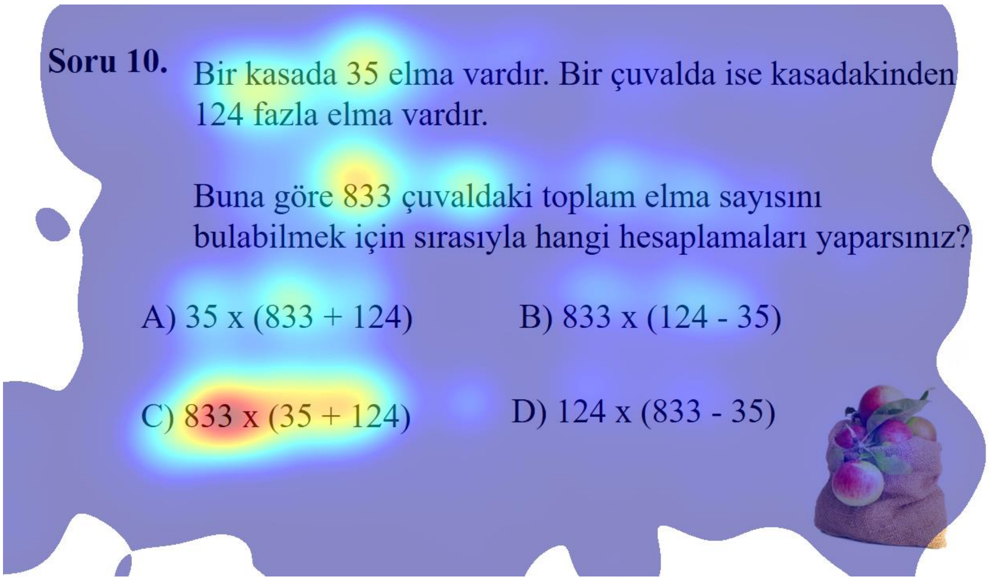


Figure 4. 43. Heatmap from a successful student who gaze shifting on an option

The gaze movement mean of the students who responded the item IV correctly (see Figure 4.44) differed from those who responded the item IV incorrectly (see Figure 4.45). (Item IV was a moderate difficulty item)

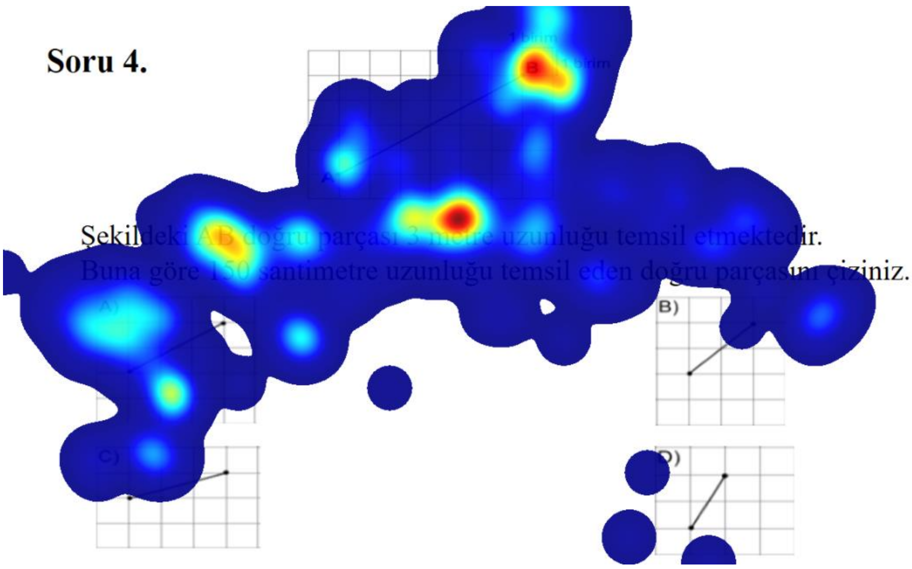


Figure 4. 44. Heatmap mean of the successful students who were gaze shifting

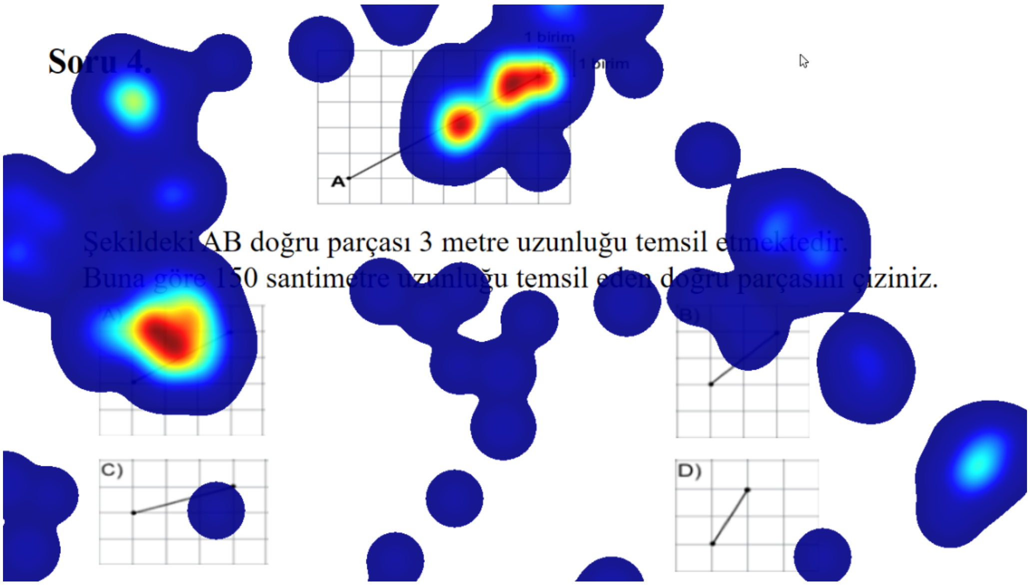


Figure 4. 45. Heatmap mean of the unsuccessful students who were gaze shifting

The gaze movement mean of the students from Group 1 who responded the item III correctly (see Figure 4.46) differed from the students from Group 1 who responded the item III incorrectly (see Figure 4.47). Heat maps variability of unsuccessful students were shown more scattered than the other (Item III was a hard OE item).

Figure 4. 46. Heatmap mean of the successful students who were gaze shifting from Group 1

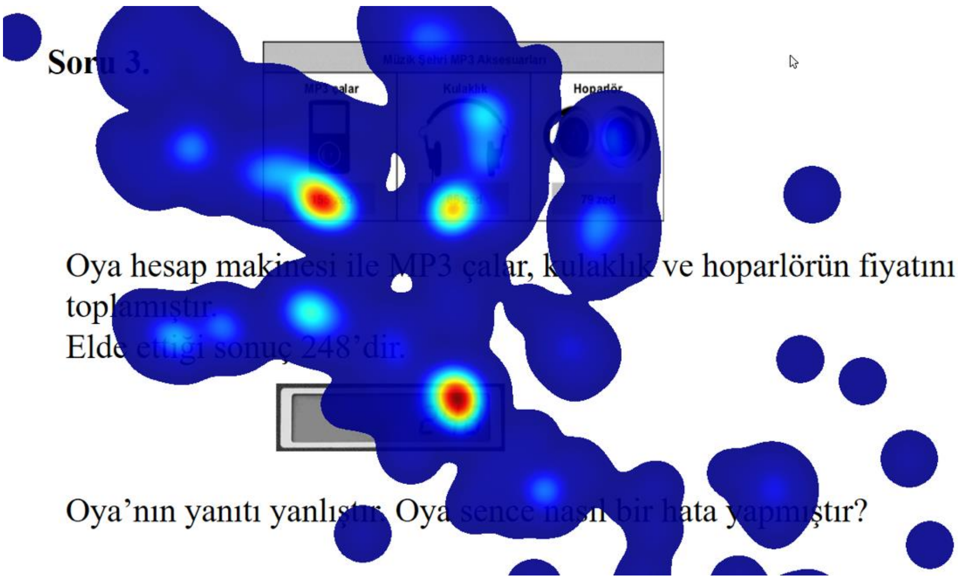


Figure 4. 47. Heatmap mean of the unsuccessful students who were gaze shifting from Group 1

The gaze movement mean of the students from Group 2 who responded the item III correctly (see Figure 4.48) differed from the students from Group 1 who responded the item III incorrectly (see Figure 4.49). Heat maps variability of successful students were shown generally on options and given shapes whereas others on question root and given shapes (Item III was a hard MC item).

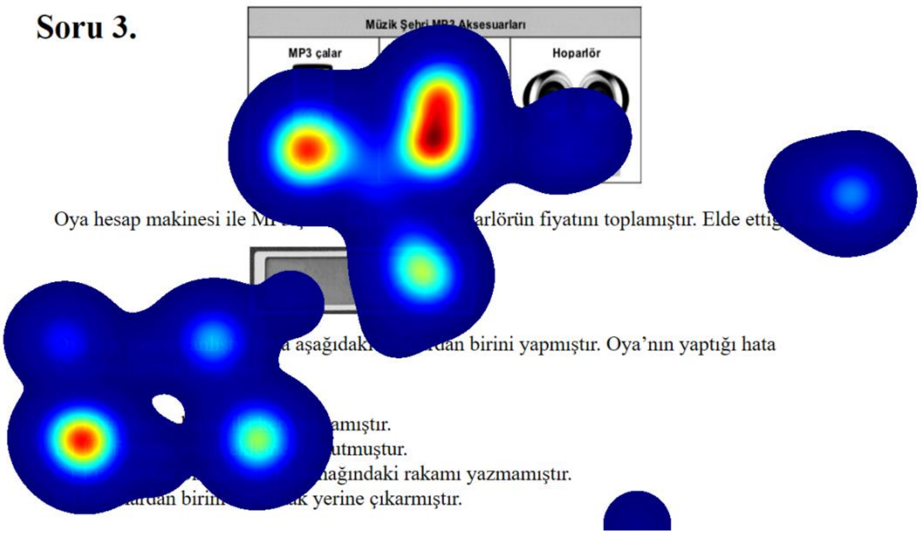


Figure 4. 48. Heatmap mean of the successful students who were gaze shifting from Group 2

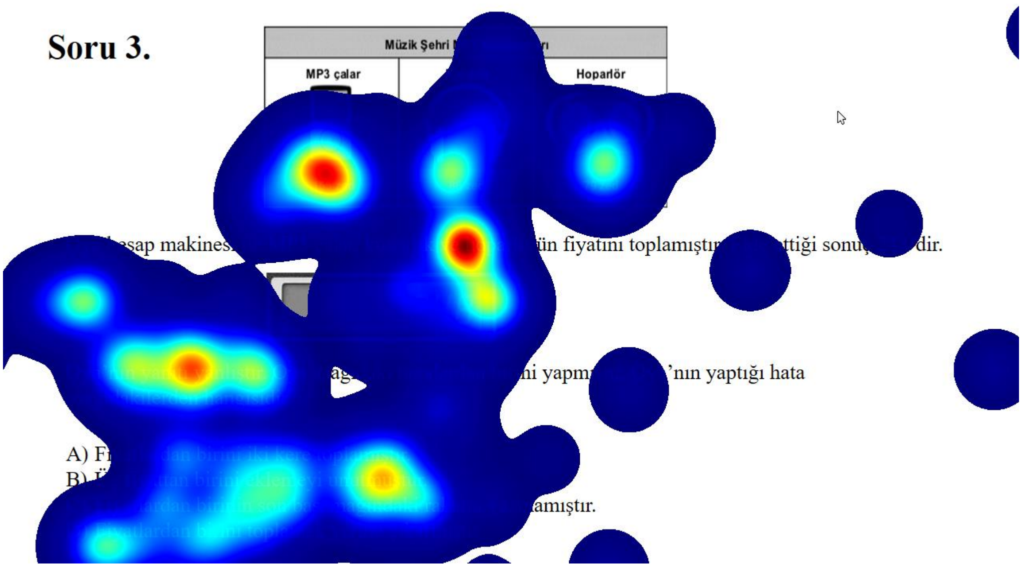


Figure 4. 49. Heatmap mean of the unsuccessful students who were gaze shifting from Group 2

4.4. Results of Modeling a Deep Data System for Metacognitive and Affective Processes

In this section, based on all the thesis deep data, my theoretical perspective on expert systems design (i.e., deep data modeling) from the perspective of an educational scientist was shared. The deep data modeling that I have been reached was concluded by summarizing the whole picture (explore Figure 4.51). In recent years, the value of data has been increasing, and there are phrases such as “data is a new oil”. IBM, which was introduced as the first admiral ship of the U.S.A in the world of informatics, started to talk about the most important data, not big data, but “deep data” (Türkoğlu, 2021). As a matter of fact, in this study, 32 student data were studied, and results were obtained as a result of deep data and their analysis. The study once again revealed the importance of deep data. It is known that scientists, computer programmers, and cognitive scientists need to train systems while designing intelligent systems. If this system data is collected from real human data and the system is trained in this context, then its authentic effect will gain importance. IBM has determined that "humanity today can only 'capture' one percent of the possible dataset that has emerged; the remainder it believes is at depths that humans have yet to reach."

According to the model design that emerged from this study, the necessity of designing systems that increase human-computer interaction for students has emerged. Thus: 1) A good deep data-based design is without a good understanding of people, 2) It is important that the question levels are determined by experts and taught to the system. 3) Recognizing the stress of students during question solving (especially when solving open-ended questions) and giving suggestions to reduce the stress level quickly, 4) When the student gives wrong answers, the question level can be lowered (from higher-order to lower-order), 5) The system receives feedback from the student and revises the content. 6) It can give personalized automated reports, 7) However, in the good design of these systems, the problems being pre-defined will make things easier. 8) Other data not used in this study can also be collected, as an illustration, EEG, fNIRS 3D brain scanning data, genetic biomarkers (Ahmad et al., 2011) from students and in-depth research on engagement. Within the scope of this study, data collection on cognitive load, dwell time, pupil dilation, self-explanation, and other emotional constructs, for which I could not reach deep data in the limited time intervals, may be good in terms of shedding light on future studies.

At best, five years for all these designs to develop with common sense and interdisciplinary approaches; it has been learned and foresighted from the studies in the world that it takes 20-30 years to reach the level that the society can reach.

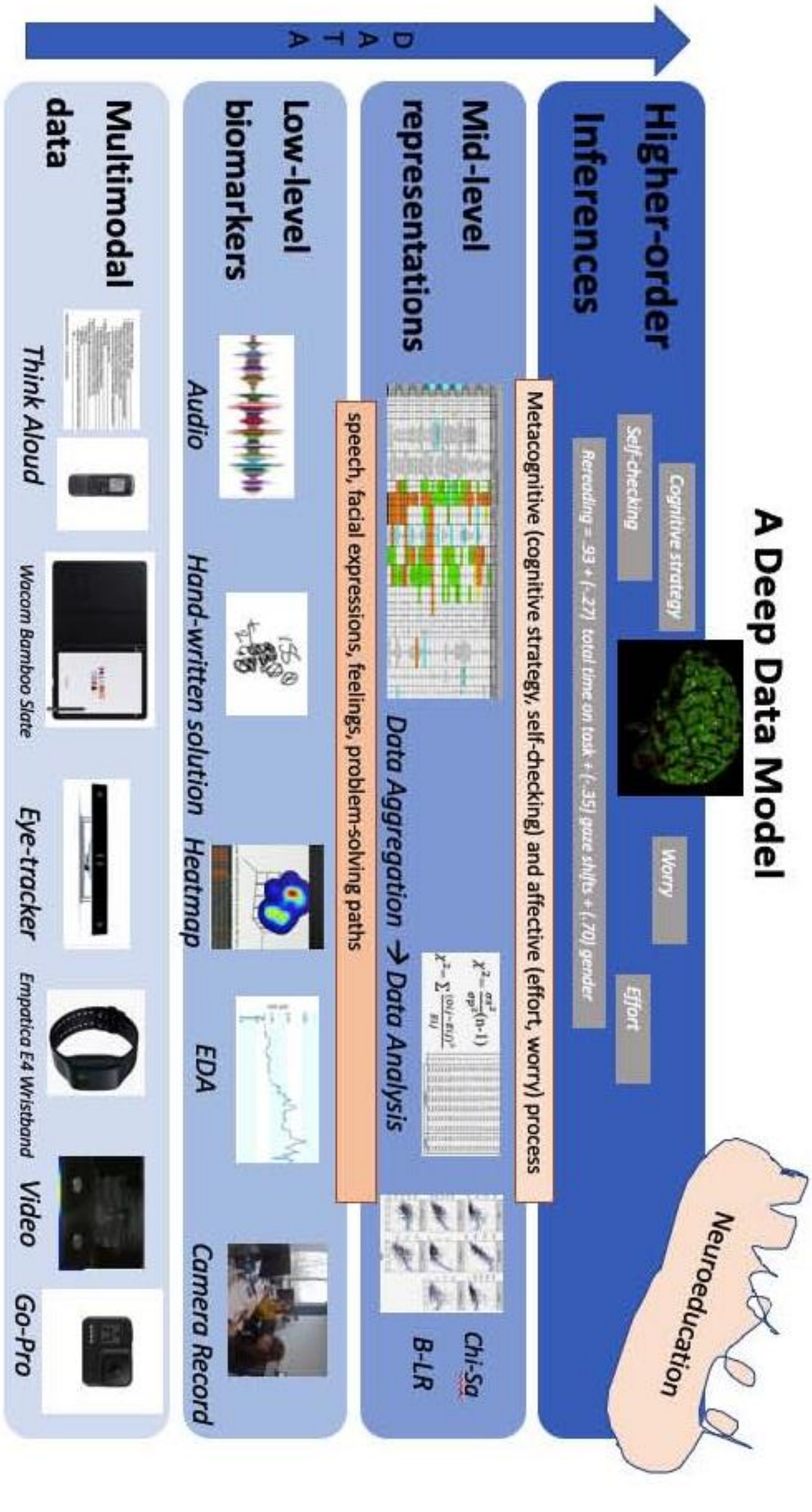


Figure 4. 50. A Deep Data Modeling for Metacognition and Affective Processes from Middle School Students

CHAPTER 5

DISCUSSION AND CONCLUSION

This mixed-methods multimodal research study produces value-added findings with regards to Turkish middle school mathematics teachers' response to curriculum change through an ecological approach, their preferences on teaching method and measurement-evaluation strategy after curriculum change, and 5th grade students' reactions and responses to different innovative item types via metacognitive skills and affective processes. The critical findings of this study are discussed in this chapter.

5.1. Teachers' Agency on Teaching Method and Assessment after Curriculum Change

The present study investigated middle school mathematics teachers' response to curriculum change through an ecological approach and 5th grade students' reactions and responses to different types of item formats with respect to metacognitive skills and affective processes utilizing an eye-tracking tool. Prior to data collection, it was hypothesized that after the curriculum change, teachers would be more adaptable and act as agents of change during enactment of preparation of in-class authentic assessment items, teaching method and measurement and evaluation preferences. Also, it was hypothesized that the curriculum change would promote mathematics teachers' adoption of dynamic and active teaching methods in the classroom, and use of more authentic, formative assessment techniques. With respect to the first research question, it was found that the middle school mathematics teachers generally tended to use traditional objective testing. Their authentic teacher-made items were periodically found to be developed in line with the procedural level of knowledge and to some extent conceptual, some factual, and rarely metacognitive. It was also found that half of the items were developed in line with “applying” the cognitive process dimension within the context of the Revised Bloom’s Taxonomy. No teacher used the “creating” level of the cognitive process dimension. Accompanying this taxonomy, the authentic teacher-made items were identified in line with the TIMSS Framework and the complementary findings show that those items mostly contributed to the development of a very basic cognitive level equal to knowing the level of the framework. The document analysis of learning outcomes in the 5th-grade national mathematics curriculum was found to be related with one-fifth HoTs whereas four-fifths were found to be related with LoTs. To conclude, the findings reveal that curriculum change did not lead to full renewal of teacher practices.

Similar findings were also reported by Delil and Özcan (2019), Grant and Gareis (2015), Hartel and Strimel (2019), İnceçam et al. (2018), Khan and Inamullah (2011), McREL International (2017), Özcan and Delil (2018), and Tofade, Elsner and Haines (2013). For instance, Özcan and Delil (2018) show that most of the items they analyzed were compatible with the mathematics learning outcomes, some were intended to measure multiple outcomes at the same time. In line with their results, the curriculum change was not reflected at that time; some mathematics teachers’ learning and teaching process and mathematics curriculum were not inspected by those teachers before its implementation. Similarly, following the analysis on the examination items related to measuring the amount of LoTs and HoTs, the findings reveal mathematics teachers’ tendency toward the preparation of examination items as LoTs, as declared by Haynes (1935, as cited in Khan & Inamullah, 2011; Gall, 1970) and Lee (2015) and the McREL International (2017). The study offers both a national and international perspective and states that teachers still prepare the items mostly based on LoTs rather than HoTs. In addition, the fact that the teachers mostly tended to ask LoT questions over HoTs supports earlier studies on the estimation that 70 to 80% of all items required the simple recall of facts (i.e. LoT items), while only 20 to 30% required the HoT items which triggered thought processes of clarifying, generalizing, and making inferences from the items (Haynes, 1935, as cited in Khan & Inamullah, 2011; Gall, 1970). Gall's (1970) findings show that teachers tended to prepare items having 60% recall, 20% procedural, and 20% thought provoking. Additionally, Lee (2015) reveals that 79% of the total items asked were lower-order questions whereas only 5% were targeted to measure higher-order items.

Research studies (e.g., Hartell & Stimel, 2019; Solanki & Evans, 2020) have discussed how higher-order questioning in examinations, testing or in-class assessment promotes learners' deeper learning. Therefore, teachers should provide better opportunities and teaching-learning experiences for their students and better awareness, ability and effort to assess their higher-order knowledge or thinking skills. In the current study, 19.1% of the items measured HoT whereas 80.8% measured LoT. Although, as Hynes (1935) claims, 20% of the items should be HoT for a satisfactory approach to assessment, there is no general acceptance of this among educational evaluators. The central idea is that teachers should have a desire, feel responsibility and experience in item analysis before implementation, and have experience with different types of measurement and evaluation strategies other than testing and should be open to such experiences. They may frequently need professional support to design appropriate learning activities, as well as assessment procedures, but not limited to tests.

This study suggests that some mathematics teachers working in private schools, and who have received theory-based training in curriculum development and assessment in cooperation with a faculty of education, were found to develop their deep knowledge and experience in preparing authentic in-class examinations and to increase the quality of HoT in their examinations. However, this exceptional result was seen only in one private school. However, mathematics teachers still tend to prepare authentic teacher-made mathematics items at the basic taxonomic levels as in the year 2021. It has been determined that the curriculum change has not as yet provided a vision-enlighten "aha moment" with which they develop these skills. Understanding the taxonomy of items (e.g., Revised Bloom's Taxonomy, TIMSS Framework) and best practice strategies on how to prepare examination items for students' deeper learning may help middle school mathematics teachers formulate a wider range of items that not only stimulate factual knowledge and recall skills but also necessitate students to trigger the use of metacognitive knowledge and evaluating skills.

Consistent with the literature, teachers' questioning skills and practices were fairly consistent, as Gall (1970) suggests and as has been evidenced by other studies, and the quality of item preparation is a practice that gains value with teachers' experience rather than their beliefs. Studies (e.g., Hartell & Stimel, 2019; Delil & Özcan, 2019; Özcan & Delil, 2018) also propose that teachers can prepare well-crafted and HoT items as they gain experience on this topic. Hence, mathematics teachers need to progress and develop an identity beyond "teacher as a professional question marker" (Gall, 1970).

5.2. Association of Teachers' Teaching Method and Measurement-Evaluation Strategy Preferences between Some Variables

The findings of this study regarding mathematics teachers' teaching method and assessment preference changes after policy change are also in accordance with earlier observations, which shows that middle school mathematics teachers percieve no specific change in the foundations of the curriculum (i.e. philosophy, purpose, content) after the curriculum policy change. The teachers prefer doing activities that provide opportunities for student creativity, to change their in-class teaching method to make their students active while checking students' readiness. Considering instructional techniques, teachers tend to prefer using more constructivist approaches such as using concrete materials, group teaching methods, educational technologies during in-class teaching, and different questioning techniques. It was further observed that the teachers may have believed that after the curriculum policy change, they needed to make some changes in their teaching method and apply what they learned from the maths curriculum. Hence, similar to the literature, in the Turkish context, renewal of curricula does not necessarily translate into the renewal of teacher behaviors in the classroom, as agreed by Bümen, Çakar and Yıldız (2014), Öztürk (2012), and Yaşar (2012). Teacher preferences are associated with different networks and variables. For example, factors such as teacher beliefs, having positive beliefs about new reforms, openness to professional development, whether or not they set their professional plans (Zhang & Shen, 2012), taking part in the curriculum development process, and number of years of experience seem to affect the network of these relationships. Having inadequate knowledge about the relationship between curriculum and instruction may affect their vision about teaching method and assessment preferences, which concurs with Kerkez's (2018) study (see p. 52 for detail) which relied on the teachers' role in the curriculum development process in Turkey. The fact that they are not seen as an agent of change (e.g., Priestley, Biesta & Robinson, 2015) may be due to the fact that they may have not yet determined their professional plans which can develop more agency (Zhang & Shen, 2012). This is also in line with Omane's (2021) prediction study in which teachers' belief, professional development and knowledge on subject content and pedagogy, was highlighted as one of the predictors of teachers' instructional experiences. All in all, teacher' instructional experiences tend not to change where changes come through a top-down approach rather than bottom-up.

There are some important constructs in the literature which can affect or are related with teachers' teaching method instructional behaviors after a curriculum change. They are the role of belief, pedagogical beliefs (e.g., König, 2012; OECD, 2009; Priestley & Drew, 2016); willingness to change (Adnan, 2020); job satisfaction (Kayır & Toraman, 2021); gender (e.g., Elemadi, 2019; Konakman, 2017); age (Kayır & Toraman, 2021); educational level, school type, large class size (Westbook, 2013); branch, seniority year, faculty from which they graduated (Konakman, 2017); professional development (Zheng & Shen, 2012); and knowledge of content and pedagogy (Omane, 2021). Prior to data collection, this study hypothesized that there is no significant difference between teachers' teaching method and measurement-evaluation strategy preferences in total or subscales in terms of their gender, seniority year (i.e. professional year), educational level, and school type. Female teachers' adoption after curriculum change was found to be higher than male teachers. The teachers who have 15-years or more experience in the profession were found to have a higher teaching method preferences total score after curriculum change than those who have less than 15-years experience.

In terms of the gender difference in instructional choices, Adnan (2020) explored teacher's belief and willingness to change after a curriculum change and Elemadi (2019) studied students' preferences in science courses in Qatar and found that female students prefer more active teaching and learning experiences than male students. Kayır and Toraman's (2021) dependent variable with Turkish teachers was job satisfaction while the independent variables were gender, age, education level, school type. They applied a regression model to explore the amount of variance in teachers' perception regarding job satisfaction by four independent variables. However, they found that those variables did not significantly explain teachers' resistance after curriculum change. Similarly, König (2012) studied the role of beliefs in the practice of teaching as well as TALIS reports conducted by OECD (2009). They explored teachers' direct transmission belief and constructivist beliefs, and conducted a regression model to reveal to what extent it was explained by teachers' gender, branch, educational level etc. The report asserts that gender should be selected as covariate because female teachers tended to prefer active learning, constructivist teaching strategies, methods and techniques in their in-class teaching more than male teachers. Their pedagogical beliefs were found to explain their instructional preferences.

Konakman's (2017) research aimed to investigate whether there is a significant difference between teachers' instructional choice and their adoption of curriculum change in terms of gender, branch, seniority year, and graduate faculty. This study found gender, faculty type, and seniority were non-significant variables. However, the branch was found to be significant so that preschool teachers' scores were higher than branch teachers. On the other hand, Westbook

(2013) states that teaching belief and large class size are two variables that may be associated with teachers' instructional preferences. Kul and Çelik's (2017) study was based on a collective case study and their results reveal that pre-service teachers may hold constructivist views about teaching and learning. Their belief system shapes their behavior during teaching and has been affected by the current teacher education program and education system. The fact that their belief system may shape their behavior was correspondingly similar to the the inferences discovered in the current study. On the other hand, Turnova (2012) revealed results which are in contrast with the results of this study. Thompson (1984) conducted a case study with three secondary school mathematics teachers. He investigated how teachers' instructional behavior in teaching is explained by belief, view and preferences. He indicated a change in personal practice. Tanguay (2020) and Paehter (2003) investigated a gender difference between teachers' method preferences and produced findings compatible with ours. In the current study the re-explanation of students was discovered to be non-significant while van Loon et al. (2021) found that students' self-monitoring skills were more accurate in classrooms where teachers more often used child-centered instructional practices.

On the other hand, the measurement and evaluation are an assessment process in which structures not only are related to the curriculum process but also provide feedback to the teacher. The Turkish education system is an examination-oriented system. Therefore, it is not a feature of teachers' characteristics nor their fault that they are not able to be as flexible as they could be when teaching with their own measurement and evaluation strategy preferences. Unfortunately, Turkey is a land of examinations. It is known from various studies that teachers may not have knowledge and experience of the principles of writing examination items and that their measurement and evaluation preferences may be related to professional development courses rather than their beliefs. Since the expectations from teachers both in terms of instructional and assessment processes are high, they cannot sufficiently divide their performance in the field into authentic assessment types. Their first priority is to focus on teaching and presenting the curriculum.

Moreover, beliefs are not revelation. Teachers personally apply a teaching method in their classroom, and evaluate themselves and their students. They believe in the benefit and usability of whichever method they have succeeded in. Therefore, it is a natural expectation that they may tend to use the teaching methods they believe in in the classroom. It is also central to think of branch group meetings as a learning community. They should see group meetings not only as completing the obligatory paperwork, but also as a process in which they share and learn teaching methods from each other. After curriculum change, if teachers already know the scope and content of the program very well, try different teaching methods, and talk about their benefits with their colleagues, such a goal-oriented approach may lead to an increase in student success.

However, in the broad perspective of the research, it has been revealed that the factors that teachers consider when choosing measurement and evaluation strategies are not beliefs and changes in the program. Teachers first need to learn from an expert in order to use the different measurement-evaluation approaches in in-class examinations or during teaching. They believe in expert opinions on measurement and evaluation. They should follow more in-service seminars and training. Teachers who trust the experts of the Ministry should pay attention to this when someone from the province and district of the Ministry provides training. If the teachers got a grade above 90 points for 6 years, they would get 1 level of progress. It can be suggested that the same method should be integrated into teaching, since grading progress was motivating for teachers. Thus, teachers can develop their beliefs about not only diversifying their teaching methods but also developing measurement-evaluation strategies.

5.3. Middle School Students' Reflection on Metacognitive Skills and Affective Processes by Way of Different Item Types

The current study also searched for evidence of 5th grade students' reflection on their metacognitive skills (cognitive strategy and self-checking) and the affective process (effort and worry) levels of their responses to different item types. To clarify, the students who responded to open-ended items used metacognitive skills such as cognitive strategy more than those who responded to multiple-choice items. The results were consistent with those of O'Neil and Brown (1998), who revealed, more than twenty years ago, that there was a significant main effect for item format, with open-ended items inducing more use of cognitive strategy than multiple-choice items (Birgili, 2014; Loh & Lee, 2019; Öksüz & Güven-Demir, 2018; Rimbatmojo et al., 2017; Sole, 2018). This study corroborates that of O'Neil and Brown (1998), who indicate that multiple-choice questions yielded greater self-checking than open-ended problems. The fact that students had positive feelings toward multiple-choice items and negative feelings toward open-ended items is in line with previous studies (O'Neil & Brown, 1998). Generally, open-ended items induce greater amounts of worry than multiple-choice items do. Only in effort did the current study not show a format difference because the data collected from students was very limited and may therefore was insufficient for statistical analysis.

There are several other explanations for similar findings, such as the fact that students who took tests that included both multiple-choice questions and short answer questions engaged in more cognitively active behaviors than those who only took multiple-choice questions, and that these active behaviors improved performance on the final exam (Chick, 2016). Similarly, students who took exams that included both short answers and multiple-choice questions had more cognitively engaged behaviors than those who only took exams that included multiple-choice questions, and these behaviors boosted performance on the final exam.. Moreover, in contrast to O'Neil and Brown (1998), who found no significant gender differences in math performance on the multiple-choice items or the open-ended problems, the findings of this study revealed that girls were less likely to use the rereading sub-skill than boys. The results may differ from the related literature because Students with high levels of visual-spatial intelligence might not struggle with each part of metacognition, but students with low levels of visual-spatial intelligence could struggle with three metacognitive skills: cognitive tasks, self-knowledge and strategy knowledge.

Although the use of multiple-choice items in assessment metacognitive skills has been seen in a speculative way, this study may show a clear perspective to measure metacognition and its subskills by advantages of open-ended items (i.e., constructed response, free response) (e.g., da Silva Soares, et al., 2021; Frenken, 2021; O'Neill & Brown; Chick, 2013; Stillman, 2020; Vuorre & Metcalfe, 2021). We should not forget that in educational sciences, the preparation of quality items (e.g., Bassett, 2016; Dutke & Barenberg, 2015; Scully, 2017) is more critical to fostering higher-order thinking skills/metacognition rather than the type of items.

Finally, one of the most innovative results of this study is the use of eye-tracking as a validity tool for the Think Aloud process (Jarodzka et al., 2013). Scientific investigations on different biomarkers in education can explicate novel aspects of teaching-learning interactions and support evidence-based student practice. Brain studies have also gained momentum in the light of scientific studies and theories that give imperative to the whole development of students (i.e. model of the learner (Willingham, 2017)) despite teachers who still believe in the classical, traditional and outdated pedagogies where students should develop only behavioral skills. The main goal is for "agent" teachers to update their practices based on brain and deep learning results. This is also an interdisciplinary study harmonizing educational sciences and neuroscience and it is expected that it will become the first methodological design in the Turkish context to receive deep data from the student, not only with Think Aloud process but also with instant/live handwriting-based problem-solving, eye-tracking, and emotion measurements. Eye-tracking is a reliable option for research with 5th-grade students as it provides a real-time measurement of students' performance by providing them with different strategies expressed during the task. In particular, the Think Aloud process has been supported by eye-trackers. This finding is also reported by Jarodzka et al. (2013), Öztürk (2021), van Gog and Jarodzka (2013), van Gog and Scheiter (2010).

Da Silva Soares et al. (2021) specifies the area of interest (AoI) as the "problem statement", "figure", and all four "multiple-choice" and reveal that the most viewed regions focused on by students were the figures. On the contrary, in the current study, the most viewed AoI region focused on was "numbers". This contradictory result may be due to all items being solely mathematics MC even though the participants were 5th-grade (10-11-year-old) students.

One of the advantages of using eye-trackers with students is to provide students with a more natural and cognitively less demanding experience, as transforming their mathematical strategy, cognitive or metacognitive process, into a verbal report may be overwhelming for low achieving students, as asserted by da Silva Soares et al. (2020). The results on assessment of metacognition skills with diagnostic instruments are also supported by Desoete (2008) and show that metacognitive skillfulness when joined with intelligence accounts for between 52.9% and 76.5% of students' mathematics performance. The choice of diagnostic instruments such as eye-tracking determines the predicted percentage to a high degree. A profound discussion on how eye-tracking technologies can be effectively applied in the educational field is needed to clarify how teachers could benefit from it before, during, and after instruction. Curiously, all students expressed positive feelings toward the technologies used in the research process.

In addition, the outcome of students' ability to use rereading subskill was significantly predicted by their total time spent on the process, gaze shifts, and gender. The most frequent constructs on which research was conducted in the related literature were not only the rereading subskill but also strategy determination, controlling, planning, deep strategy usage, perseverance (i.e.

“give up” in the current study), mood, fixation, gaze shifts, feeling of difficulty, gender, pupil dilation, cognitive load, dwell time, amount of visual attention, time on task and self-explanation. Although Jarodzka and her colleagues (2017), who have been known for their innovative educational sciences versus eye-tracking studies, predicted the self-explanation skills of gaze shift and time on task in their study, this model was not significant in the current study because 5th-grade students were trying to solve ten different kinds of OE and MC item in 40 minutes and their exposure to Think Aloud at the same time may have affected their self-explanation level. The reason for this result is not clear but it may have something to do with self-checking, re-expression, finding error, process control, amount of effort etc. sub-skills of metacognition. Previous studies have highlighted the gender differences in eye-tracking studies as similar as Apaydin and Hossary (2017), Zhang (2018). In turn, it was emphasized by Azevedo and Aleved (2013) and Jian (2016) that the greatest challenge for students is to monitor and control basic cognitive and metacognitive processes during learning. Students' efforts to use metacognitive subskills naturally were appreciated and different observations have been noticed in some studies (e.g., Roderer & Roebers, 2014).

People make decisions based on results. It is hoped that this conclusion section is not the last word, but that it leaves the door open for new future research. The results of this deep and interdisciplinary study are a vision-enhancing beginning for cognitive scientists and engineers working with educators, teachers, evaluators, and junior educational scientists.

In the literature, studies analyzing teachers' authentic teacher-made examination items and studies investigating what kind of changes they made in their teaching method and measurement-evaluation process in the classroom after curriculum change were examined separately. This study aimed to examine mathematics teachers' authentic teacher-made examination items according to certain taxonomies and its results are generally consistent with the literature. The teachers tended to prefer and prepare mathematics items that measure lower-order levels in 2021, while they tended to change their teaching methods autonomously with their own will and belief. However, it is not prepared with the same consistency in the measurement, evaluation, and assessment. If items were prepared at high-order levels, it was observed that the 5th graders successfully could reflect different types of metacognitive and affective processes in their natural response to item processes. It is believed that this educational sciences study, which examines middle school mathematics teachers’ and 5th-grade students’ dimension in a holistic manner with a mixed-method research design, will shed light for researchers with its innovative aspect. The research questions were investigated through this multimodal study in which document analysis, semi-structured interviews, survey delivery, and deep student analyzes were handled. Metacognitive and affective processes of the students were examined from the perspective of an educational scientist. Accordingly, it has been noticed that the students can reflect their metacognitive skills and affective processes with great motivation and consistency in open-ended items. Consistent with most of the literature, it was observed that the students were able to use different metacognitive sub-skills in item types and tended to be nervous and worried about open-ended items.

At the last stage, a theoretical search was sought for the following question, which is believed to be innovative and useful for Turkey: “How can we design a deep data modal that can measure and evaluate the responses and emotions of middle school students to open-ended items? Which biomarkers do we also need to collect from the students?” Persistent and long-term studies are required to seek answers to these research questions. In the light of real data collected and analyzed from human nature (i.e. students), it has been realized that deep model designs require a very long time (at least 5 years), and the designs that include both researchers and academics from cognitive scientists and computer engineers should be made. The possibility of designing, researching, as well as applying, intelligent deep data models adapted to the Turkish language developed and designed for Turkish children can only be done by being on the same page. This is one of the future predictions.

Implications for Future Research. Research implications suggest how the findings of this study may be important for educational policy, theory, and subsequent research. This study was based on mixed-methods research including five phases. First of all, document analysis and interviews were conducted with a sample of pilot mathematics teachers, and then data were collected from a wider teacher sample through the survey study. In the brain dynamics laboratory, where I studied with 5th-grade students, both qualitative and quantitative data were collected, aggregated, synthesized and interpreted. Therefore, due to the nature of the study and the flow of the research, the qualitative approach was dominant. Yet, quantitative multimodal data supported the research questions. Finally, I illustrated a theoretical and original deep data analysis model. As a multi-layered and very comprehensive study was designed, I encountered several difficulties while conducting it. I needed in-class observations from some middle school mathematics teachers. For this reason, I recommended making observations of teachers while determining their teaching methods and measurement-evaluation preferences in future studies. Teachers who cannot allocate time for individual meetings can be invited to focus group meetings as another data collection methods. In this study, some mathematics teachers who did not want to participate in the interview showed up and said, "This research is of no use to me. I don't think it will benefit my students", "There is no project in this study and I have no gain." They used sentences such as "I did not want to help". As a researcher, please try to arrange a meeting with another teacher without losing your motivation. School principals and teachers should better understand the importance of research and more action should be taken in this direction. You can motivate participating teachers with small gifts or incentives for participating in the research.

Future researchers can use the TMMESP questionnaire, which I developed for theoretical and originally for mathematics teachers, or they can use the three-dimensional scale of Sadler-Smith and Riding (1999)'s scale, which measures teaching method and assessment-evaluation preferences. It is a matter of curiosity how the results will change when other scales in the related literature that propose to measure similar dimensions would be used. The literature shows that the adaptation of teachers and young teachers with less than 15-years experience to curriculum change can be accelerated and they are open to more changes in their classroom practices such as teaching methods. Although this study revealed that senior teachers (with more than 15-years experience) tend to have more autonomy in pre and post policy change, understanding the curriculum development process in Turkey from the perspectives and experiences of those (with less than 15-years experience) remains a matter for the future. It is still a matter of interest for the literature to determine what kind of adaptations novice middle school mathematics teachers (being within the first 5 years of their profession) improvised their in-class practices during the teaching and learning process after the curriculum change.

Face recognition, emotion recognition, screen recording, self-expression sessions, and linguistic analysis of discourses can be used innovatively in experimental stages where multimodal data collected from students are used for artificial intelligence/expert systems design. Thus, we can strengthen the quality of deep data and collect more student data. In particular, the deep data obtained from such laboratory studies with innovative technological tools are useful for the progression of students' learning and redesigning the learning processes. The kind of multimodal data sheds light on *learning analytics* studies and is recommended for future studies. The research laboratory process can be repeated with other metacognitive skills (e.g., awareness, monitoring) and the affective process (e.g., anxiety, mood) variables that are not covered in the context of this study. In addition, students' self-regulation sub-skills can also be selected as dependent variables during physiological and biological measurements. It should be noted that there might be tradeoffs in neuroeducation studies when changing the methods or selecting the study group according to certain predefined criteria. It is an important principle not to impair generalizability. Knowing this, with the selection of more communicative students, the difference in re-explanation sub-skills between them and others can be examined more clearly.

Mixed-methods studies have various sub-designs according to different researchers (i.e. Cresswell, Kuckartz, Onwuegbuzie Plano-Clark). Researchers first ask, “do you believe that you can best answer your research question(s) through the use of mixed-methods research?” and “do you believe that mixed-methods research will offer you the best design for the amount and kind of evidence that you hope to obtain as you conduct your research study?” etc. They should answer the research questions then carefully follow these steps: selection of sampling design [random/purposive]; sampling arrangement, covering “the mixing dimension [partial/fully], time dimension [concurrent/sequential], emphasis dimension [dominant/equal status], and relationship among/between samples; sample size; and a number of sampling units [e.g., of people, cases, words, texts, observations, activities, or any other object of study] in order to obtain meta-inferences” consistent with the future study's goals (Corrigan & Onwuegbuzie, 2020; Onwuegbuzie & Teddlie, 2003 p. 787). Researchers who might work with similar research questions may need more teacher and student groups so they may deal with complex mixed methods designs and may need to receive more expert opinions when managing and combining qualitative and quantitative processes. Sometimes, the qualitative and quantitative phases may produce conflicting results. Finally, most mixed-method studies in the world, especially in Turkey, have problems integrating quantitative and qualitative results (Toraman, 2021). It can be suggested that researchers be careful in the integration process.

Implications for Practice. Research implications also suggest how the findings of this study may be important for educational practice. In the survey phase, which is the second step of this study, data was collected from 350 teachers by visiting various districts of Istanbul one by one. However, considering that there are over 1 million students and teachers in Istanbul, the teacher group is thought to be small. Therefore, the results of this study can only be generalized to a population similar to the study group. On the other hand, the study has shown that while teachers’ teaching method preferences mostly depend on their beliefs, their measurement-evaluation preferences depend on whether they have received professional education or not. The majority of the mathematics teachers in this study did not remember their in-service training periods, and those who did entered limited data. It would be beneficial for the literature to repeat the first dimension of this study with mathematics teachers who receive in-service training under the MoNE or Measurement-Evaluation Academies.

Belief is a skill that requires practice. Deconstructing faith is not easy and is a process requiring intensive practice. Teachers' practice of a teaching method and seeing its effect on their students' success/performance/achievement can make that teaching method credible in the eyes of the teachers. Since the first results of this study are for middle school mathematics teachers, it is recommended that mathematics teachers in particular try different teaching methods and try classroom practices for constructivist approaches. Nevertheless, Özeren and Akpunar (2019) highlight that "modern constructivist education programs in Turkey are practiced in a traditional sense". That is to say, educational policies remain insufficient in renewing classroom practices and using teaching strategies, methods and techniques for educational philosophy because curriculum change studies, regrettably, continue in a top-down manner in Turkey. The study also highlights that it is the teachers with less than 15-years seniority who have the most difficulty in adapting to curriculum change. When the draft curricula are presented to the view of teachers, educators, parents and students, even before the publicised, those teachers should be involved in the process of draft curriculum making as future change agents. Thus, it might be possible to shift from a top-down approch, which is still developed with the traditional approach, rather than a bottom-up one.

Another interesting result of the study is that the ability of mathematics teachers to use innovative approaches in their measurement-evaluation preferences depends on the level of education they received as professional development. In particular, there was no difference in the subdimensions of measurement-evaluation preferences. It was discovered that mathematics teachers still use objective-testing in preparing examination items. Both data sets support each other. In Turkey, it is recommended that mathematics teachers in particular should take courses that focus on measurement and evaluation, such as different item types, preparing innovative items, and making tables of specifications in preparing items, and improve themselves in this regard. Even in their authentic examinations, there is a lack of knowledge and skill in preparing items that would measure students' higher-order cognitive skills. It has been determined that innovative items, computerized adaptive tests, and intelligent tutoring systems, which are used extensively around the world, need to make more preparations for developing the types of items, preparing such items, answering them, and even preparing their students for the future and the professions of the future. The results of the research have shown that when qualified and higher-order items are prepared, students can use their awareness, explanations, and metacognitive skills more.

In addition, it is important for teachers to know the technological systems that can measure natural human capabilities in a useful and multimodal way. It is good to increase their awareness of using such devices in the classroom. They may be aware of different metacognitive skills and emotional processes that their students can experience/display during different types of items. Their awareness of interpreting student behaviors may increase, they can interpret the differences between the students who do not show worry and who have high performance and the students who have mild worry but high performance, and can internalize such principles. For example, teachers may realize the importance of using different item types and items at different cognitive levels while solving questions in the measurement-evaluation process in the classroom. They may teach their students how to use their worry in a positive way during the assessment, and to internalize what kind of information the students should focus on. When responding to an item, students may understand how many ways they could solve the problem and where they should focus on the items. It is expected that students may be aware of the fact that there is not a single correct solution to a problem and that questions can be solved in more than one way and via thinking skills, and teachers should convey this to their students. By interpreting the results of this study, teachers may understand 5th-grade students' thinking process, realize their psychology, be aware that they can solve the questions in different ways, and develop insight. In particular, they are expected to make progress in integrating these results into their teaching methods.

The number of in-class observations should be increased in order to determine more decisively the changes made by teachers in classroom teaching with their own autonomy after curriculum change. Observing the teachers while they are teaching in their natural flow will be good for a more reliable evaluation of the results. However, teachers who have reservations about being observed should meet with the researcher and get information about this subject and answer their questions. Researchers should also guarantee that these observations will be in the natural environment, not as an inspection, and will be for informational purposes only. Otherwise, teachers may be concerned that the researcher could report to the Ministry of National Education. It should be emphasized that observations are the goal of contributing to scientific research.

Recommendations for further practice. Educators should help students to acquire metacognitive knowledge and skills while responding to open-ended items and implement self-checking strategies by using a variety of mathematics problems. To do so, educators should point out some task-based problem-solving processes as well as encouraging students to check their understanding and use metacognitive skills such as self-explanation/re-explanation and self-checking/finding errors.

In the light of this study, it is suggested that a series of precautions should be taken and put into practice in order for mathematics teachers to prepare their authentic teacher-made items at a level that can measure students' higher order thinking (HoT) skills by complying with scientific research and reports. It is thought that it is good to examine the contemporary "journal of papers", examine the official letter from the ministry of national education by the inspectors, and take suggestions, if any, while preparing the teachers' in-class examinations.

However, they can receive in-service education on how to prepare for examinations that are able to measure HoT during the summer holidays. With the cooperation of measurement and evaluation centers opened in 81 provinces in Turkey affiliated to the Ministry of National Education, education on item quality and in-class authentic examination preparation should be added to the Education Information Network (EBA) platform. Teachers should be able to receive this training from where they are at their own pace with an active learning approach.

REFERENCES

ABIDE (2015). *Akademik becerilerin izlenmesi ve değerlendirilmesi [Monitoring and evaluating academic skills].* http://abide.meb.gov.tr/

ABIDE (2016). *8. sınıflar raporu. [8th Grade Report].* MoNE General Directorate of Measurement, Evaluation and Examination Services.

Akgün, H. (2018). *The effect of item format on the choice of reading and test-taking strategies* [Unpublished master's thesis]. Boğaziçi University.

Aldım, Ü. F. (2010). İlköğretim 7. sınıflarında uygulanan sbs (seviye belirleme sınavı) ingilizce sorularının bazı değişkenlere göre incelenmesi. [Unpublished master's thesis]. Fırat University.

Alemdağ, E., & Çağıltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125,* 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Allahverdi, N. (2002). *Uzman sistemler bir yapay zeka uygulaması [Expert systems an artificial intelligence application].* Atlas Publication.

All-Party Parliamentary Group for Education [APPG] (2017). *How well do schools prepare children for the future?* http://www.educationappg.org.uk/wp-content/uploads/2017/04/Preparing-for-the-future-inquiry-report.pdf

Altun, M., & Akkaya, R. (2014). Mathematics teachers' comments on pisa math questions and our country's students' low achievement levels [Matematik öğretmenlerinin pisa matematik soruları ve ülkemiz öğrencilerinin düşük başarı düzeyleri üzerine yorumları]. *Hacettepe University Journal of Education*, *29*(1), 19–34.

Alvunger, D. (2018). Teachers' curriculum agency in teaching a standards-based curriculum. *The Curriculum Journal, 29*(4), 479–498. https://doi.org/10.1080/09585176.2018.1486721

Amer, A. (2006). Reflections on Bloom's revised taxonomy. *Electronic Journal of Research in Educational Psychology, 4*(1), 213–230. https://doi.org/10.25115/ejrep.v4i8.1217

American Association of Colleges for Teacher Education (AACTE) (2010). *The clinical preparation of teachers: A policy brief.* AACTE Publication.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (2015). *Standards for educational and psychological testing*. American Psychological Association.

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Educational Evaluation, 31*(2-3), 102–113. https://doi.org/10.1016/j.stueduc.2005.05.004

Anderson, L., & Krathwohl, D.A. (2001). *Taxonomy for learning, teaching and assessment. A revision of Bloom's taxonomy of educational objectives*. Longman.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., et al. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's educational objectives (abridged edition)*. Longman.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Andrews, P., Ryve, A., Hemmi, K., & Sayers, J. (2014). PISA, TIMSS and Finnish mathematics teaching: An enigma in search of an explanation. *Educational Studies in Mathematics, 87*(1), 7–26. https://doi.org/10.1007/s10649-014-9545-3

Ankara University (2017). Report on views toward 2017 draft curriculum. (Ebru Oğuz, E-mail, April 22, 2019).

Anne Çocuk Eğitim Vakfı [Mother Child Education Foundation] [AÇEV] (2017, February). *AÇEV, taslak müfredat hakkındaki görüşlerini açıkladı [AÇEV explained their views on draft curriculum]* http://www.hayatimizegitim.com/genel/acev-taslak-mufredat-hakkindaki-goruslerini-acikladi-h14282e.html

Anteneh, M. M., & Silesh, B. D. (2019). Assessment practices and factors for the disparity between students' academic scores at teacher-made and regional exams: The case of Bench Maji zone grade 8 students. *Educational Research and Reviews, 14*(1), 1–24. https://doi.org/10.5897/ERR2018.3581

Apaydin, M., & Hossary, M. (2017). Achieving metacognition through cognitive strategy instruction, *International Journal of Educational Management, 31*(6), 696–717. https:// doi.org/10.1108/IJEM-05-2016-0130

Apple, M. W. (2018) Critical curriculum studies and the concrete problems of curriculum policy and practice. *Journal of Curriculum Studies, 50*(6), 685–690. https://doi.org/10.1080/00220272.2018.1537373

Ariasi, N., & Mason, L. (2011). Uncovering the effect of text structure in learning from a science text: An eye-tracking study. *Instructional Science*, *39*(5), 581–601. https://doi.org/10.1007/s11251-010-9142-5

Arieli-Attali, M., Ou, L., & Simmering, V. R. (2019). Understanding test takers' choices in a self-adapted test: A hidden markov modeling of process data. *Frontiers in Psychology, 10*(83), 1–15. https://dio.org/ 10.3389/fpsyg.2019.00083

Aryadoust, V. (2019). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: an eye-tracking study. *Computer Assisted Language Learning, 33*(5–6), 510–537. https://doi.org/10.1080/09588221.2019.1574267

Atila, M. E. (2012). Fen ve teknoloji dersi öğretim programındaki yapılandırmacılığa dayalı öğelerin öğretmenler tarafından algılanışı ve uygulanışı [Teachers' perception and application of the elements based on constructivism in science and technology curriculum]. [Unpublished doctoral dissertation]. Atatürk University.

Awang-Hashim, R., O'Neil, H. F., & Hocevar, D. (2010). Ethnicity, effort, self-efficacy, worry, and statistics achievement in Malaysia: A construct validation of the state-trait motivation model, educational assessment. *Educational Assessment*, *8*(4), 341–364. https://doi.org/10.1207/S15326977EA0804_3

Ayan, C. (2018). Bilişsel tanı modelinde geleneksel ve bilgisayarlı sınıflamalı test uygulamalarının psikometrik özelliklerinin karşılaştırılması [Comparing the psychometric features of traditional and computerized adaptive classification test in the cognitive diagnostic model]. [Unpublished doctoral dissertation]. Ankara University.

Ayan, C., & Çıkrıkçı, R. N. (2021). A study on formative assessment with cognitive diagnostic model application for math proficiency. *The International Journal of Assessment and Evaluation, 29*(1), 37–55. https://doi.org/10.18848/2327-7920/cgp/v29i01/37-55

Ayhan, E. (2019). Determining the levels of reading texts in teaching turkish as a foreign language: eye tracking findings during reading [Unpublished doctoral dissertation]. Hacettepe University.

Azevedo, R. (2002). Beyond intelligent tutoring systems: Using computers as METAcognitive tools to enhance learning? *Instructional Science 30*, 31–45. https://doi.org/10.1023/A:1013592216234

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*(1), 87–95. https://doi.org/10.1007/s11409-009-9035-7

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: An overview of current interdisciplinary research. *Springer International Handbooks of Education*, 1–16. doi:10.1007/978-1-4419-5546-3_1

Azevedo, R., & Gašević, D. (2019). Analyzing Multimodal Multichannel Data about Self-Regulated Learning with Advanced Learning Technologies: Issues and Challenges. *Computers in Human Behavior, 96*, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

Azevedo, R., Mudrick, N., Taub, M., & Wortha, F. (2017). Coupling between metacognition and emotions during stem learning with advanced learning technologies: A critical analysis, implications for future research, and design of learning systems. *Teachers College Record, 119*(3), 1–18, Article e1143699.

Bağcan Büyükturan, E. & Çıkrıkçı Demirtaslı, N. (2013). Comparing the psychometric characteristics of multiple choice tests and structural communication grids. *Ankara University Journal of Faculty of Educational Sciences (JFES), 46*(1), 395–416. https://doi.org/10.1501/egifak_0000001289

Bannert, M., & Mengelkamp, C. (2007). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*(1), 39–58. https://doi.org/10.1007/s11409-007-9009-6

Barootchi, N., & Keshavarz, M. H. (2002). Assessment of achievement through portfolios and teacher-made tests. *Educational Research*, *44*(3), 279–288. https://doi.org/10.1080/00131880210135313

Bascia, N., Carr-Harris, S., Fine-Meyer, R., & Zurzolo, C. (2014). Teachers, curriculum innovation, and policy formation. *Curriculum Inquiry, 44*(2), 228–248. https://doi.org/10.1111/curi.12044

Bassett, M. H. (2016). Teaching critical thinking without (much) writing: Multiple-choice and metacognition. *Teaching Theology & Religion,* 19(1), 20–40. Wiley. https://doi.org/10.1111/teth.12318

Başokçu, O. T. (2011). Bağıl ve Mutlak Değerlendirme ile DINA Modele Göre Yapılan Sınıflamaların Geçerliğinin Karşılaştırılması [Comparison of the Validity of Classifications by Criterion- Referenced and Normative Assessments with that of DINA Model]. [Unpublished doctoral dissertation]. Hacettepe University

Bayraktar, D. (2014). *Investigation of the effect of multimedia designed according to different types of attention on recovery performance in condition of cognitive characteristics by eye-tracking method* [Unpublished doctoral dissertation]. Marmara University.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1/2), 1–21. https://doi.org/10.1023/A:1007051109754

Berberoglu, G. (2009). CİTO Türkiye öğrenci izleme sistemi (ÖİS) öğrenci sosyal gelişim programı'na (ÖSGP) ilişkin ön bulgular. [CITO Turkey student follow- up system (OIS) pre-findings about student social development program.] *CITO Education: Theory and Practice Journal*, November-December Issue, 32–42.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. [Examination of student's success according to years, school type and region: Analysis of OSS and PISA.] *Educational Sciences and Practice, 22* (4), 21–35.

Besold, T. R., & Kühnberger, K. U. (2014). Applying AI for modeling and understanding analogy-based classroom teaching tools and techniques. In: Proceedings of the 37th annual German conference on AI, KI. Advances in artificial intelligence. [Lecture notes in computer science], 8736, 55–61, Springer.

Biesta, G., Priestley, M., & Robinson, S. (2017). Talking about education: Exploring the significance of teachers' talk for teacher agency. Journal of *Curriculum Studies*, *49*(1), 38–54.

Biesta, G. J. J., & Tedder, M. (2007). Agency and learning in the lifecourse: Towards an ecological perspective. *Studies in the Education of Adults, 39*, 132–149. https://doi.org/10.1080/00220272.2016.1205143

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.

Bilgeç, İ. (2016). Açık uçlu soruların kullanıldığı matematik sınavlarının ölçme değerlendirme perspektifinden incelenmesi [Inspection of mathematic exams with open ended questions from the perspective of measurement and evaluation]. [Unpublished master thesis]. Marmara University.

Binder M. D., Hirokawa N., Windhorst U. (eds) (2009). Gaze shift. *Encyclopedia of Neuroscience*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-29678-2_1941

Birgili, B. (2014). Open-ended questions as an alternative to multiple-choice: Dilemma in Turkish examination system. [Unpublished master thesis]. Middle East Technical University.

Birgili, B., Akar, H., & Çakar, T. (2021, April 9-12). An Ecological Approach to Identifying Mathematics Teachers' Response to Curriculum Change in Their Achievement Tests. Paper presented at American Educational Research Association Meeting [AERA], 2021 Annual Meeting, April 9-12, San Francisco, USA, 2021.

Birgili, B., & Kiraz, E. (2017). A Dilemma in the Turkish Examination System: Open-Ended or Multiple-Choice? In M. Gammone, M. A. Icbay & H. Arslan (Eds.), *Recent Developments in Education*. (1st ed., pp. 72-80). Bialystok.

Björn, P. M., Äikäs, A., Hakkarainen, A., Kyttälä, M., & Fuchs, L. S. (2019). Accelerating mathematics word problem-solving performance and efficacy with think-aloud strategies. *South African Journal of Childhood Education*, *9*(1). https://doi.org/10.4102/sajce.v9i1.716

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.

Bloom, B. S. (1964). *Stability and change in human characteristics*. John Wiley & Sons.

Bloom, B. S. (1968). Learning for mastery. *Evaluation comments*, *1*(2) (pp. 1-12). Los Angeles, California: UCLA-Center for Study of Evaluation and Instructional Programs.

Bloom, B. S., & Rakow, H. A. (1969). Higher mental processes. In R. C. Ebel (Ed.), *Encyclopedia of educational research* (pp. 594-601). MacMillan Publishing Co., Inc.

Bloom, B. S. (1971). *Mastery learning*. In J. H. Block (Ed.), Mastery learning: Theory and practice (pp. 47–63). Holt, Rinehart and Winston.

Bloom, B. S. (1974). *Time and learning*. *American Psychologist, 29*(9), 682–688.

Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.

Bloom, B. S. (Ed.) (1985). *Developing talent in young people*. Ballantine Books.

Bloom, B. S. (1987). A response to Slavin's Mastery Learning reconsideration. *Review of Educational Research*, 57(4), 507–508.

Boardman, A. G., & Woodruff, A. L. (2004). Teacher change and ''high-stakes'' assessment: what happens to professional development? *Teaching and Teacher Education, 20*(1), 545–557. https://doi.org/10.1016/j.tate.2004.06.001

Bobbitt, F. (1918). *The Curriculum*. Riverside.

Bolat, Y. (2016). Kavram Temelli Disiplinler Arasi Yaklaşıma Göre Tasarlanan Ünitenin Otantik Değerlendirmesine Yönelik Bir Eylem Araştırması [An Action Research on the Authentic Assessment of a Unit Designed According to Concept-Based Interdisciplinary Approach]. [Unpublished doctoral dissertation]. Çukurova University.

Boesen, J. (2006). Assessing mathematical creativity: Comparing national and teacher-made tests, explaining differences and examining impact. [Unpublished doctoral dissertation]. Umea University.

Bogazici University (2017). Report on the 2017 Turkish educational system curricular change. http://www.fed.boun.edu.tr/form_files/Mufredat_degisikligi_fakulte_gorusleri_2017.pdf

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats, *The Journal of Experimental Education*, *81*(3), 409–428. https://doi.org/10.1080/00220973.2012.727886

Briggs, D. C., & Circi, R. (2017). Challenges to the Use of Artificial Neural Networks for Diagnostic Classifications with Student Test Data. *International Journal of Testing, 17*(4), 302–321. https://doi.org/10.1080/15305058.2017.1297816

Broekkamp, H., Hout-Wolters, B. H. A. M., Bergh, H., & Rijlaarsdam, G. (2004). Students' expectations about the processing demands of teacher-made tests. *Studies in Educational Evaluation, 30*(1), 281–304. https://doi.org/10.1016/j.stueduc.2004.11.002

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

BTTO (2017). Yapay Zeka eğitimde çığır açacak [Artificial Intelligence will break ground in education]. https://btto.yasar.edu.tr/yapay-zeka-egitimde-cigir-acacak/

Budak, B. (2017). Yapay Zekanın bugünü ve geleceğine DARPA gözüyle bakış [Today and Future of Artificial Intelligence from the View of DARPA]. https://webrazzi.com/2017/04/11/yapay-zekanin-bugunu-gelecegine-darpha-gozuyle-bakis/

Bullens, D. (2002). Authentic assessment: Change for the future. [Unpublished Master Thesis]. Saint Xavier University.

Burfitt, J. (2019). Cognitive interviews for reviewing multiple-choice items in mathematics. *Issues in Educational Research, 29*(2), 346–362.

Burul, C. (2018). Öğretmenlerin eğitim programı tasarım yaklaşımı tercihlerinin öğretim programına bağlılıklarıyla olan ilişkisinin incelenmesi [Investigation of the relationships between curriculum design approach preferences of teachers and their curriculum fidelity]. [Unpublished Master Thesis]. Balıkesir University.

Bümen, N. T., Çakar, E. & Yıldız, D. G. (2014). Curriculum fidelity and factors affecting fidelity in the Turkish context. *Educational Sciences: Theory & Practice, 14*(1), 219–228. https://doi.org/10.12738/estp.2014.1.2020

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*, 294–320.

Carless, D., & Zhou, J. (2015). Starting small in assessment change: Short in-class written responses. *Assessment & Evaluation in Higher Education, 41*(7), 1114–1127. https://doi.org/10.1080/02602938.2015.1068272

Casalino, G., Cafarelli, B., del Gobbo, E., Fontanella, L., Grilli, L., Guarino, A., ... & Taibi, D. (2021). Framing automatic grading techniques for open-ended questionnaires responses. A short survey. In CEUR Workshop Proceedings. In teleXbe—*Technology Enhanced Learning Environments for Blended Education*.

Catrysse, L., Gijbels, D., Donche, V., Maeyer, S. D., Lesterhuis, M., & Van den Bossche, P. (2018). How are learning strategies reflected in the eyes? Combining results from self-reports and eye-tracking. *British Journal of Educational Psychology, 88*, 118–137.

Cheong, Y.-Y. (2018). Assessment in Singapore: Perspectives for classroom practice. *Assessment in Education: Principles, Policy & Practice, 25*(6), 678–682. https://doi.org/10.1080/0969594X.2017.1309354

Chick, N. (2013). Metacognition. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/metacognition/.

Ching-En, C. (2018). Metacognitive experience modeling using eye-tracking. *Artificial Intelligence in Education,* 503–507. https://doi.org/10.1007/978-3-319-93846-2_94

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge.

Commission of the European Communities [CEC]. (2008). Improving competences for the 21st century: An agenda for European cooperation on schools. Communication from the Commission to the European Parliament, the council, the European Economic and Social Committee and the Committee of the region, Brussels, 3 July 2008.

Conati, C. (2016). Commentary on: "toward computer-based support of metacognitive skills: A computational framework to coach self-explanation". *International Journal Artificial Intelligence in Education, 26*(1), 183–192. https://doi.org/10.1007/s40593-015-0074-8

Corrigan, J. A., & Onwuegbuzie, A. J. (2020). Toward a meta-framework for conducting mixed methods representation analyses to optimize meta-inferences. *The Qualitative Report*, *25*(3), 785–812. https://doi.org/10.46743/2160-3715/2020.3579

Coşkun, A. (2019). Investigation of classroom management skills by using eye-tracking technology [Unpublished master's thesis]. Middle East Technical University.

Creswell, J. W. (2012). *Educational research Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Creswell, J. W. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches.* Apple Books.

Creswell, J. W. & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Creswell, J. W., Plano-Clark, V., Gutmann, M., & Hanson, W. (2003). Advanced mixed methods research designs. In A. Tashakkori and C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Holt, Rinehart and Winston.

Çağlar, M., & Kılıç, A. (2019). Merkezi sınav ve öğretmen yapımı sınavların bazı değişkenler açısından incelenmesi: ortaöğretime geçiş sınavı örneği [Investigation of the central examination and teacher-made exams for some variables: the transition to secondary education example]. Journal of *Bolu Abant İzzet Baysal University Faculty of Education, 19*(4), 1288–1305. https://doi.org/10.17240/aibuefd.2020..-416761

Çelik, S., & Kasapoglu, H. (2014). Implementing the recent curricular changes to English language instruction in Turkey: Opinions and concerns of elementary school administrators. *South African Journal of Education, 34*(2), 1–14. https://doi.org/10.15700/201412071144

Çetin, A., & Ünsal, S. (2018). Merkezi sınavların öğretmenler üzerinde sosyal, psikolojik etkisi ve öğretmenlerin öğretim programı uygulamalarına yansıması [Social, psychological effects of central examinations on teachers and their reflections on teachers' curriculum implementations]. *Hacettepe University Journal of Education*, 1–20. https://doi.org/10.16986/huje.2018040672

Çevik, E. (2009). İlköğretim II. kademe sosyal bilgiler dersi öğretmenlerinin yazılı sınav soruları ile seviye belirleme sınavı sorularının programa uygunluğunun incelenmesi. [Unpublished master thesis]. Niğde University.

Çıkrıkçı-Demirtaşlı, N. (2010). Açık uçlu soru formatı ve öğrenci izleme sistemi (ÖİS) akademik gelişimi izleme ve değerlendirme (AGİD) modülündeki kullanımı. [Open-ended question fotedrmat and student monitoring system (OIS) academic development monitoring and evaluation] *CITO Education: Theory and Practice Journal*, April-June Issue, 21–30.

Çıkrıkçı-Demirtaşlı N. (2017). Öğrenme, öğretim ve değerlendirme arasındaki ilişkiler [The relation between learning, teaching and evaluation]. N. Çıkrıkçı-Demirtaşlı (Ed.). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in Education]* (s. 4-29).

Çınar, A., Ince, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies, 25*(5), 3821–3844.

Çolakoğlu, M. H. (2018). Teachers' views and recommendations about PISA 2015 results. *Journal of Research in Informal Environments, 3*(1), 46–66. https://doi.org/10.1007/s10639-020-10128-0

Dağlı, M. (2014). Investigation of the usability of z-books prepared within the scope of FATİH educational project with eye-following and retrospective thinking-aloud techniques. [Unpublished master’s thesis]. Marmara University.

Damayanti, R., Sunardi, Yuliati, N., Karimah, R., & Albab, A. U. (2020). Students’ metacognitive ability in solving quadrilateral problem based on adversity quotient. *Journal of Physics: Conference Series*, 1538, 012077. https://doi.org/10.1088/1742-6596/1538/1/012077

Darling-Hammond, L. (1999). *Teaching as the Learning Profession: Handbook of Policy and Practice*. Jossey-Bass Education Series.

Darling-Hammond, L. (2006). *Powerful teacher education: Lessons from exemplary programs*. Jossey-Bass: A Wiley Imprint.

Darling-Hammond, L. (2012). Teacher preparation and development in the United States. In Darling-Hammond, L. and Lieberman, A. (Eds.). *Teacher education around the world: Changing policies and practices*.: Routledge Taylor and Francis.

Darling-Hammond, L., & Lieberman, A. (Eds.) (2012). *Teacher education around the world: Changing policies and practices*. Routledge Taylor and Francis.

Darling-Hammond L. (2017). *Teacher education around the world: What can we learn from international practice?, European Journal of Teacher Education, 40*(3), 291–309. https://doi.org/10.1080/02619768.2017.1315399

Darling-Hammond, L., & Oaks, J. (2019). *Preparing teachers for deeper learning*. Harvard Education Press.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5), 523–545.

da Silva Soares, R., Jr, Lukasova, K., Carthery-Goulart, M. T., & Sato, J. R. (2021). Student's perspective and teachers' metacognition: Applications of eye-tracking in education and scientific research in schools. *Frontiers in Psychology, 12,* 673615. https://doi.org/10.3389/fpsyg.2021.673615

Deardorff, D. K. (2007). Principles of international education assessment. *IIENetworker* (Spring 2007), 51–52.

de la Torre, J. (2012). Application of the DINA Model Framework to Enhance Assessment and Learning. In M. Mok (Ed.), *Self-directed learning oriented assessments in the Asia- Pacific* (pp. 92-110). Springer.

de la Torre, J. & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa, 20*(1), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

Delil, A., & Ozcan, B. N. (2019). How 8th graders are assessed through tests by mathematics teachers? *International Journal of Educational Methodology*, *5*(3), 479–488. https://doi.org/10.12973/ijem.5.3.479

Deluca, V W., & Lori, N. (2013). Developing students' metacognitive skills in a data-rich environment. *Journal of STEM Education, 14*(1), 45–55.

Demir, M. D. (2004). Bir eğitim yazılımının temelini oluşturan yapay zeka program geliştirme [Developing an artificial intelligence based instructional software]. [Unpublished master thesis]. Çukurova University.

Demirel, Ö. (2012). *Eğitimde Program Geliştirme: Kuramdan Uygulamaya [Curriculum Development in Education: From Theory to Practice].* Pegem Academy.

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: how you test is what you get. *Metacognition and Learning, 3,* 189–206. https://doi.org/10.1007/s11409-008-9026-0

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

Dimitrov, D. M. (1999). Gender differences in science achievement: Differential effect of ability, response format, and strands of learning outcomes. *School Science and Mathematics, 99*(8), 445–450. https://doi.org/10.1111/j.1949-8594.1999.tb17507.x

Dixon, M. A., & Cunningham, G. B. (2009). Data aggregation in multilevel analysis: A review of conceptual and statistical issues. *Measurement in Physical Education and Exercise Science*, *10*(2), 85–107. https://doi.org/10.1207/s15327841mpee1002_2

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

Doll, R. (1996). *Curriculum improvement: Decision making and process* (9th ed.). Allyn and Bacon.

Donoghue, G. M., & Horvath, J. C. (2022). Neuroeducation: A brief history of an emerging science. In *Encyclopedia of Behavioral Neuroscience*, (2nd ed.) (pp. 632–637). Elsevier. https://doi.org/10.1016/b978-0-12-819641-0.00077-3

Drew, V. & Mackie, L. (2011). Extending the constructs of active learning: Implications for teachers' pedagogy and practice. *Curriculum Journal, 22*(4), 451–467. https://doi.org/10.1080/09585176.2011.627204

Driscoll, J. R., Rajala, D. A., Shaffer, W. H., & Thomas, D. W. (1991). The operation and performance of an artificially intelligent keywording system. *Information Processing and Management, 27*(1), 43–54. https://doi.org/10.1016/0306-4573(91)90030-P

D'Souza, S. M., & Wood, L. N. (2003). Secondary students' resistance toward incorporating computer technology into mathematics learning. *Mathematics & Computer Education, 37*(3), 284–295.

Dunstone, T., & Yager, N. (Eds.). (2009). *Biometric System and Data Analysis*. Springer US. https://doi.org/10.1007/978-0-387-77627-9

Duran, E., & Tufan, B. S. (2017). The effect of open-ended questions and multiple choice questions on comprehension. *International Journal of Languages' Education and Teaching, 5*(1), 242–254. https://doi.org/10.18298/ijlet.1676

Dutke, S., & Barenberg, J. (2015). Easy and informative: Using confidence-weighted true–false items for knowledge tests in psychology courses. In *Psychology Learning & Teaching*, *14*(3), pp. 250–259). SAGE Publications. https://doi.org/10.1177/1475725715605627

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: knowing the test format enhanced metacognitive monitoring. *Metacognition Learning, 5,* 195–206. https://doi.org/10.1007/s11409-010-9057-1

Efklides, A. (2001). Metacognitive experiences in problem solving: Metacognition, motivation and self-regulation. In A. Efklides, J. Kuhl & R. M. Sorrentino (Eds.). *Trends and prospects in motivation research* (pp. 297-323). Kluwer.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*, 3–14. https://doi.org/10.1016/j.edurev.2005.11.001

Efklides, A. (2009). The role of metacognitive experiences in the learning process. *Psicothema*, *21*(1), 76–82.

Einbender, L., & Wood, D. (1995). An authentic journey: Teachers' emergent understandings about authentic assessment and practice. Columbia University. Teachers College National Center for Restructuring Education, Schools and Teaching.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education*, *27*(1), 31–45. https://doi.org/10.1080/08957347.2013.853070

Elliott, L. J., Lum, H. C., Aqlan, F., Zhao, R., & Lasher, C. D. (2019). A study of metacognitive problem solving in undergraduate engineering students. *Advances in Human Factors in Training, Education, and Learning Sciences*, 95–102. https://doi.org/10.1007/978-3-030-20135-7_9

Ellis, A. K. (2004). *Exemplars of curriculum theory*. Routledge: Taylor & Francis.

Ellis, L. B. (2003). Illuminative case study design: A new approach to the evaluation of continuing professional education. *Nurse Researcher, 10* (3), 48–59. https://doi.org/10.7748/nr2003.04.10.3.48.c5897

Emirbayer, M., & Mische, A. (1998). 'What is agency?'. *The American Journal of Sociology, 103*, 962–1023. https://doi.org/10.1086/231294

Eraslan, A. (2009). Finlandiya'nın PISA'daki başarısının nedenleri: Türkiye için alınacak dersler. [Reasons behind the Success of Finland in PISA: Lessons for Turkey]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 3*(2), 238–248.

Erdoğdu, B. (2009). Bilgisayar destekli sınav: Bilgisayar uyarlamalı sınav için soru sınıflandırma değerlendirmesi [Computer based testing: Evaluation of question classification for computer adaptive testing]. [Unpublished master thesis]. Bahçeşehir University.

Eren, B. (2015). Çoktan seçmeli ve karma test uygulamalarına ilişkin öğrenci başarıları ile öğrenci ve öğretmen görüşlerinin karşılaştırılması [The comparison of student achievements, students' and teachers' views for multiple choice and mixed format test applications]. [Unpublished master thesis]. Ankara University.

ERG (2017). Eğitim reformu girişiminin milli eğitim bakanlığı taslak öğretim programları inceleme ve değerlendirmesi. [Examination and evaluation of erg regarding ministry of national education draft curriculum]. https://bilimakademisi.org/wp-content/uploads/2017/02/EĞİTİM-REFORMU-GİRİŞİMİ'NİN-MİLLİ-EĞİTİM-BAKANLIĞI-TASLAK-ÖĞRETİM-PROGRAMLARI-İNCELEME-VE-DEĞERLENDİRMESİ-.pdf

Eriş, H., & Kılıçoğlu, A. (2019). Curriculum development competencies of form teacher candidates. *The Turkish Online Journal of Educational Technology, 18*(1), 25–31.

Ernest, P. (March, 1999). Social constructivism as a philosophy of mathematics: Radical constructivism rehabilitated? http://socialsciences.exeter.ac.uk/education/research/centres/stem/publications/pmej/soccon.htm

Erümit, A. (2014). Polya'nin problem çözme adımlarına göre hazırlanmış yapay zeka tabanlı öğretim ortamının öğrencilerin problem çözme süreçlerine etkisi [Artificial intelligence-based learning environments which preparing Polya's problem solving steps effect on students' problem-solving processes]. [Unpublished doctoral dissertation]. Karadeniz Technical University].

European Qualifications Framework (2008). European Qualifications Framework for Lifelong Learning (EQF) http://ecompetences.eu/wp-content/uploads/2013/11/EQF_broch_2008_en.pdf

European Strategy Framework (2015). Achieving the Millennium Goals Learning for Sustainability Maastricht, The Netherlands, November 15th – 17th 2002, Europe-wide Global Education Congress.

Falmagne, J. C., Cosyn, E., Doignon, J. P., & Thiery, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmid (Eds.), *Formal concept analysis* (pp. 61–79). Springer.

Fetters, M. K., Czerniak, C. M., Fish, L., & Shawberry, J. (2002). Confronting, challenging, and changing teachers' beliefs: Implications from a local systemic change professional development program. *Journal of Science Teacher Education, 13* (2), 101–130. https://doi.org/10.1023/A:1015113613731

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Flake, L. H. (2017). A look at the relationship of curriculum and instruction and the art and science of teaching. *Asian Journal of Education and Training, 3*(2), 82–85. https://doi.org/10.20448/journal.522.2017.32.82.85

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L.B. Resnick (Ed.), *The nature of intelligence* (pp. 231-325). Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066x.34.10.906

Fleming, S. M., & Frith, C. D. (Eds.). (2014). *The cognitive neuroscience of metacognition*. Springer-Verlag.

Foley, D. P. (1981). Instructional potential of teacher-made tests. *Teaching of Psychology, 8*(4), 243–244.

Forehand, M. (2005). Bloom's taxonomy: Original and revised. In M. Orey (Ed.), Emerging perspectives on learning, teaching, and technology. https://www.d41.org/cms/lib/IL01904672/Centricity/Domain/422/BloomsTaxonomy.pdf

Forgas, J. P. (1994). The role of emotion in social judgments: An introductory review and an Affect Infusion Model (AIM). *European Journal of Social Psychology*, *24*, 1–24. https://doi.org/10.1002/ejsp.2420240102

Fraenkel, J., Wallen, N., & Hyun, H. (2014). *How to design and evaluate research in education* (9th ed.). McGraw-Hill Education.

Frenken, L. (2021). Measuring students' metacognitive knowledge of mathematical modeling. In: Leung F.K.S., Stillman G.A., Kaiser G., Wong K.L. (eds) Mathematical Modeling Education in East and West. *International Perspectives on the Teaching and Learning of Mathematical Modeling.* Springer, Cham. https://doi.org/10.1007/978-3-030-66996-6_18

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation, 17*(2), 1–18. https://doi.org/10.7275/sxbs-0829

Gall, M. D. (1970). The use of questions in teaching. *Review of Educational Research*, *40*(5), 707–721. American Educational Research Association (AERA). https://doi.org/10.3102/00346543040005707

Gareis, C. R., & Grant, L. W. (2008). *Teacher-made assessments: How to connect curriculum, instruction, and student learning.* Eye on Education.

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

Garrido, A. (2010). Mathematics and Artificial Intelligence, two branches of the same tree. *Procedia Social and Behavioral Sciences, 2*(1), 1133–1136. https://doi.org/10.1016/j.sbspro.2010.03.160

Gascoine, L., Higgins, S., & Wall, K. (2016). The assessment of metacognition in children aged 4-16 years: a systematic review. *Review of Education, 5*(1), 3–57. https://doi.org/10.1002/rev3.3077

Gelbal, S. (2013). *Ölçme ve Değerlendirme [Measurement and Evaluation*]. Anadolu University.

Gerstner, J. J. & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research & Practice in Assessment, 8*(1), 15–28.

Glas, C. A., W., & van der Linden, W. J. (2001). Modeling variability in item parameters in item response models (Research Report 01–11). University of Twente.

Grammatikopoulos, V. (2012) Integrating program theory and systems-based procedures in program evaluation: a dynamic approach to evaluate educational programs. *Educational Research and Evaluation, 18*(1), 53–64.

Goodlad, J. I. (1968). Curriculum: a Janus look. *The Record, 70,* 95–107.

Goodlad, J. I. (1994). *What schools are for*. Phi Delta Kappa International Incorporated.

Goodlad, J. (2007). Kudzu, rabbits, and school reform. In A. Ornstein, E. Pajak, & S. Ornstein (Eds.), *Contemporary issues in curriculum* (4th ed., pp. 51-58). Allyn ve Bacon.

Goswami, U. (2006). Neuroscience and education: from research to practice? *Nature Reviews Neuroscience, 7*(5), 406–413). Springer Science and Business Media LLC. https://doi.org/10.1038/nrn1907

Göçmez, H. (February, 2017). 2017'de yapay zeka eğitim sektörüne nasıl etki yapacak? *Tech Insider*. https://www.techinside.com/2017de-yapay-zeka-egitim-sektorune-nasil-etki-yapacak/

Greene, M. (1995). *Releasing the imagination: Essays on education, the arts, and social change*. Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Sage.

Gullie K. A. (2011). Open-ended (extended/constructed) response questions as predictors of success on subsequent state mathematics examinations: The influence of mathematical awareness and conceptual knowledge. University at Albany. Department of Educational Theory and Practice, USA.

Guskey, T. R. (2010). Lessons of mastery learning. Educational, School, and Counseling Psychology Faculty Publications, 14. https://uknowledge.uky.edu/edp_facpub/14

Gültekin, S., & Çıkrıkçı-Demirtaşlı, N. (2012). Comparing the test information obtained from multiple-choice, open-ended and mixed item tests based on item response theory. *Elementary Online, 11* (1), 251–263.

Günel, K., Polat, R. & Kurt, M. (2016). Analyzing learning concepts in intelligent tutoring systems. *The International Arab Journal of Information Technology, 13*(2), 281–286.

Güneş, B. & Baki, A. (2011). Dördüncü sınıf matematik dersi öğretim programının uygulanmasından yansımalar [Reflections on the implementation of the fourth grade mathematics curriculum]. *Hacettepe University Journal of Faculty of Education, 41*(1), 192–205.

Güvendir, M.A. & Ozkan, Y.O. (2021). Uzaktan eğitimin değerlendirmeye yansımaları: Çevrimiçi sınavlar mı sınıf içi sınavlar mı? [Distance education's impact on Assessment: Online exams or classroom exams?]. *Journal of Digital Measurement and Evaluation Research*, *1*(1), 22–34. https://doi.org/10.29329/dmer.2021.285.2

Haberport. (2019). When will the MIS exam be? https://haberport.com/egitim/2019-mis-sinavi-ne-zaman-kac-gun-kaldi-sinav-tarihi-h1430.html

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(1), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hand, D. J. (2004). *Measurement theory and practice*. A Hodder Arnold.

Hansen, K. T., Heckman, J. J., & Mullen, K. J. (2003). The effect of schooling and ability on achievement test scores. [Working Paper], 2003(13), *Institute for Labor Market Policy Evaluation* (IFAU), Uppsala.

Hartell, E. & Strimel, G. J. (2019). What is it called and how does it work: examining content validity and item design of teacher-made tests. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-018-9463-2

Hayati, A., Bentri, A., & Rahmi, U. (2017). Analyzing the issues in the implementation of authentic assessment in the 2013 curriculum. *Al-Ta Lim Journal, 24*(1), 53–59. https://doi.org/10.15548/jt.v24i1.256

Haynes, (1935). The relation of teacher intelligence, Teacher, type of school to type of questions [Unpublished doctoral dissertation]. University of George.

Heck, J. L., & Stout, D. E. (1998). Multiple-choice vs. open-ended exam problems: Evidence of their impact on student performance in introductory finance. *Financial Practice and Education*, *8*, 83–93.

Heffernan, N. T., Turner, T. E., Lourenco, A. L. N., Macasek, M. A., Nuzzo-Jones, G. & Koedigner, K. R. (2006). The ASSISTment Builder: Towards an Analysis of Cost Effectiveness of ITS Creation. In Proceedings of the 19th International FLAIRS Conference, Melbourne Beach, FL, 515–520.

Hendrickson, K. A. (2017). Assessment in Finland: A scholarly reflection on one country's use of formative, summative, and evaluative practices. Mid-Western Educational Researcher, 25(1/2), 33–43.

Herman, J. L., Klein, D. C. D., & Wakai, S. T. (1997) American students' perspectives on alternative assessment: do they know it's different? *Assessment in Education: Principles, Policy & Practice, 4*(3), 339–352. https://doi.org/10.1080/0969594970040302

Hernandez-Orallo, J. (2016). Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artificial Intelligence Review, 48*(3), 397–447. https://doi.org/10.1007/s10462-016-9505-7

Ho, O. N. K. (2018). Rethinking the curriculum: *The epistle to the Romans as a pedagogic text.* Springer.

Hilebowitsh, P. S. (2005). *Designing the school curriculum* (1st ed.). Pearson Custom.

Holmqvist, K., & Andersson, R. (2017). Eye tracking: A comprehensive guide to methods, paradigms and measures. Lund Eye-Tracking Research Institute.

Hoogerheide, V., Staal, J., Schaap, L. & van Gog, T. (2019). Effects of study intention and generating multiple choice questions on expository text retention. *Learning and Instruction, 60*(1), 191–198. https://doi.org/10.1016/j.learninstruc.2017.12.006

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal, 27*(3), 406–422. https://doi.org/10.1080/09585176.2016.1156004

Hopman, M., DeWinter, M., & Koops, W. (2014). Analyzing the hidden curriculum. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10*(1), 12–20. https://doi.org/10.1027/1614-2241/a000063

Huang, T.-W., & Wu, P.-C. (2013). Classroom-based Cognitive Diagnostic Model For a Teacher-made Fraction- Decimal Test. *Educational Technology & Society, 16* (3), 347–361.

Human-Centered AI Institute (March, 2019). Infographic: The age of Artificial Intelligence. https://hai.stanford.edu/news/infographic-age-artificial-intelligence

IBM Corp (2012). IBM SPSS Statistics for MAC, Version 21.0. IBM Corp.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346–368. https://doi.org/10.16986/huje.2016015182

IMOTIONS.COM (2021). Eye-tracking. https://imotions.com/eye-tracking/

International Handbook of Metacognition and Learning Technologies. (2013). In R. Azevedo & V. Aleven (Eds.), Springer International Handbooks of Education. Springer. https://doi.org/10.1007/978-1-4419-5546-3

Interpress (September, 2015). *Artificial Intelligence in Education*.

İnceçam, B., Demir, E. & Demir, E. (2018). Ortaokul öğretmenlerinin sınıf içi ölçme ve değerlendirmelerde yazılı yoklamalarda kullandıkları açık uçlu maddeleri hazırlama yeterlikleri [Competencies of middle school teachers to prepare open-ended items used in open-ended test for in-classroom assessment]. *Elementary Education Online, 17*(4), 1912–1927. https://doi.org/10.17051/ilkonline.2019.506900

İnci, T. (2014). Sekizinci sinif fen ve teknoloji dersine ilişkin ortak sinav sorularinin değerlendirilmesi [Unpublished maser thesis]. Eskişehir Osmangazi University.

İş-Güzel, Ç., Kızıltan, A. & Gökçe, S. (2010). Matematik dersinde üst düzey düşünme becerilerinin geliştirilmesi [Development of higher-order thinking skills in mathematics]. *CITO Education: Theory and Practice Journal*, October-December Issue, 16–30.

Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in Educational Science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research, 10*(1). https://doi.org/10.16910/jemr.10.1.3

Jarodzka, H., Van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a Model's eye movements fosters learning. *Learning and Instruction, 25,* 62–70. https://doi.org/10.1016/j.learninstruc.2012.11.004

Jian, Yu-Cin (2016). Fourth graders' cognitive processes and learning strategies for reading illustrated biology texts: Eye movement measurements. *Reading Research Quarterly, 51*(1), 93–109. https://doi.org/10.1002/rrq.125

jMetrik™ (2021). Psycho measurement systems. https://itemanalysis.com/

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102%2F0013189X033007014

Jorgensen, M., Falster, K., Hanly, M., Goldfeld, S., Jorm, L., & Craven, R. (2017). Preschool attendance and early childhood development outcomes in the first year of school for Aboriginal and non-Aboriginal children in New South Wales, Australia. *International Journal of Population Data Science, 1*(1), 103. https://doi.org/10.23889/ijpds.v1i1.123

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Educational Sciences: Theory & Practice, 17*(2), 573–596. https://doi.org/10.12738/estp.2017.2.0280

Karadayı, Z. (2004). Computers supported preschool education and artificial intelligence [Unpublished master thesis]. Çanakkale Onsekiz Mart University.

Karadeniz, A. (2016). Kitlesel açık ve uzaktan öğrenmede başarının açık uçlu sorularla ölçülmesine yönelik bir sistemin tasarımı, uygulanmasi ve değerlendirilmesi [Design, evaluation and implementation of a system intended to assess the learners' achievement through open-ended questions in massive open and distance learning] [Unpublished doctoral dissertation]. Anadolu University, Eskişehir.

Karakaya-Özyer, K. (2018). Açıköğretim sistemindeki açık uçlu soruların çok yüzeyli Rasch modeli analizi yöntemiyle puanlanması. *AUAd, 4*(1), 61–77.

Kaya, E., Çetin, P. S., & Yıldırım, A. (2012). Transformation of centralized curriculum into classroom practice: An analysis of teachers' experiences. *International Journal of Curriculum and Instructional Studies, 2*(3), 103–113.

Kerkez, B. (2018). Mesleki ve teknik eğitimde program geliştirme ve eğitim programina ilişkin öğretmenler görüşleri [Teachers' opinions about curriculum development and curriculum in vocational and technical education] [Unpublished master thesis]. Hacettepe University.

Khan, W. B., & Inamullah, H. M. (2011). A study of lower-order and higher-order questions at secondary level. *Asian Social Science*, 7(9), Canadian Center of Science and Education. https://doi.org/10.5539/ass.v7n9p149

Kinay, İ. (2015). Otantik değerlendirme yaklaşımının öğretmen adaylarının problem çözme becerileri iile öğrenmeye ve katılımcı değerlendirmeye yönelik inançlarına etkisinin incelenmesi [the investigation of effects of authentic assessment approach on prospective teachers' problem - solving skills, beliefs towards learning and participative assessment] [Unpublished doctoral dissertation]. Gaziantep University.

Kirby, P. C. & Oescher, J. (1989). Assessing teacher-made math and science tests. *American Secondary Education, 18*(1), 6–12.

Kılıç, R. (2014). İlköğretim I. kademe öğretmenlerinin otantik ölçme ve değerlendirme yöntemleri ile ilgili bilgi, tutum ve görüşlerinin değerlendirilmesi [evaluation of knowledge, attitudes and opinions on the authentic measurement and evaluation methods of primary education teachers] [Unpublished doctoral dissertation]. Marmara University.

Koyuncu, E. (2017). Temel eğitimden ortaöğretime geçiş (teog) sınavında açık uçlu soruların kullanılmasına ilişkin öğrenci görüşlerinin incelenmesi. [Determination of student views on the use of open-ended questions on transition from primary to secondary education exam (teog)] [Unpublished master thesis]. Gaziantep University.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–264. https://doi.org/10.1207/s15430421tip4104_2

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. (2nd ed.) Sage.

Krippendorff, K. (2011). Computing Krippendorff 's alpha-reliability. http://repository.upenn.edu/asc_papers/43

Kuckartz, U. (2019). Mixed methods research in the social sciences [Powerpoint Document], Marburg University.

Kul, U., & Çelik, S. (2017). Exploration of pre-service teachers' beliefs in relation to mathematics teaching activities in classroom-based setting. *International Journal of Research in Education and Science (IJRES), 3*(1), 245-257. https://files.eric.ed.gov/fulltext/EJ1126698.pdf

Kyriakides, L., Creemers, B. P. M., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education, 50*(3), 381–393. https://doi.org/10.1007/s11858-018-0919-3

Lee, D. L. (2015). Using questions to develop students' higher-order thinking skills: a primary English teacher's beliefs and practices [Unpublished doctoral dissertation]. University of Hong Kong.

Lee, Y.-J., Kim, M., Jin, Q., Yoon, H.-G., & Matsubara, K. (2017). *East-Asian primary science curricula: An overview using revised Bloom's Taxonomy.* Springer.

Lee, Y.-J., Kim, M., & Yoon, H.-G. (2015). The intellectual demands of the intended primary science curriculum in Korea and Singapore: An analysis based on revised Bloom's taxonomy. *International Journal of Science Education, 37*(13), 2193–2213. https://doi.org/10.1080/09500693.2015.1072290

Lee, Y., Park, Y. & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177. https://doi.org/10.1080/15305058.2010.534571

Leijen, Ä., Pedaste, M., & Lepp, L. (2019). Teacher agency following the ecological model: how it is achieved and how it could be strengthened by different types of reflection. *British Journal of Educational Studies, 68*(3), 295–310. https://doi.org/10.1080/00071005.2019.1672855

Lemaignan, S., Warnier, M., Akin Sisbot, E., Clodic, A. & Alami, R. (2016). Artificial cognition for social Human–Robot interaction: An implementation, *Artificial. Intelligence,* http://dx.doi.org/10.1016/j.artint.2016.07.002

Leuffen, D., Shikano, S., & Walter, S. (2013). Measurement and data aggregation in small-n social scientific research. *European Political Science*, *12,* 40–51. https://doi.org/10.1057/eps.2012.8

Li, X., & Wang, W. (2020). Exploring spatial cognitive process among stem students and its role in stem education: A cognitive neuroscience perspective. *Science & Education, 30,* 121–145. https://doi.org/10.1007/s11191-020-00167-x

Lin, C.-J. & Chang, H.-H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement, 79*(2), 335–357. https://doi.org/10.1177%2F0013164418790634

Linacre, J. M. (1993). Generalizability theory and many facet rasch measurement. Annual Meeting of The American Educational Research Association. (April, 13, 1993), (ED 364 573). Atlanta Georgia

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Lingel, K., Lenhart, J. & Schneider, W. (2019). Metacognition in mathematics: Do different metacognitive monitoring measures make a difference?. *ZDM Mathematics Education 51*, 587–600. https://doi.org/10.1007/s11858-019-01062-8

Liu, F., Zhang, Y., Shi, Y., Chen, Z. & Feng, X. (2018). Analyzing the impact of characteristics on artificial intelligence iq test: A fuzzy cognitive map approach. *Procedia Computer Science, 139*(1), 82–90. https://doi.org/10.1016/j.procs.2018.10.221

Loh, M. Y. & Lee, N. H. (2019). The impact of various methods in evaluating metacognitive strategies in mathematical problem solving. In Liljedahl P., Santos-Trigo M. (Eds) *Mathematical Problem Solving*. ICME-13 Monographs. Springer, Cham. https://doi.org/10.1007/978-3-030-10472-6_8

Long, C., Dunne, T., & De Kock, H. (2014). Mathematics, curriculum and assessment: The role of taxonomies in the quest for coherence. *Pythagoras, 35*(2), Art. #240. http://dx.doi.org/10.4102/pythagoras.v35i2.240

Low, A. R., L., & Aryadoust, V. (2021). Investigating Test-taking strategies in listening assessment: A comparative study of eye-tracking and self-report questionnaires. *International Journal of Listening*, (pp. 1-20) https://doi.org/10.1080/10904018.2021.1883433

Lufi, D., Okasha, S., & Cohen, A. (2004). Test anxiety and its effects on the personality of students with learning disabilities. *Learning Disability Quarterly, 27*, 176–184. https://doi.org/10.2307/1593667

Lundahl, C. (2009). Varför nationella prov? Framväxt, dilemman, möjligheter [Why national tests? Growth, dilemmas, opportunities]. Studentlitteratur AB.

Macmillan, P. D. (2000). Classical, generalizability, and multifaceted rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68*(2), 167–190. https://doi.org/10.1080/00220970009598501

Malcı, E. (2021). Examining 10th grade students' problem-solving processes in geometry using eye tracking technology [Unpublished master's thesis]. Middle East Technical University.

Mangaroska, K., Sharma, K., Gašević, D., & Giannakos, M. (2021). Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity. Journal of *Computer Assisted Learning, 38*(1), 40–59. https://doi.org/10.1111/jcal.12590

Marso, R. N. & Pigge, F. L. (1991). An analysis of teacher-made tests: item types, cognitive demands, and item construction errors. *Contemporary Educational Psychology, 16*(1), 279–286. https://doi.org/10.1016/0361-476X(91)90027-I

Marulis, L. M., Baker, S. T., & Whitebread, D. (2020). Integrating metacognition and executive function to enhance young children's perception of and agency in their learning. *Early Childhood Research Quarterly, 50*, 46–54. https://doi.org/10.1016/j.ecresq.2018.12.017

Marzano, R. J. (2003). What works in schools: Translating research into action. Association for Supervision and Curriculum Development.

Mayer, R. E. & Alexander, P. A. (2011). *Handbook of research on learning and instruction*. Routledge: Taylor & Francis.

McCarthy, J (2007). What is artificial intelligence? [Technical report], Stanford

University. http://www-formal.stanford.edu/jmc/whatisai.html

McCowan, R. J., & McCowan, S. C. (1999). Item analysis for criterion-referenced tests. Research Foundation of SUNY/Center for Development of Human Services.

Mcewen, S. E., Huijbregts, M. P. J., Ryan, J. D., & Polatajko, H. J. (2009). Cognitive strategy use to enhance motor skill acquisition post-stroke: A critical review. *Brain Injury*, *23*(4), 263–277. https://doi.org/10.1080/02699050902788493

McHugh M. L. (2013). The chi-square test of independence. *Biochem Med (Zagreb), 23*(2), 14–39. https://doi.org/10.11613/bm.2013.018

McMillan, J. H., Myran, S. & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(1), 203–213. https://doi.org/10.1080/00220670209596593

McGraw Hill Education (2017). What makes ALEKS unique. https://www.aleks.com/about_aleks/overview

McREL International (2017, January). Higher-order questioning inspires higher-level thinking. https://www.mcrel.org/higher-order-questioning-inspires-higher-level-thinking/

MEB (February, 2017). Ankara'da taslak öğretim program çalıştayı düzenlendi [Draft curriculum workshops were organized in Ankara]. https://ankara.meb.gov.tr/www/ankarada-taslak-ogretim-programi-calistayi-duzenlendi/icerik/986

Mesker, P., Wassink, H., Akkerman, S., & Bakker, C. (2018). Student teacher's boundary experiences during an international teaching internship. *Teacher Education and Development, 5*(1), 1–16. https://doi.org/10.1080/2331186X.2018.1498577

Meyer, H.-D., Strietholt, R., & Epstein, D. Y. (2017). Three models of global education quality: The emerging democratic deficit in global education governance. In M. Akiba & G. K. LeTendre (Eds.), *International Handbook of Teacher Quality and Policy* (1st ed.), Routledge.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in pisa science multiple choice items. *International Journal of Testing, 20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* (2nd Edition). SAGE Publications.

Miller, J. (2007). *The holistic curriculum*. University of Toronto Press.

Middle East Technical University [METU] (2017). Report on the 2017 Turkish educational system curricular change https://eds.metu.edu.tr/en/announcement/our-departments-report-2017-turkish-educational-system-curricular-change-turkish

Ministry of National Education [MoNE/MEB]. (2017a) Açık uçlu sorular önemli bir dönüm noktası olacaktır [Open-ended questions will be an important milestone] http://www.meb.gov.tr/bakan-yilmaz-anadolu-ajansi-editor-masasi8217na-konuk-oldu/haber/12887/tr

Ministry of National Education [MoNE/MEB]. (2017b). Millî izleme sistemi için 23 il temsilcisi ile istişare toplantısı yapıldı [Consultation meeting with 23 provincial representatives for national monitoring system]. http://odsgm.meb.gov.tr/www/mill-izleme-sistemi-icin-23-il-temsilcisi-ile-istisare-toplantisi-yapildi/icerik/263

Ministry of National Education [MoNE/MEB] (2018). *Mathematics curriculum (primary and middle school 1,2,3,4,5,6,7 and 8th grades).* http://mufredat.meb.gov.tr/Dosyalar/201813017165445-MATEMATİK%20ÖĞRETİM%20PROGRAMI%202018v.pdf

Ministry of National Education [MoNE/MEB] (2019). Minister of Education's Twitter Mention, April 6, 2019. https://twitter.com/tcmeb/status/1114581415138603009

Ministry of National Education [MoNE/MEB] (July, 2021). High School Entrance System (LGS) Placed by Central Examination Student Performance. Education Analysis and Evaluation Report Series, No:17. Retrieved from: https://cdn.eba.gov.tr/icerik/2021/07/rapor/No_17-LGS_2021-merkezi_yerlestirme_211730.pdf

Ministry of National Education [MoNE/MEB] PISA Pre-report (2019, December). Education Analysis and Evaluation Report Series 10. http://pisa.meb.gov.tr/www/pisa-2018-turkiye-on-raporu-yayimlandi/icerik/3

Ministry of National Education [MoNE/MEB] (2019). Turkey's 2023 Education Vision. Ministry of National Education.

Minsky, M. L. (ed) (1968). Semantic information processing. MIT Press, Cambridge.

Mulkey, J. C., & O'Neil, H. F. (1999). The effects of test item format on self-efficacy and worry during a high-stakes computer-based certification examination. *Computers in Human Behavior, 15*(1), 495–509. https://doi.org/10.1016/S0747-5632(99)00033-3

Mullis, I. V. S., & Martin, M. O. (2019). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), USA.

Muna, K., & Bahit, M. (2020). Eye-tracking and metacognitive skills: A review on the use of eye-tracking for measuring students' metacognitive skills in chemistry learning. *Journal of Physics: Conference Series*, 1422, 012033. https://doi.org/10.1088/1742-6596/1422/1/012033

Munzar, B., Muis, K. R., Denton, C. A., & Losenno, K. (2020, February 10). Elementary students' cognitive and affective responses to impasses during mathematics problem solving. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000460

Nagel, N. G. (1993). Authentic learning/authentic assessment: let's begin with tomorrow's teachers. Clark College Teacher Education, USA.

National Education Statistics (2018). *Milli eğitim istatistikleri: Örgün eğitim 2017-2018 [National education statistics: formal education 2017-2018].* Ministry of National Education.

National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics.* Key Curriculum.

National Research Council. (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. Committee on Defining Deeper Learning and 21st Century Skills, James W. Pellegrino and Margaret L. Hilton, Editors. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. The National Academies Press.

Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. Routledge.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Nohl, A. M., & Somel, R. N. (2016). Curricular change in Turkey: Time, sequentiality, and differential power of actors in establishing a new knowledge path. *The Journal of Educational Change, 17*(1), 303–318. https://doi.org/10.1007/s10833-015-9265-1

Norqvist, M., Jonsson, B., Lithner, J., Qwillbard, T., & Holm, L. (2019). Investigating algorithmic and creative reasoning strategies by eye tracking. *The Journal of Mathematical Behavior, 55,* Article 100701. https://doi.org/10.1016/j.jmathb.2019.03.008

Nunnally, C. (1967). *Psychometric theory.* (pp. 172-235, see especially formulas 6-26, p. 196.). McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory.* (2nd Ed.). McGraw-Hill.

Nunez, R., Allen, M., Gao, R.C., Miller R., Rigoli, C. M., Relaford-Doyle, J., & Semenuks, A. (2020). For the sciences they are a-changin': A response to commentaries on Nunez et al.'s (2019) "What happened to Cognitive Science?" *Topics in Cognitive Science*, https://doi.org/10.1111/tops.12511

Nunn, L. M. (2014). *Defining student success: The role of school and culture*. Rutgers University.

Nyman, T., & Kaikkonen, P. (2013). What kind of learning environment do newly qualified teachers create?, *Scandinavian Journal of Educational Research, 57*(2), 167–181. https://doi.org/10.1080/00313831.2011.628690

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 160940691989922. https://doi.org/10.1177/1609406919899220

ODSGM (2017) Milli İzleme Sistemi [National Monitoring System] https://odsgm.meb.gov.tr/www/mill-izleme-sistemi/icerik/257

Oliva, P. F. (2009). *Developing the curriculum* (7th ed.), Pearson Education, Inc.

Omoteso, K. (2012). The application of artificial intelligence in auditing: Looking back to the future. *Expert Systems with Applications, 39*(1), 8490–8495. https://doi.org/10.1016/j.eswa.2012.01.098

O'Neil, H. F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *The Journal of Educational Research, 89*, 234–245. https://doi.org/10.1080/00220671.1996.9941208

O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Assessment in Education, 11*(4), 331–351. https://doi.org/10.1207/s15324818ame1104_3

Ornstein, A. C., & Hunkins, F. P. (2004). *Curriculum: Foundations, principles and issues* (4th ed.). Pearson Education.

Ort, V. K. (1967) Teacher-made tests. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 41*(7), 396–399. https://doi.org/10.1080/00098655.1967.11477388

O'Shea, J., & Leavy, A. M. (2013). Teaching mathematical problem-solving from an emergent constructivist perspective: the experiences of Irish primary teachers. *Journal of mathematics teacher education, 16*(4), 427–449. https://doi.org/10.1007/s10857-013-9235-6

Onwuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351-383). Sage.

Öksüz, Y., & Güven-Demir, E. (2018). Açık uçlu ve çoktan seçmeli başarı testlerinin psikometrik özellikleri ve öğrenci performansı açısından karşılaştırılması [Comparison of open-ended questions and multiple-choice tests in terms of psychometric features and student performance]. *Hacettepe University Journal of Education*. https://doi.org/10.16986/huje.2018040550

ÖSYM (2019). Exam calendar. https://www.osym.gov.tr/TR,8797/takvim.html

Özcan, B. N., & Delil, A. (2018). İlköğretim matematik öğretmenlerinin hazırladıkları testlerin öğretim programı kazanımları açısından bir analizi [An analysis of the tests prepared by primary mathematics teachers in terms of teaching program acquisitions]. *Kastamonu Education Journal, 26*(6), 1909–1917. https://doi.org/10.24106/kefdergi.2212

Özcan, M. (2011). *Bilgi çağında öğretmen: Eğitimi, nitelikleri, gücü. [Teachers in the information age: Education, qualities and power of teachers]* TED.

Özcan, M. (2013). *Okulda üniversite: öğretmen eğitimini yeniden yapılandırmak için bir model önerisi [University within school: A model to restructure* *teacher education in Turkey].* TÜSİAD. Information about TÜSİAD can be seen at www.**tusiad**.us or at *www.**tusiad**.org.*

Özcan, H., & Düzgünoğlu, H. (2017). Teachers' views on science draft curriculum 2017. *International Journal of Active Learning (IJAL), 2*(2), 28–47.

Özçelik, D. A. (1998). *Ölçme ve Değerlendirme* [Measurement and Evaluation]. (8th ed.). ÖSYM.

Özeren, E., & Akpunar, B. (2019). Analysis of teacher measurement-evaluation preferences from the perspective of epistemological belief. *Journal of Education and Training Studies, 7*(6), 84–92. https://doi.org/10.11114/jets.v7i6.4227

Özkaya, A. (2021). Türkiye ile Kazakistan ortaokul matematik dersi öğretim programlarının karşılaştırılması. [Comparison of Turkey and Kazakhstan middle school mathematics teaching programmes]. *e-Kafkas Journal of Educational Researh, 8*(3), 592–611. https://doi.org/10.30900/kafkasegt.987453

Öztürk, İ. H. (2012). Öğretimin planlanmasında öğretmenin rolü ve özerkliği: Ortaöğretim tarih öğretmenlerinin yıllık plan hazırlama ve uygulama örneğI [The role and autonomy of the teacher in the planning of the teaching: The example of preparation and implementation of the annual plan]. *Educational Sciences: Theory & Practice, 12*(1), 271–299.

Öztürk, M. (2021). An embedded mixed method study on teaching algebraic expressions using metacognition-based training. *Thinking Skills and Creativity, 39*, 100787. https://doi.org/10.1016/j.tsc.2021.100787

Öztürk, M., & Kaplan, A. (2019). Cognitive analysis of constructing algebraic proof processes: A mixed method research. *Education & Science, 44*(197), 25–64. https://doi.org/10.15390/eb.2018.7504

Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Sage.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Peters, R. S. (1970). Freedom to learn: A view of what education might become. [Review Article]. *Interchange*, *1*(4), 111–114. https://doi.org/10.1007/BF02214887

PISA (2015) PISA 2015 results in focus. ttps://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Pinar, W. (2012). *What is curriculum theory?* Taylor & Francis.

Polat, R. (2016). Tübitak araştırma projesi gelişme raporu (bilimsel rapor) [Tübitak research project development report (scientific report)]. Report Term: 15/09/2015-15/09/2016, Project No: 115E472.

Polya, G. (1990). *How to solve it?.* Princeton University.

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141-159). Lawrence Erlbaum.

Priestley, M. (2010). Schools, teachers and curriculum change: A balancing act*? Journal of Educational Change*, *12*(1), 1–23. https://doi.org/10.1007/s10833-010-9140-z

Priestley, M. (2016). A perspective on learning outcomes in curriculum and assessment. *the National Council for Curriculum and Assessment*, 1–7.

Priestley, M. (2017). Approaches to specifying curriculum areas of learning. University of Stirling/Welsh Government.

Priestley, M., & Biesta, G. J. J. (Eds.) (2013). *Reinventing the curriculum: new trends in curriculum policy and practice*. Bloomsbury Academic.

Priestley, M., Biesta, G., & Robinson, S. (2015). *Teacher agency: An ecological approach*. Bloomsbury Academic.

Priestley, M., & Drew, V. (2016). Teachers as agents of curriculum change: closing the gap between purpose and practice. *European Conference for Educational Research*, Dublin, 23–26 September 2016. http://www.eera-ecer.de/ecer-programmes/conference/21/contribution/39307/

Priestley, M., Edwards, R., Priestley, A., & Miller, K. (2012). Teacher agency in curriculum making: Agents of change and spaces for maneuver. *Curriculum Inquiry, 42*(2), 191–214. https://doi.org/10.1111/j.1467-873X.2012.00588.x

Radmehr, F., & Drake, M. (2017). Revised Bloom's taxonomy and integral calculus: Unpacking the knowledge dimension. *International Journal of Mathematical Education in Science and Technology*, *48*(8), 1206–1224. https://doi.org/10.1080/0020739X.2017.1321796

Radmehr, F., & Drake, M. (2018). An assessment-based model for exploring the solving of mathematical problems: Utilizing revised bloom's taxonomy and facets of metacognition. *Studies in Educational Evaluation, 59*(1), 41–51. https://doi.org/10.1016/j.stueduc.2018.02.004

Radmehr, F., & Drake, M. (2019). Revised Bloom's taxonomy and major theories and frameworks that influence the teaching, learning, and assessment of mathematics: A comparison. *International Journal of Mathematical Education in Science and Technology*, *50*(6), 895–920. https://doi.org/10.1080/0020739X.2018.1549336

Rahimi, M., & Alavi, J. (2017). The role of teaching experience in language teachers' perceptions of a top-down curriculum change. *The Curriculum Journal, 28*(4) 479–503. https://doi.org/10.1080/09585176.2017.1344134

Random, H. (2016). Random house unabridged dictionary. http://dictionary.reference.com/browse/instruction

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distributed Learning, 15*(5), 83–98.

Reynolds, S. T. D. (2019). Validating Bloom's Revised Taxonomy as a Rubric for Assessing Middle School Students' Levels of Thinking. [Unpublished doctoral dissertation]. Walden University.

Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology*, *19*(6), 551–554. https://doi.org/10.1037/h0033456

Rimbatmojo, S., Kusmayadi, T. A., & Riyadi, R. (2017). Metacognition difficulty of students with visual-spatial intelligence during solving open-ended problems. In Journal of Physics: Conference Series (Vol. 895, p. 012034). IOP Publishing. https://doi.org/10.1088/1742-6596/895/1/012034

Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. *Educational Researcher*, *49*(2), 80–89. American Educational Research Association (AERA). https://doi.org/10.3102/0013189x19890600

Roderer, T., & Roebers, C. M. (2014). Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. In Metacognition and Learning (Vol. 9, Issue 1, pp. 1–23). Springer Science and Business Media LLC. https://doi.org/10.1007/s11409-013-9109-4

Rogers, C. R. (1969). *Freedom to learn: A view of what education might become*. Charles E. Merrill.

Rose, E. (2004). Instructional design and curriculum development: deconstructing the difference. *Educational Technology*, *44*(2), 3–13.

Rossi, M. (2017). Artificial Intelligence Forces us to Change Math. https://view.joomag.com/mag/0904196001509906528?page=20

Rotherham, A. J., & Willingham, D. T. (Spring, 2010). “21-st Century” skills: Not new, but a worthy challenge. American Educator, 17–20. https://dbweb01.aft.org/sites/default/files/periodicals/RotherhamWillingham.pdf

Russell, S., & Norvig, P. (2009). Artificial intelligence: a modern approach. Prentice Hall: Upper Saddle River.

Russell, S. (2019). *İnsanlık için yapay zeka: Yapay zeka ve kontrol problemi [Human compatible: Artificial intelligence and the problem of control].* Buğdağı.

Ryan, A. M. (2001). Explaining the black-white test score gap: the role of test perceptions. *Human Performance, 14*(1), 45–75. https://doi.org/10.1207/S15327043HUP1401_04

Sadler-Smith, E., & Riding, R. (1999). Cognitive style and instructional preferences. *Instructional Science, 27*, 355–371. https://doi.org/10.1007/BF00892031

Šafranj, J. (2019). The effect of meta-cognitive strategies on self-efficacy and locus of control of gifted in foreign language learning. *Research in Pedagogy, 9*(1), 40–51. https://doi.org/10.17810/2015.90

Sahlberg, P. (Summer, 2017). Small data for big change: What does it mean for teachers? ATA Magazine. https://pasisahlberg.com/ata-small-data-2017/

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00. the symposium. https://doi.org/10.1145/355017.355028

Saxe, J. G. (1872). The Blind Men and the Elephant. [A Poem].

Say, C. (2018). *Yapay Zeka [Artificial Intelligence]*. Science and Future Publication.

Schleicher, A. (2017). Türk Eğitim Sistemi Dünyaya Uyum Sağlayamadı [Turkish education system could not adapt to the world]. https://tr.sputniknews.com/turkiye/201711131030970907-pisa-direktoru-andreas-schleicher-turk-egitim-sistemi-dunya-uyum/

Schmoelz, A., Swertz, C., Forstner, A. & Barberi, A. (2015). Does artificial tutoring foster inquiry-based learning? *Science Education International, 25*(1), 123–129.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *University of Massachusetts Amherst*. https://doi.org/10.7275/SWGT-RJ52

Shaughnessy, M. F., Veenman, M., & Kleyn-Kennedy, C. (2008). Meta-cognition: A recent review of research, theory, and perspectives. Nova Science Publishers, Inc.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessment: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27. https://doi.org/10.2307/1177207

Shawer, S. F. (2017). Teacher-driven curriculum development at the classroom level: Implications for curriculum, pedagogy and teacher training. *Teaching and Teacher Education, 63*(1), 296–313. https://doi.org/10.1016/j.tate.2016.12.017

Shuilleabhain, A. N. & Seery, A. (2018). Enacting curriculum reform through lesson study: A case study of mathematics teacher learning. *Professional Development in Education, 44*(2), 222–236. https://doi.org/10.1080/19415257.2017.1280521

Sireci, S. G., & Zenisky, A. L. (2011). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In *Handbook of test development* (pp. 343-362). Routledge.

Slavin, R. E. (2003). *Educational psychology: Theory and practice, 7/E*. Allyn & Bacon/Longman.

Smawley, R. B. (1962). Teacher-made Tests Improve by Use of Shortcut Item Analysis. *The Clearing House: A Journal of Educational Strategies,* *Issues and Ideas, 37*(3), 138–141. https://doi.org/10.1080/00098655.1962.11476230

Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology, 27*(1), 65–77. https://doi.org/10.1080/0020739960270109

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies guidance for educators.* RAND Corporation.

Solanki, V., & Evans, B. R. (2020). A comparative study of higher-order thinking and teaching in the United States and United Kingdom. *Curriculum and Teaching 35*(2), 45–63. https://doi.org/10.7459/ct/35.2.04

Sole, M. A. (2016) Multiple problem-solving strategies provide insight into students' understanding of open-ended linear programming problems. *PRIMUS, 26*(10), 922–937. https://doi.org/10.1080/10511970.2016.1199621

Sole, M. A. (2018). Open-ended questions: A critical class component. *The Mathematics Teacher*, *111*(6), 462–465. Retrieved from http://www-jstor-org.uaccess.univie.ac.at/stable/10.5951/mathteacher.111.6.0462

Sözcü (September, 2015). Yapay Zeka eğitimde çığır açacak [Artificial Intelligence will break ground in education]. https://www.sozcu.com.tr/egitim/yapay-zeka-egitimde-cigir-acacak.html

Stark, L., Brünken, R., & Park, B. (2018). Emotional text design in multimedia learning: A mixed-methods study using eye tracking. *Computers & Education, 120,* 185–196. https://doi.org/10.1016/j.compedu.2018.02.003

Stein, M. K., Remillard, J. T., & Smith, M. S. (2007). How curriculum influences student learning. In F. Lester (Ed.). *Second handbook of research on mathematics teaching and learning* (pp. 319–369). Information Age Publishing.

Stepankova, B., & Emanovsky, P. (2011). An open-ended and closed-ended question in didactic tests of mathematics. *Problems of Education in The 21stcentury*, 28, 114–122.

Sternberger, L., Pysarchik, D. T., Yun, Z.-S., & Deardorff, D. (2009). Designing a model for international learning assessment. https://www.aacu.org/publications-research/periodicals/designing-model-international-learning-assessment

Stillman, G. (2020) Metacognition. In S. Lerman (Eds) *Encyclopedia of Mathematics Education*. Springer, Cham. https://doi.org/10.1007/978-3-030-15789-0_166

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). Strategies and processes for developing innovative items in large-scale assessments. [Pearson Assessment Report]. http://images.pearsonassessments.com/images/tmrs/StrategiesandProcessesforDevelopingInnovativeItems.pdf

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics, 104*(2), 147–200. https://doi.org/10.1007/s10649-020-09948-1

Şahin, A. (2014). Öğretmenlerin Öğretmenler Tarafından Hazırlanan Sınavların Güvenilirliği ve Geçerliliği Iile İlgili Algıları [Teachers' Perceptions of Reliability and Validity in Teacher Made Examinations]. [Unpublished master thesis]. Çağ University.

Şimşek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, *4*(4), 477–489.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*, (6th ed.). Allyn and Bacon.

Talim Terbiye Kurulu Başkanlığı [TTKB] (2017). Müfredatta yenileme ve değişiklik çalışmalarımız üzerine [On the renovation and differentiation studies of curriculum]. https://ttkb.meb.gov.tr/meb_iys_dosyalar/2017_07/18160003_basin_aciklamasi-program.pdf

Tan, L. S. (2019). East-Asian primary science curricula: an overview using revised Bloom's taxonomy. *Asia Pacific Journal of Education, 39*(1), 154–155. https://doi.org/10.1080/02188791.2019.1585053

Tan-Şişman, G. & Karsantık, Y. (2017). An investigation of the curriculum development process in Singapore and Turkey regarding administrative structure and reforms [Singapur ve Türkiye'de program geliştirme süreçlerinin yönetsel yapı ve reformlar açısından incelenmesi]. Paper presented at the 26. International Educational Sciences Congress, April 20-23, 2017, Antalya, Turkey.

Tashakkori, A. & Teddlie, C. (1998). Mixed methodology combining qualitative and quantitative approaches (Vol. 46). CA Sage Publications.

Taub, M., & Azevedo, R. (2018). How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive SRL processes during learning with an intelligent tutoring system? *International Journal of Artificial Intelligence in Education, 29*(1), 1–28. https://doi.org/10.1007/s40593-018-0165-4

TEDMEM (2013). Kazak modeli nedir? [What is Kazakh model?] http://www.tedmem.org/haberler/2013/05/28/kazak_modeli_nedir.html

TEDMEM (February, 2017). Öğretim programı taslaklarında değerler eğitimi [Value education in draft curricula]. https://tedmem.org/mem-notlari/gorus/ogretim-programi-taslaklarinda-degerler-egitimi

Tekbıyık, A. & Akdeniz, A. R. (2008). İlköğretim fen ve teknoloji dersi öğretim programını kabullenmeye ve uygulamaya yönelik öğretmen görüşleri [Teachers' views about adoption and application of primary science and technology curriculum]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 2*(2), 23–37.

The Government of Istanbul Socio-Economic Status Analysis (2019-2020). The Government of Istanbul.

Thompson, C. (2003). Introduction to artificial intelligence [Course Notes]. http://www.cs.utah.edu/classes/cs55300

The Organization for Economic Cooperation and Development [OECD] (2012). Evaluation and Assessment. https://www.oecd.org/education/school/NLD_CBR_Evaluation_and_Assessment.pdf

The Organization for Economic Cooperation and Development [OECD] (2016). Education at a Glance 2016: OECD Indicators, OECD Publishing, Paris. http://dx.doi.org/10.187/eag-2016-en

The Organization for Economic Cooperation and Development [OECD] (2018a). Turkey: Overview of the education system (EAG 2018). http://gpseducation.oecd.org/CountryProfile?primaryCountry=TUR&treshold=10&topic=EO

The Organization for Economic Cooperation and Development [OECD] (2018b). Education at a Glance 2018: OECD Indicators. https://doi.org/10.1787/eag-2018-69-en

The Organization for Economic Cooperation and Development [OECD] (2018c). *The Future of Education and Skills: Education 2030*. Retrieved from: https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

The Organization for Economic Cooperation and Development [OECD] (2019). *Trends Shaping Education 2019*, OECD Publishing, Paris. https://doi.org/10.1787/trends_edu-2019-en

The Partnership for 21st Century Learning [P21] (2018). *21st Century Learning Exemplar Program.* http://www.battelleforkids.org/networks/p21/21st-century-learning-exemplar-program

Ticha, M., & Hospesova, A. (2006). Qualified pedagogical reflection as a way to improve mathematics education. *Journal of Mathematics Teacher Education, 9*(2), 129–156. https://doi.org/10.1007/s10857-006-6893-7

Tiven, M. B., Fuchs, E. R., Bazari, A., & MacQuarrie, A. (2018). Evaluating Global Digital Education: Student Outcomes Framework. Bloomberg Philanthropies and the Organization for Economic Co-operation and Development.

Tiekstra, M., Minnaert, A. & Hessels, M. G. P. (2016) A review scrutinizing the consequential validity of dynamic assessment. *Educational Psychology, 36*(1), 112–137. https://doi.org/10.1080/01443410.2014.915930

Tobias, S. (1987). Math anxiety. *Science*, *237*(4822), 1556. https://doi.org/10.1126/science.237.4822.1556.a

Tofade, T., Elsner, J., & Haines, S. T. (2013). Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education*, *77*(7), 155. https://doi.org/10.5688/ajpe777155

Toraman, S. (2021). Mixed methods research: a brief history, definitions, perspectives, and key elements. *Qualitative Social Sciences*, *3*(1), 1–29. https://doi.org/10.47105/nsb.847688

Tsai, P.-Y., Yang, T.-T., She, H.-C., & Chen, S.-C. (2019). Leveraging college students' scientific evidence-based reasoning performance with eye-tracking-supported metacognition. *Journal of Science Education and Technology*, *28*(6), 613–627. https://doi.org/10.1007/s10956-019-09791-x

TTKB (July, 2017). Müfredatta yenileme ve değişiklik çalışmaları üzerine [On the studies of renewal and change of curriculum]. http://ttkb.meb.gov.tr/meb_iys_dosyalar/2017_07/18160003_basin_aciklamasi-program.pdf

Tuomi, I. (2018). The impact of artificial intelligence on learning, teaching, and education policies for the future, Eds. Cabrera, M., Vuorikari, R. & Punie, Y., EUR 29442 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-97257-7. https://doi.org/10.2760/12297

Turing, A. M. (1950). COMPUTING MACHINERY AND INTELLIGENCE. Mind: Vol. LIX (Issue 236, pp. 433–460). Oxford University Press (OUP). https://doi.org/10.1093/mind/lix.236.433

Turkey’s 2023 Education Vision (October, 2018). 2023 Education Vision [Educational Policy Report]. http://2023vizyonu.meb.gov.tr

Turkey’s Qualifications Framework (2016). published in Official Gazette No. 29581 dated 2/1/2016 and entered into force. https://www.myk.gov.tr/index.php/en/turkiye-yeterlilikler-cercevesi

TUSIAD (2013). *PISA 2012 değerlendirmesi: Türkiye için veriye dayalı eğitim reform önerileri.* [Evaluation of PISA 2012: Education Reform Suggestions Depending on Data for Turkey.]

Türkoğlu, T. (2021). Deep data. https://www.herkesebilimteknoloji.com/yazarlar/tanol-turkoglu/derin-veri

Tyler, R. (1949). *Basic principles of curriculum and instruction.* University of Chicago.

Unal D. S., Arrington C.M., Solovey E., & Walker, E. (2020). Using Think Alouds to Understand Rule Learning and Cognitive Control Mechanisms Within an Intelligent Tutoring System. In: Bittencourt I., Cukurova M., Muldner K., Luckin R., Millán E. (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science, vol 12163. Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_40

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2013). *National Assessment of Educational Progress (NAEP), 2013 Mathematics Assessment.* http://nces.ed.gov/nationsreportcard/ itemmaps/index.asp

Usiskin, Z. (2012). Misidentifying factors underlying Singapore's high test scores: A strong curriculum is not the sole reason for Singaporean students' success on international assessments. Mathematics Teacher, 105(9), 666–670. https://doi.org/10.5951/mathteacher.105.9.0666

Uymaz, M., & Çalışkan, H. (2019). An investigation on the teacher-made social studies course exam questions in terms of revised Bloom's taxonomy. *Kastamonu Education Journal, 27*(1), 331–346. https://doi.org/10.24106/kefdergi.2637

Uysal, I., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 28–54. https://doi.org/10.21031/epod.817396

Üstkan, S. (2007). Uzman sistemler [Expert systems]. Directed Project, Sakarya University.

van Gog, T., & Jarodzka, H. M. (2013). Eye tracking as a tool to study and enhance cognitive and Metacognitive processes in computer-based learning environments. In R. Azevedo, & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (Vol. 28, pp. 143-156). Springer. Springer International Handbooks of Education https://doi.org/10.1007/978-1-4419-5546-3_10

van Gog, T., & Scheiter K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction, 20*, 95–99. 10.1016/j.learninstruc.2009.02.009

van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2020). Determinants of 21st-century skills and 21st-century digital skills for workers: A systematic literature review. *SAGE Open, 10*(1). https://doi.org/10.1177/2158244019900176

van Loon, M. H., Bayard, N. S., Steiner, M., & Roebers, C. M. (2020). Connecting teachers' classroom instructions with children's metacognition and learning in elementary school. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09248-2.

VanSledright, B. (2008). Narratives of nation-state, historical knowledge, and school history education. *Review of Research in Education*, *32*, 109–146. https://doi.org/10.3102%2F0091732X07311065

Vaismoradi, M., Jones, J., Turunen, H., & Snelgrove, S. (2016). Theme development in qualitative content analysis and thematic analysis. *Journal of Nursing Education and Practice, 6*(5), 100–110. https://doi.org/10.5430/jnep.v6n5p100

van Loon, M. H., Bayard, N. S., Steiner, M., & Roebers, C. M. (2021). Connecting teachers' classroom instructions with children's metacognition and learning in elementary school. In *Metacognition and Learning* (Vol. 16, Issue 3, pp. 623–650). Springer Science and Business Media LLC. https://doi.org/10.1007/s11409-020-09248-2

Verhoeven, P., & Verloop, N. (2002). Identifying changes in teaching practice: Innovative curricular objectives in classical languages and the taught curriculum. *Journal of Curriculum Studies, 34*(1), 91–102. https://doi.org/10.1080/00220270110069190

Voogt, J., & Roblin, N. P. (2010). 21st century skills. [Discussion paper]. University of Twente. https://www.bvekennis.nl/wp-content/uploads/documents/10-0405-White-Paper-21stCS-Final-ENG-def2-1.pdf

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition Learning*. https://doi.org/10.1007/s11409-020-09257-1

Walker, K. (2006). Teacher made exam designs. [Research Brief]. https://files.eric.ed.gov/fulltext/ED537929.pdf

Wallace, C. S. & Priestley, M. R. (2017). Secondary science teachers as curriculum makers: mapping and designing Scotland's new curriculum for excellence. *Journal of Research in Science Teaching*, *54*(3), 324–349. https://doi.org/10.1002/tea.21346

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

Wang, C. (2021). On interim cognitive diagnostic computerized adaptive testing in learning context. *Applied Psychological Measurement, 45*(4), 235–252. https://doi.org/10.1177/0146621621990755

Wang, Y. & Jiang, W. (2018). An automatic classification and clustering algorithm for online learning goals based on cognitive thinking. *International Journal of Emerging Technologies in Learning (iJET), 13*(11), 54–66. https://doi.org/10.3991/ijet.v13i11.9587

Wang, S., Yang, Y., Culpepper, S. A. & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1), 57–87.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

White, J. W., Moye, J. J., Gareis, C. R. & Hylton, S. P. (February, 2018). Improving teacher-made assessments in technology and engineering education. *Technology and Engineering Teacher*, 23–28.

Willingham, D. T. (2017). A mental model of the learner: Teaching the basic science of educational psychology to future teachers. *Mind Brain Educ., 11,* 166–175. 10.1111/mbe.12155

Yalçın, S. (2018). 21. yüzyıl becerileri ve bu becerilerin ölçülmesinde kullanılan araçlar ve yaklaşımlar [21st century skills and tools and approaches that are used to measure these skills]. Ankara University, *Journal of Faculty of Educational Sciences, 51*(1), 183–201. https://doi.org/10.30964/auebfd.405860

Yamaguchi, K. & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE, 13*(2), 1–17. https://doi.org/10.1371/journal.pone.0188691

Yaşar, M. D. (2012). 9*. Sınıf kimya öğretim programındaki yapılandırmacılığa dayalı öğelerin öğretmenler tarafından algılanışı ve uygulamasına yönelik bir inceleme: Erzurum örneği* [An examination of the perception and application of the elements based on constructivism in the 9th grade chemistry curriculum: Erzurum case] [Unpublished doctoral dissertation]. Atatürk University.

Yaşar University (2017). Yapay Zeka eğitimde çığır açacak [Artificial Intelligence to break ground in education]. http://btto.yasar.edu.tr/yapay-zeka-egitimde-cigir-acacak/

Yayla, K. & Yayla, T. (2018). 2017 Fizik öğretim programının öğretmen görüşleri doğrultusunda değerlendirilmesi (Ordu ili örneği) [Evaluation of the new physics instruction program in direction of the teacher opinions (Ordu example)]. *Ordu University Journal of Social Science Research, 8*(1), 89–94.

Yates, R. (2000). Curriculum overview. http://www.multiage-education.com/russportfolio/curriculumtopics/curoverview.html

Yılmaz, N. (2019). *The effect of the hypothetical learning trajectories and the contribution of eye-tracking technology in understanding young children's mathematical patterning recognition and generalization* [Unpublished master's thesis]. Middle East Technical University.

Zhang, L. (2018). *Metacognitive and cognitive strategy use in reading comprehension*. Springer. https://doi.org/10.1007/978-981-10-6325-1.

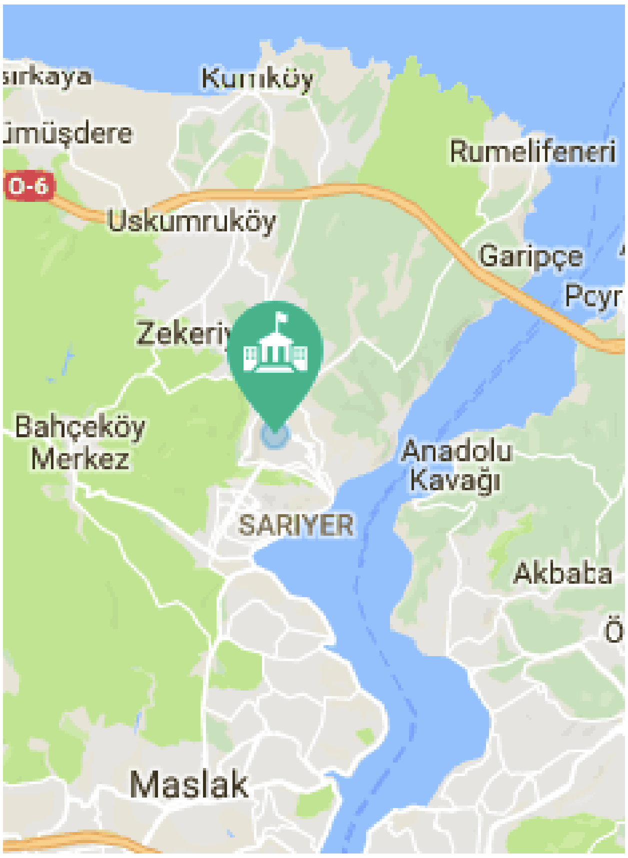
Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342. https://doi.org/10.1207/S15324818AME1604_4

Zhang, N., & Shen, J. (2012). Comparison of characteristics of agency of primary school teachers' professional development. *Educational Research and Experiment, 3*(1), 72–76.

Zimmerman, B. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional methods. In D. Schunk & B. Zimmerman (Eds.) *Self-regulated learning: From teaching to self-reflective practice*. The Guilford.

APPENDICES

A. SARIYER DISTRICT IN TURKEY



B. PUBLIC SCHOOLS IN NEIGHBOURHOOD OF SARIYER DISTRICT

Public School Name	Neighborhood	Number of Teachers	Number of Students	Classroom
Osman Saçmacı	Reşitpaşa mah.	28	660	20
Tuncay Artun İMKB Doğanevler	Reşitpaşa	23	536	21
Yeniköy İHO	Yeniköy mah.	-	-	-
İstanbul Ticaret Odası Mehmetçik	Yeniköy	21	367	16
Yeniköy	Yeniköy	16	313	10
Orgeneral Emin Alpkaya	Yeniköy mah	22	374	25
Şehit Muharrem Kerem Yıldız İHO	Yeniköy	45	787	36
Şehit Uğur Taşçı	Ferahevler mah.	18	390	19
Hürriyet	Ferahevler	-	-	-
Nilüfer Gökay	Ayazağa mah.	40	1209	29
Gümüşdere	Gümüşdere mah.	13	112	6
Zekeriyaköy	Zekeriyaköy	22	680	20
Kazım Karabekir	Poligon	11	226	9
Mehmet Akif	Pınar	36	600	24
Fahrettin Arslan	Cumhuriyet	30	450	20
Mehmet İpkin	Büyükdere	26	275	12
Mehmet Sevim Ulusal İHO	İstinye mah	30	685	20
Recaizade Ekrem	İstinye mah	21	387	24
Alpaslan	Tarabya mah	27	490	20
Prof Ali Kemal Yiğitoğlu	Maslak mah	17	300	10
Sarıyer	Merkez mah	72	1079	18
Hatemoğlu	Çamlıtepe	31	402	20
Zübeyde Hanım	PTT Evleri mah	-	-	-
Hacı Cemal Öğüt İHO	PTT Evleri mah	15	327	12
Turgut Akan	Kazım Karabekir paşa	-	-	-
Kumköy Ferhan Bedii Feyzioğlu	Kumköy	7	156	6
Anafartalar	Uskumruköy	16	220	9
Bahçeköy Türkan Efe	Bahçeköy	31	400	20
Türkan Şoray	Rumelihisarı	23	298	15
Şair Nigar	Rumelihisarı mah	9	180	10
MEV Dumlupınar	Fatih Sultan Mehmet mah	17	657	29
Kocataş Barbaros	Kocataş mah	18	368	12
R.Güney Kıldıran	Rumeli Kavağı mah	15	181	8
Veysel Vardal Görme Engelliler	Kumköy	20	79	11
Ayazağa	Huzur mah	23	857	22
Süleyman Çelebi	Huzur mah	35	580	21
Hacı Mehmet Şalgamcıoğlu	Baltalimanı	23	375	13

C. PRIVATE SCHOOLS IN NEIGHBOURHOOD OF SARIYER DISTRICT

Private School Name	Neighborhood
Özel Cent Koleji	Tarabya
Özel Mürüvvet Evyap	Maden mah
Özel Enka	İstinye
Özel Sarıyer Açı	Bahçeköy
Özel İstek Okulları Kemal Atatürk	Tarabya
Özel Sarıyer Doğa Koleji	Rumeli Feneri cad
Özel Erol Altaca koleji	Ferahevler mah
Özel Boğaziçi Fatih Koleji	Yeniköy mah
Özel Darüşşafaka Eğitim Kurumları	Darussafaka cad Maslak
Özel Boğazhisar Koleji	Maden mah
Özel Tarabya İngiliz Okulları	Yeniköy
İTÜ GVO Özel Dr Natuk Birkan	Maslak
FMV Işık Okulları Ayazağa	Maslak

D. APPROVAL OF METU HUMAN SUBJECT ETHICS COMMITTEE: FIRST PHASE

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr

Sayı: 28620816 / 231

05 NİSAN 2018

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Doç.Dr. Hanife AKAR

Danışmanlığını yaptığınız doktora öğrencisi Bengi BİRGİLİ'nin ***"Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"*** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2018-EGT-045** protokol numarası ile **06.04.2018 - 30.09.2018** tarihleri arasında geçerli olmak üzere verilmiştir.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Ş. Halil TURAN
Başkan V

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR
Üye

Doç. Dr. Yaşar KONDAKÇI
Üye

Doç. Dr. Zana ÇITAK
Üye

Doç. Dr. Emre SELÇUK
Üye

Dr. Öğr. Üyesi Pınar KAYGAN
Üye

E. APPROVAL OF METU HUMAN SUBJECT ETHICS COMMITTEE: SECOND PHASE

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr

Sayı: 28620816 / 110

06 MART 2019

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Doç.Dr. Hanife AKAR

Danışmanlığını yaptığınız Bengi BİRGİLİ'nin **"Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülmüş ve **101-ODTÜ-2019** protokol numarası ile onaylanmıştır.

Saygılarımızla bilgilerinize sunarız.

Prof. Dr. Tülin GENÇÖZ
Başkan

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR (4.)
Üye

Prof. Dr. Yaşar KONDAKÇI
Üye

Doç. Dr. Emre SELÇUK
Üye

Doç. Dr. Pınar KAYGAN
Üye

Dr. Öğr. Üyesi Ali Emre TURGUT
Üye

F. APPROVAL OF METU HUMAN SUBJECT ETHICS COMMITTEE: THIRD AND FOURTH PHASE

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59

Sayı: 28620816 / 101

16 MART 2020

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Prof.Dr. Hanife AKAR

Danışmanlığını yaptığınız **Bengi BİRGİLİ'nin "Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülmüş ve **101 ODTU 2020** protokol numarası ile onaylanmıştır.

Saygılarımızla bilgilerinize sunarız.

Prof.Dr. Mine MISIRLISOY
Başkan

Prof. Dr. Tolga CAN
Üye

Doç.Dr. Pınar KAYGAN
Üye

Dr. Öğr. Üyesi Ali Emre TURGUT
Üye

Dr. Öğr. Üyesi Şerife SEVİNÇ
Üye

Dr. Öğr. Üyesi Müge GÜNDÜZ
Üye

Dr. Öğr. Üyesi Süreyya Özcan KABASAKAL
Üye

G. APPROVAL OF MINISTRY OF NATIONAL EDUCATION: FIRST PHASE

T.C.
İSTANBUL VALİLİĞİ
İl Millî Eğitim Müdürlüğü

Sayı : 59090411-44-E.12469270 28.06.2018
Konu: Anket ve Araştırma İzin Talebi

Sayın: Bengi BİRGİLİ

İlgi: a) 11.06.2018 tarihli ve 11330076 Gelen Evrak No'lu dilekçeniz.
b) Valilik Makamının 26.06.2018 tarih ve 12320435 sayılı oluru.

"Öğretmenlerin Öğretim Yönetimi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi" konulu araştırma çalışmanız hakkındaki ilgi (a) dilekçe ve ekleri ilgi (b) valilik onayı ile uygun görülmüştür.

Bilgilerinizi ve söz konusu talebiniz; bilimsel amaç dışında kullanmaması, **uygulama sırasında bir örneği müdürlüğümüzde muhafaza edilen mühürlü ve imzalı veri toplama araçlarının kurumlarımıza araştırmacı tarafından ulaştırılarak uygulanılması,** katılımcıların gönüllülük esasına göre seçilmesi, araştırma sonuç raporunun müdürlüğümüzden izin alınmadan kamuoyuyla paylaşılmaması koşuluyla, gerekli duyurunun araştırmacı tarafından yapılması, okul idarecilerinin denetim, gözetim ve sorumluluğunda, eğitim-öğretimi aksatmayacak şekilde ilgi (b) Valilik Onayı doğrultusunda uygulanması ve işlem bittikten sonra 2 (iki) hafta içinde sonuçtan Müdürlüğümüz Strateji Geliştirme Bölümüne rapor halinde bilgi verilmesini rica ederim.

M. Nurettin ARAS
Müdür a.
Müdür Yardımcısı

EK:1- Valilik Onayı
2- Ölçekler

İl Milli Eğitim Müdürlüğü Binbirdirek M. İmran Öktem Cad.
No:1 Eski Adliye Binası Sultanahmet Fatih/İstanbul
E-Posta: sgb34@meb.gov.tr

A. BALTA VHKİ
Tel: (0 212) 455 04 00-239
Faks: (0 212)455 06 52

Bu evrak güvenli elektronik imza ile imzalanmıştır. https://evraksorgu.meb.gov.tr adresinden c8fa-f4ab-3bc9-acc1-4303 kodu ile teyit edilebilir.

T.C.
İSTANBUL VALİLİĞİ
İl Millî Eğitim Müdürlüğü

Sayı : 59090411-20-E.12320435 26/06/2018
Konu : Anket ve Araştırma İzin Talebi

VALİLİK MAKAMINA

İlgi: a) 11.06.2018 tarihli ve 11330076 Gelen Evrak No'lu dilekçe.
b) MEB. Yen. ve Eğ. Tk. Gn. Md. 22.08.2017 tarih ve 12607291/ 2017/25 No'lu Gen.
c) Milli Eğitim Müdürlüğü Araştırma ve Anket Komisyonunun 19.06.2018 tarihli tutanağı.

Orta Doğu Teknik Üniversitesi Sosyal Bilimler Enstitüsü doktora programı öğrencisi Bengi BİRGİLİ'nin **"Öğretmenlerin Öğretim Yönetimi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"** konulu tezi kapsamında, ilimiz Sarıyer ilçesinde bulunan ortaokullarda görev yapan matematik öğretmenlerine; yarı yapılandırılmış görüşme formunu uygulama istemi hakkındaki ilgi (a) dilekçe ve ekleri Müdürlüğümüzce incelenmiştir.

Araştırmacının söz konusu talebi; bilimsel amaç dışında kullanılmaması, **uygulama sırasında bir örneği müdürlüğümüzde muhafaza edilen mühürlü ve imzalı veri toplama araçlarının kurumlarımıza araştırmacı tarafından ulaştırılarak uygulanılması, katılımcıların gönüllülük esasına göre seçilmesi, araştırma sonuç raporunun müdürlüğümüzden izin alınmadan kamuoyuyla paylaşılmaması koşuluyla, okul idarelerinin denetim, gözetim ve sorumluluğunda, eğitim-öğretimi aksatmayacak şekilde ilgi (b) Bakanlık emri esasları dâhilinde uygulanması, sonuçtan Müdürlüğümüze rapor halinde (CD formatında) bilgi verilmesi kaydıyla** Müdürlüğümüzce uygun görülmektedir.

Makamlarınızca da uygun görülmesi halinde olurlarınıza arz ederim.

Ömer Faruk YELKENCİ
Milli Eğitim Müdürü

OLUR
26/06/2018

Ahmet Hamdi USTA
Vali a.
Vali Yardımcısı

Ek:1- Genelge
2- Komisyon Tutanağı

İl Milli Eğitim Müdürlüğü Binbirdirek M. İmran Öktem Cad.
No:1 Eski Adliye Binası Sultanahmet Fatih/İstanbul
E-Posta: sgb34@meb.gov.tr

A. BALTA VHKİ
Tel: (0 212) 455 04 00-239
Faks: (0 212)455 06 52

H. APPROVAL OF MINISTRY OF NATIONAL EDUCATION: SECOND PHASE

T.C.
İSTANBUL VALİLİĞİ
İl Milli Eğitim Müdürlüğü

Sayı : 59090411-20-E.8335765 26/04/2019
Konu : Anket ve Araştırma İzin Talebi

VALİLİK MAKAMINA

İlgi: a) 29.03.2019 tarihli ve 6504710 Gelen Evrak No'lu dilekçe.
b) MEB. Yen. ve Eğ. Tk. Gn. Md. 22.08.2017 tarih ve 12607291/ 2017/25 No'lu Gen.
c) Milli Eğitim Müdürlüğü Araştırma ve Anket Komisyonunun 25.04.2019 tarihli tutanağı.

Orta Doğu Teknik Üniversitesi Sosyal Bilimler Enstitüsü doktora programı öğrencisi Bengi BİRGİLİ'nin **"Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"** konulu tezi kapsamında, ilimiz genelinde bulunan özel/resmi ortaokullarda görev yapan matematik öğretmenlerine; anket istemi hakkındaki ilgi (a) dilekçe ve ekleri Müdürlüğümüzce incelenmiştir.

Araştırmacının söz konusu talebi; bilimsel amaç dışında kullanılmaması, **uygulama sırasında bir örneği müdürlüğümüzde muhafaza edilen mühürlü ve imzalı veri toplama araçlarının kurumlarımıza araştırmacı tarafından ulaştırılarak uygulanılması, katılımcıların gönüllülük esasına göre seçilmesi, araştırma sonuç raporunun müdürlüğümüzden izin alınmadan kamuoyuyla paylaşılmaması koşuluyla, okul idarelerinin denetim, gözetim ve sorumluluğunda, eğitim-öğretimi aksatmayacak şekilde ilgi (b) Bakanlık emri esasları dâhilinde uygulanması, sonuçtan Müdürlüğümüze rapor halinde (CD formatında) bilgi verilmesi kaydıyla** Müdürlüğümüzce uygun görülmektedir.

Makamlarınızca da uygun görülmesi halinde olurlarınıza arz ederim.

Levent YAZICI
İl Milli Eğitim Müdürü

Ek:
1- Genelge.
2- Komisyon Tutanağı.

OLUR
26/04/2019

Ahmet Hamdi USTA
Vali a.
Vali Yardımcısı

I. APPROVAL OF MINISTRY OF NATIONAL EDUCATION: THIRD AND FOURTH PHASE

T.C.
SARIYER KAYMAKAMLIĞI
Prof. Ali Kemal Yiğitoğlu Ortaokulu Müdürlüğü

Sayı : 47654154-605.01-E.16887727 18.11.2020
Konu : Doktora Tez Çalışma İzin Talebi

T.C MİLLİ EĞİTİM BAKANLIĞINA
(Strateji Geliştirme Başkanlığı)
ANKARA

27326038944 T.C Kimlik no'lu Bengi BİRGİLİ, MEF Üniversitesi Eğitim Fakültesi Araştırma Görevlisi olup, Orta Doğu Teknik Üniversitesi Eğitim Bilimleri Bölümü Doktora öğrencisidir. 'Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi' başlıklı doktora tez çalışması kapsamında etik kurul evrakları ile Müdürlüğümüze başvurmuştur. Milli Eğitim Bakanlığı Yenilik ve Eğitim Teknolojileri Genel Mudurluğu'nün Araştırma Uygulama İzinleri kapsamındaki genelgesi okunmuş, "Madde 14: Deneme modelli araştırmalar için yapılacak başvurular öncesinde, araştırmacıların il/ilçe milli eğitim mudurlukleri ile uygulamanın gerçekleştirileceği okullar ve ilgili kişilerle önceden irtibata geçerek bilgilendirme yapmaları ve alınan okul onayını başvuru esnasında ibraz etmeleri gerekmektedir." maddesi tarafımızca incelenmiştir. Doktora öğrencisi Bengi BİRGİLİ'nin okulumuz öğrencilerine ulaşabilmesi, örneklem bulabilmesi ve deneme çalışması yapmasının tarafımızca herhangi bir sakıncası bulunmamaktadır. Okulumuz çalışmaya katılmaya gönüllüdür. İl ve İlçe Milli Eğitim Müdürlüklerinden izin alındıktan sonra okulumuzda çalışmaya başlayabilir. Bu kapsamda biz de İstanbul / Sarıyer Prof.Ali Kemal Yiğitoğlu Ortaokulu Müdürlüğü olarak duyurumuzu velilerimize online olarak yapacağız.

Süleyman BİNBOĞA
Okul Müdürü

T.C.
İSTANBUL VALİLİĞİ
İl Milli Eğitim Müdürlüğü

Sayı : E.59090411-20-18144888 16/12/2020
Konu : Anket ve Araştırma İzin Talebi

VALİLİK MAKAMINA

İlgi : a) Orta Doğu Teknik Üniversitesinin 30.11.2020 tarihli ve 255 sayılı yazısı.
b) Bakanlığımızın 21.01.2020 tarih ve 1563890/ 2020/2 No'lu genelgesi.
c) Müdürlüğümüz Araştırma ve Anket Komisyonunun 15.12.2020 tarihli tutanağı.
d) Bakanlığımız Ortaöğretim Genel Müdürlüğünün 19.11.2020 tarihli ve 16964289 sayılı yazısı.

Orta Doğu Teknik Üniversitesi Eğitim Bilimleri Doktora Öğrencisi Bengi BİRGİLİ 'nin "Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi" konulu tezine dair araştırma çalışmasını ilimiz Sarıyer ilçesinde yer alan ortaokulllarda öğrenim gören öğrencilerine yönelik; görüşme soruları uygulama isteği hakkındaki ilgi (a) yazı ve ekleri müdürlüğümüzce incelenmiştir.

Bakanlığımızın ilgi (d) yazısında **04.01.2021** tarihine kadar uzaktan eğitime devam edileceği belirtilmiştir.

Araştırmacının söz konusu talebi **yüz yüze eğitime geçilmesine müteakiben**, bilimsel amaç dışında kullanılmaması, uygulama sırasında bir örneği müdürlüğümüzde muhafaza edilen mühürlü ve imzalı veri toplama araçlarının kurumlarımıza araştırmacı tarafından ulaştırılarak uygulanılması, katılımcıların gönüllülük esasına göre seçilmesi, araştırma sonuç raporunun kamuoyuyla paylaşılmaması koşuluyla, okul idarelerinin denetim, gözetim ve sorumluluğunda, eğitim-öğretimi aksatmayacak şekilde ilgi (b) Bakanlık emri esasları dâhilinde uygulanması, araştırma bittikten sonra 2 (iki) hafta içerisinde araştırma sonuç raporu hakkında Müdürlüğümüz istmemanket@meb.gov.tr adresine mail yoluyla bilgi verilmesi kaydıyla Müdürlüğümüzce uygun görülmektedir.

Makamlarınızca da uygun görülmesi halinde olurlarınıza arz ederim.

Levent YAZICI
İl Milli Eğitim Müdürü

OLUR
16/12/2020
Dr. Hasan Hüseyin CAN
Vali a.
Vali Yardımcısı

Ek:
1- Genelge.
2- Komisyon Tutanağı.

Adres: İstanbul Milli Eğitim Müdürlüğü - Strateji Geliştirme Şb.Md
Binbirdirek Mh.İmran Öktem Cd. No:1 Sultanahmet-Fatih/İstanbul
Elektronik Ağ: istanbul.meb.gov.tr
e-posta: ist.sgb34@gmail.com
Bilgi için: Aykut ÇELİK
Tel: 0 (212) 384
Faks: 0 (__) ___

J. APPROVAL OF A SAMPLE PUBLIC SCHOOL IN SARIYER DISTRICT

T.C.
İSTANBUL VALİLİĞİ
İl Milli Eğitim Müdürlüğü

Sayı : E.59090411-44-18306715 22.12.2020
Konu : Anket Araştırma İzni

ORTA DOĞU TEKNİK ÜNİVERSİTESİ REKTÖRLÜĞÜNE

İlgi : a) 30.11.2020 tarihli ve 255 sayılı yazınız.
b) Valilik Makamının 16.12.2020 tarihli ve 18144888 sayılı oluru.

Üniversiteniz Doktora Öğrencisi Bengi BİRGİLİ'nin **"Öğretmenlerin Öğretim Yöntemi Tercihlerinin Ekolojik Bir Yaklaşımla İncelenmesi"** konulu tez araştırma çalışması hakkındaki ilgi (a) yazınız ilgi (b) valilik onayı ile uygun görülmüştür.

Bilgilerinizi ve araştırmacının söz konusu talebi; bilimsel amaç dışında kullanmaması, **uygulama sırasında bir örneği müdürlüğümüzde muhafaza edilen mühürlü ve imzalı veri toplama araçlarının kurumlarımıza araştırmacı tarafından ulaştırılarak uygulanılması,** katılımcıların gönüllülük esasına göre seçilmesi, araştırma sonuç raporunun kamuoyuyla paylaşılmaması koşuluyla, gerekli duyurunun araştırmacı tarafından yapılması, okul idarecilerinin denetim, gözetim ve sorumluluğunda, eğitim-öğretimi aksatmayacak şekilde ilgi (b) Valilik Onayı doğrultusunda uygulanması ve işlem bittikten sonra 2 (iki) hafta içinde sonuçtan Müdürlüğümüz Strateji Geliştirme Bölümüne rapor halinde bilgi verilmesini arz ederim.

Levent ÖZİL
İl Milli Eğitim Müdürü a.
Müdür Yardımcısı

Ek:
1- Valilik Onayı
2- Ölçekler

K. MATHEMATICS TEACHERS INTERVIEW PROTOCOL FORM

Görüşme Protokol Formu

Değerli Katılımcı,

Bu çalışmanın amacı, 2017-2018 eğitim öğretim yılından itibaren okullarda uygulanmakta olan matematik eğitim programı çerçevesince sınıf içerisinde öğretim yöntemi ve ölçme değerlendirme süreci olarak ne tür değişiklikler yaptığınızı ortaya çıkarmaktır.

Çalışma kapsamında yapılacak görüşmeler gönüllülük esasına dayanmaktadır. Araştırmacı ile yapılacak görüşme yaklaşık 30 dakika sürmektedir. Çalışma sonunda paylaştığınız her türlü bilgi araştırmacılar tarafından saklanacak ve kişisel bilgileriniz tamamen gizli tutulacaktır. Görüşmeler sonucu toplanan veriler sadece araştırma amaçlı kullanılacaktır. Görüşme esnasında herhangi bir gerekçe bildirmeden istediğiniz anda geri çekilme hakkınız vardır.

Çalışma grubumuz Sarıyer ilçesine bağlı devlet ve özel okulda çalışan matematik öğretmenlerinden oluşmaktadır. Görüşme protokol formunda iki bölüm mevcuttur. Birinci bölümde kişisel bilgilere ilişkin sorular, ikinci bölümde ise öğretim yönteminiz ve ölçme-değerlendirme süreciniz ile ilgili görüşlerinizi almaya yönelik sorular sorulacaktır.

Katkılarınız için teşekkür ederiz.

Arş. Gör. Bengi BİRGİLİ Doç. Dr. Hanife AKAR

MEF Üniversitesi Orta Doğu Teknik Üniversitesi

Teacher’s Semi-structured Interview Form

Tarih ve Saat:___ / ___ / ____   ___: ___
Cinsiyet: ☐ Kadın ☐ Erkek
Kurum: ____
Branş: ____________________________
Mezun olduğu bölüm: ☐ Öğretmenlik:__________ ☐ Diğer:_________
Formasyon alınan yer:_____________ Süresi: ____________________
Kıdem Yılı: ________(ay/yıl)
Kurumda geçirilen süre: ________________(ay/yıl)
Hizmet içi Eğitim Süresi: _______________ (gün/hafta/saat)
Haftalık ders saatiniz:__________________ (saat)
Ders verdiğiniz sınıflar: ☐ 5. sınıf ☐ 6. sınıf ☐ 7. sınıf ☐ 8. sınıf
Sınıfınızdaki ortalama öğrenci sayısı: __________

Görüşme Soruları

1. 2017-2018 eğitim öğretim yılında uygulamaya konulan matematik eğitim programı yenilenmiştir. Bu konuyla ilgili size bilgilendirme nasıl oldu? Kimler tarafından, nereden ve nasıl bilgi verildi?

2. Değişen matematik eğitim programı ile ilgili ilk incelemeleri nasıl gerçekleştirdiniz?
   a. Kendiniz mi yoksa zümre ile mi incelediniz?

3. Program ile ilgili genel değerlendirmeniz nelerdir?
   a. Amaç-felsefe,
   b. içerik,
   c. işleyiş, ve
   d. değerlendirme süreci hakkında neler söylersiniz?

4. Yeni programın 2016-2017 eğitim öğretim yılında kullandığınız matematik eğitim programından ayrılan yanları nelerdir?
   a. Eski ve yeni programların arasında nasıl farklılıklar görüyorsunuz?

   b. Kullandığınız kaynaklar, ders kitabı ve benzeri öğretmen el kılavuzu bağlamında ne tür değişiklikler yaptınız?

   c. Önceki seneyi dikkate alırsanız bu sene derslerinizi nasıl işliyorsunuz?
      a. öğretim yönteminizde herhangi bir değişiklik oldu? Örnek vererek ayrıntılı açıklar mısınız?

5. 2017-2018 eğitim öğretim yılında uygulamaya konulan eğitim programına uygun konuyu işlerken ölçme değerlendirme süreci hakkında deneyimlerinizi paylaşır mısınız?
   a. Önceki seneyi dikkate alırsanız bu sene nasıl bir ölçme-değerlendirme süreci uyguluyorsunuz?

   b. Hangi farklı ölçme-değerlendirme yöntemlerini kullanıyorsunuz? Neden?

c. Kullandığınız ölçme-değerlendirme yöntemlerinden hangisinin en etkili olduğunu düşünüyorsunuz? Neden?

d. Öğrenci cevaplarını değerlendirirken dikkat ettiğiniz noktalar neler?

6. Açık uçlu sorulara dayalı bir sınav hazırlarken ve uygularken yaşadığınız
   a. olumlu deneyimler nelerdir?

   b. olumsuz deneyimler nelerdir?

7. Öğrencileriniz açık uçlu soruları çözerken nasıl deneyimler yaşamaktadır?
   a. Öğrencilerin üst bilişsel becerilerini ölçmesi hakkında neler düşünüyorsunuz? Bugüne kadar neler gözlemlediniz?

   b. Öğrencilerin üst bilişsel becerilerden olan bilişsel strateji ve öz kontrol becerilerini kullanması konusunda neler söyleyebilirsiniz?

   c. Öğrenciler problem çözme süreçlerinde nasıl bir yaklaşım uyguluyorlar? Gözlemleriniz nelerdir?

8. Öğrencilerinizin açık-uçlu problemler çözerken yaşadıkları duyguları nelerdir? Gözlemlerinizde bahseder misiniz?
   a. Kaygı belirtileri
   b. Rahat tutum
   c. Sürekli yazdıklarını değiştiriyor/siliyor

Deneyimlerinizi paylaştığınız ve araştırmaya zaman ayırdığınız için çok teşekkür ediyoruz. Sözlerinizi bitirmeden önce önemli bulup eklemek istediğiniz bir şey varsa paylaşabilirseniz memnun olurum.
Teşekkürler ☺

L. CODEBOOK FROM MATHS TEACHERS INTERVIEW PROTOCOL FORM

Soru 3. Genel Değerlendirme		
Kazanım (yetkinlik, beceri)	Kazanım fazlalığı dolayısı ile yetiştirme zorluğu	
İşleyiş	Sorgulayıcı mantık yürütme eğitiminin zorluğu, somutlaştırma, zaman yönetimini kullanabilme	
İçerik	Konu içeriğinin sadeleştirilmesi, sarmallık, konu derinliğine inilmesi	
Amaç_İçerik	İçeriğin genel amaçlar ile uyumsuzluğu, tutarsızlık	
Amaç_Değerlendirme	Eğitim program ile değerlendirme arasındaki tutarsızlık	
Değerlendirme	Soru Çözümüne vakit ayrılması, öncekinden fazla soru çözülebilmesi	
Amaç_İşleyiş	Etkinliğin azalması, amaç ve uygulamanın örtüşmemesi	
Soru 4.b. Kaynak, Ders kılavuzu, El kitabı		
Soru_kitap	Soru kalıpları farklı olan kitapların seçilmesi, örnekler içeren	
Sınav_Soru	Sınav sistemine hazırlayıcı sorular içeren: problem çözümü, muhakeme	
Deneme	Deneme sınavları	
Kaynak	Ek kaynak ihtiyacı, el kitabı, Kılavuz kitap	
Ders_Kit	Ders kitabını kullanma	

Soru 4.c. Öğretim Yöntemi		
Benzer	Herhangi bir değişikliğin olmaması	
Aktif	Aktif öğrenme, öğrencinin merkezde olduğu, aktif rol aldığı öğrenme yaklaşımları, etkinlik, dikkat çekme, rol yapma, tartışma	
Günlük	Günlük hayattan örnekler kullanma, ilişkilendirme yapma, gerçek yaşam örnekleri, somut materyal kullanma, problem çözme	
Anlatım	Doğrudan öğretim yöntemini kullanma	
Grup	Grup çalışması yöntemi	
Gezi Gözlem	Öğretim sırasında sınıf ortamına getirilemeyen cisim, olgu ve olayların yerinde ve planlı olarak incelenmesini gerektiren bir teknik olan gezi tekniği	
Doğaçlama	Role play, Drama ile öğrenme yöntemi önceden bir hazırlık yapmadan o an doğaçlamalarla bir öğrenme gerçekleşiyorsa bu drama "yaratıcı drama" denir. Bu yöntem öğrencilerin düşünmesini olaylara bakış açısını ve hayal gücünü geliştirir.	
Tartışma	Tartışmaların yaşandığı öğretim süreci	
Araştırma-İnceleme		"Kümeler konusunda neyi araştırıp isterim ben genelde, şey yaparlar kümesinin kesişimi, şunlar bunlar gerçi yok, onlar yok işte lise bire atmışlar gene. Orda iç içe şeyler geçer ya kümeler geçer falan işte niye geçer, sonra oradan biri olimpiyatlardan çıkar, mutlaka çıkar, olimpiyatın sembolü vardır ya, araştırım gelin, oradaki niye altı tane ya da niye yedi tane. Böyle, konu geldikçe." (Teacher A)

Soru 5.a. Ölçme Değerlendirme Süreci		
OE	Açık uçlu soru tarzları, kısa cevaplı sorular	
Procedural	Mantık ve işlem odaklı olan sorular	
St_Need	Öğrenciye uygun/bireysel sorular sormak, öğrenci ihtiyaçlarını dikkate almak	"Her sene sınavlarımda açık uçlu sorular sordum. Mantık, işlem odaklı olan ve olmayan sorular sorarak her öğrenciye uygun sormaya çalışıyorum." (Teacher E)
Quiz	Küçük ara sınavlar	
Download	Hazır online kaynaklardan indirmek, hazır soruları kullanmak	
Fill	Boşluk doldurma tarzında (test kitabında vs)	"Evet, sonra ben o kitabı dolduruyorum boşluk doldurma kitabını dolduruyorum çocukla. Onu, o doldurdukları yerleri birebir kontrol etmeye dikkat ediyorum. Yani herkes, bitiren getirsin gibi." (Teacher B
Proje		
Sunum		
Soru 5.b. Hangi Farklı ÖD Yöntemleri		
Soru_Cevap2	Soru cevap yöntemi ile bilinmeyeni belirleme	
Ödev	Günlük hayat problemleri ile ilişkilendirme	"Evin dikdörtgen, kare biçiminde olan odalarının ve halılarının alanlarını ve çevresini hesaplamaları için ödev veriyorum." (Teacher E)
Merak	Öğrencide merak uyandırma	
Katılım	Öğrencinin derse katılım göstermesi	
Karma	OE, MC, T/F, karma soru tiplerini içeren ölçme süreci	

Beyin_F	Beyin fırtınası yöntemi	
HOM	Üst düzey matematiksel bilgi ve becerileri ölçen sorular	
Gözlem	Öğretmen olarak sınıf içerisinde gözlem yapmak	Gözlem yapmak her zaman benim için mesela, kendim öğretmen olarak aldığım en büyük veri. Çünkü bi çocuğu birden fazla yerde gözlemleyebilirsiniz. Grup çalışması içerisinde yaptığı yorumlarda…
Grup	Grupla öğretim yöntem ve tekniklerini sınıf içerisinde kullanmak	

M. DIMENSIONS OF TMMESP-Q IN TURKISH

ÖĞRETİM YÖNTEMİ TERCİHİ

1. **Boyut: Program Tasarımı ile ilgili Görüşler**

Item 1. Programın amacında bir değişiklik olmadığını düşünüyorum (Ornstein & Hunkins, 2004).
Item 2. Programın felsefesinde bir değişiklik olmadığını düşünüyorum (Ornstein & Hunkins, 2004; Ozmon & Craver, 2008).
Item 3. Kazanım sayılarının azaldığını fark ediyorum (Ornstein & Hunkins, 2004).
Item 4. Konu içeriklerinin zenginleştiğini düşünüyorum (Oliva, 2009; Ornstein & Hunkins, 2004).
Item 7. Milli Eğitim Bakanlığının Matematik kaynak kitabının içeriğinde değişiklik olmadığını düşünüyorum (Yüksel, 2000).
Item 8. Öğretmen el kitabını kullanmak tercihimdir (Demirel, 1992; 2012).

2. **Boyut: Genel Sınıf içi Uygulama ile ilgili Görüşler**

Item 6. Öğrencilerin yaratıcılıklarına fırsat sağlayan aktiviteler yapmak tercihimdir (Özmantar, Bingölbali, Demir, Sağlam & Keser, 2009).
Item 9. Sınıf içi öğretim yöntemimi öğrencilerimi aktif kılacak şekilde değiştiririm (Phillipson, Riel & Leger, 2018).
Item 11. Öğrencilerime; matematiği etkin bir şekilde keşfederek öğrenmelerini sağlayacak şekilde bir ders tasarlıyorum (Bruner, 1961, 1996; Abrahamson & Kapur, 2018).
Item 13. Ders öncesinde öğrencilerin hazırbulunuşluklarını test ediyorum (Özer & Anıl, 2011).

3. **Boyut: Teknik ile ilgili Görüşler**

Item 5. Sınıf içi öğretimlerde somut materyal (örn. matematiksel obje) kullanmak bana çok yardımcı oluyor (Baroody, 2017).
Item 10. Grupla öğretim yöntemi tekniklerini (örn. işbirlikli öğrenme, düşün-eşleş-paylaş vs.) kullanmayı tercih ederim (Mayer & Alexander, 2011).
Item 12. Sadece anlatım tekniğini kullanırım (Moore,1986; Westerhof, 1992).
Item 14. Sınıf içi öğretimde eğitim teknolojilerini kullanmaya çalışırım (Bos, 2009).
Item 16. Farklı soru sorma tekniklerini (Neden? Niçin? vb.) kullanmaya ihtiyaç duyuyorum (Michaels, Connor, Hall, & Resnick, 2010).
Item 17. Anlattığım konuya günlük hayattan örnek veririm (Kitchen, 2016; NCTM, 2014).

4. **Boyut: Yapılandırmacılık ile ilgili Görüşler**

Item 15. öğretim esnasında yapılandırmacı yaklaşım (örn. bilgiyi araştırma, yorumlama ve analiz etme, düşündürme sürecini geliştirme vs.) tekniklerini kullanmak tercihimdir (Von Glasersfeld, 1995).
Item 18. Öğrencilere üzerinde çalıştıkları konu hakkında düşünmelerini sağlayan bir öğrenme ortamı tasarlamayı tercih ederim (Von Glasersfeld, 1995).
Item 19. Öğrencilerin kendi öğrenme süreçlerinde sorumluluk aldıkları öğretim tekniklerini (gösteri, soru-cevap, beyin fırtınası, tartışma) kullanırım (Mayer & Alexander, 2011).
Item 20. öğrencileri araştırma yapmaya teşvik ediyorum (Bruner, 1961;Clabaugh, 2010).

ÖLÇME-DEĞERLENDİRME SÜRECİ

1. **Boyut: Genel Ölçme-Değerlendirme Süreci ile ilgili Görüşler**

Item 1. önceki seneye göre ölçme-değerlendirme sürecinde değişiklik yapıyorum (Alkharusi, Kazem & Al-Musawai, 2011).
Item 2. sınavlarımı çevrimiçi kaynaklardan (forum, paylaşım siteleri vb.) indirerek uyguluyorum.
Item 4. dersin kazanımlarını ölçmek için biçimlendirici değerlendirmeyi (formatif değerlendirme) kullanıyorum (DeLuca, Valiquette, Coombs, LaPointe-McEwan & Luhanga, 2006; Marzano, 2006).
Item 11. öğrencilere sınavlardaki soru türleri hakkında tercih hakkı tanıyorum (Gelbal & Kelecioğlu, 2001; Zhang & Burry-Stock, 2003).
Item 12. sınavlarda işlemsel becerilerini (prosedürel) kullanacakları soru türlerini tercih ediyorum.

2. **Boyut: Soru Formatları ile ilgili Görüşler (kuramsal geçerlilik)**

Item 3. çoktan seçmeli sorular içeren ölçme araçları kullanıyorum (Kanatlı, 2008).
Item 5. çoktan seçmeli ve kısa cevaplı soruları içeren sınavlar hazırlıyorum (Çakan, 2004; Kilmen & Çıkrıkçı-Demirtaşlı, 2009).
Item 7. öğrencilerin dönem sonunda performanslarını gösterecekleri öğrenci portfolyo dosyalarını kullanıyorum (Barootchi & Keshavarz, 2002; Özbaşı, 2008).
Item 8. kısa sınavları (quiz) uyguluyorum (Çakan, 2004).
Item 9. sınıf içi sınavlarımda açık uçlu soruları kullanıyorum (Birgili, 2014).
Item 14. sınıf içi sınavlarımda hem açık uçlu, hem de çoktan seçmeli soruları kullanıyorum (Özbaşı, 2008).

3. **Boyut: Sınıf içi Öğretimi Ölçme ve Değerlendirme Teknikleri ile ilgili Görüşler**

Item 6. günlük hayattaki problemleri çözmeye yönelik sorular soruyorum (Shepard et al., 2005).
Item 10. öğretim sırasında soru-cevap tekniğini kullanıyorum (Özbaşı, 2008).
Item 13. soruları öğrencilerin üst düzey bilişsel becerilerini (ne bildiğini bilme, düşüncenin farkındalığı) kullanacakları soru türlerinden (PISA, TIMSS gibi uluslararası sınav soruları) seçiyorum (Kilmen & Çıkrıkçı-Demirtaşlı, 2009).
Item 15. uyguladığım sınıf içi sınavların sayısını öğrencilerle birlikte belirliyorum (Acar-Erdol & Yıldızlı, 2018).

N. TRANSLATION OF TMMESP-Q

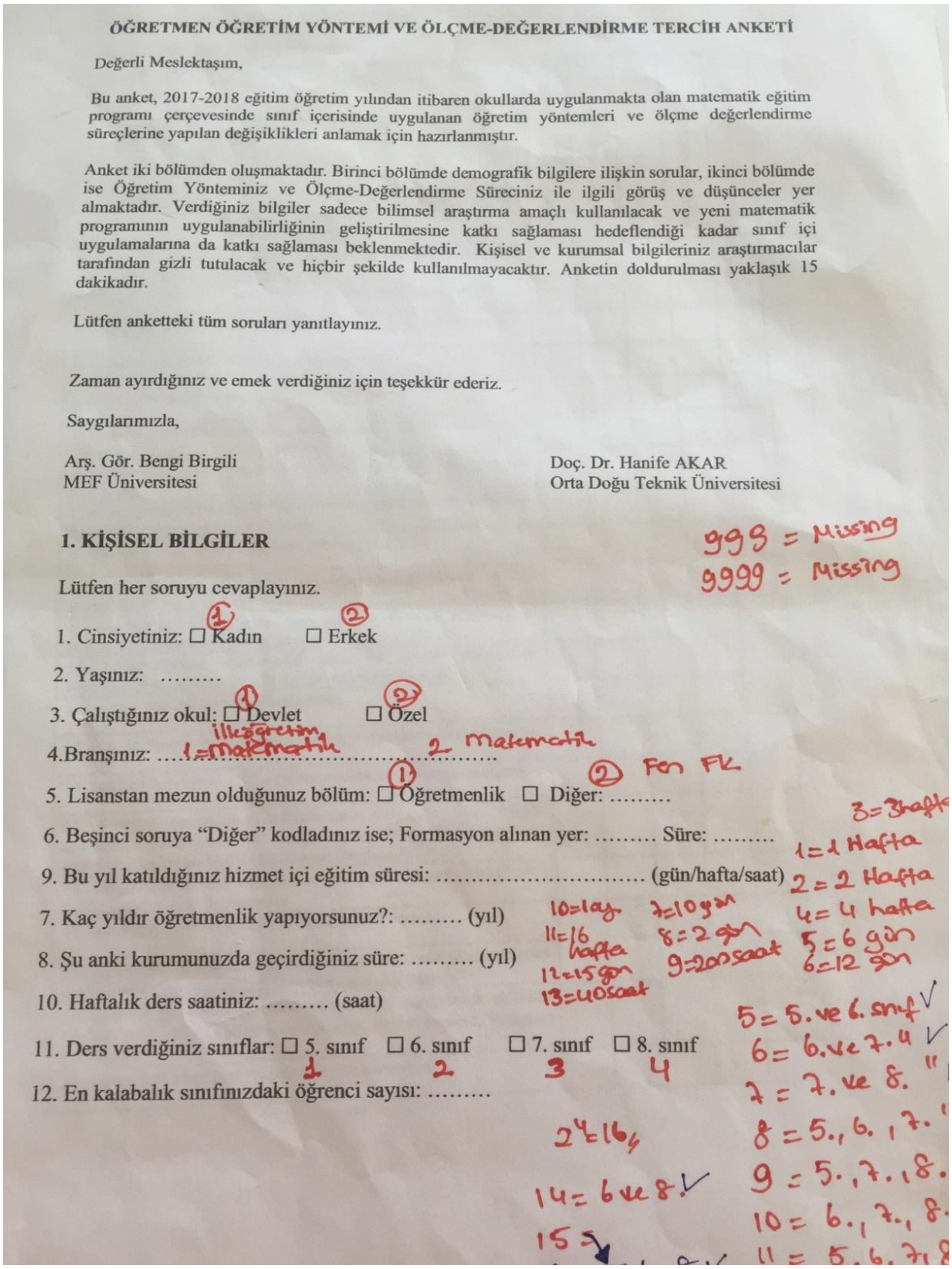
TÜRKÇE	ENGLISH
Teaching Methods	
1. programın amacında bir değişiklik olmadığını düşünüyorum.	1. I think there is no change in the purpose of the curriculum.
2. programın felsefesinde bir değişiklik olmadığını düşünüyorum.	2. I think there is no change in the philosophy of the curriculum.
3. kazanım sayılarının azaldığını fark ediyorum.	3. I have noticed the decrease in the number of learning outcomes.
4. konu içeriklerinin zenginleştiğini düşünüyorum.	4. I think the content of the subject has been enriched.
5. sınıf içi öğretimlerde somut materyal (örn. matematiksel nesne) kullanmak bana çok yardımcı oluyor.	5. Using concrete materials (e.g. mathematical objects) during classroom teaching helps me a lot.
6. öğrencilerin yaratıcılıklarına fırsat sağlayan aktiviteler yapmak tercihimdir.	6. I prefer doing activities that provide opportunities for student creativity.
7. Milli Eğitim Bakanlığının Matematik kaynak kitabının içeriğinde değişiklik olmadığını düşünüyorum.	7. I think there is no change in the content of the Mathematics resource books of the Ministry of National Education.
8. öğretmen el kitabını kullanmak tercihimdir.	8. I prefer using the teacher's handbook.
9. sınıf içi öğretim yöntemimi öğrencilerimi aktif kılacak şekilde değiştiririm.	9. I change my in-class teaching method to make my students active.
10. grupla öğretim tekniklerini (örn. işbirlikli öğrenme, düşün-eşleş-paylaş vs.) kullanmayı tercih ederim.	10. I prefer using group teaching methods (e.g. cooperative learning, think-pair-share etc.).
11. öğrencilerime; matematiği etkin bir şekilde keşfederek öğrenmelerini sağlayacak şekilde bir ders tasarlıyorum.	11. I design lessons that enable my students to learn by exploring mathematics effectively.
12. sadece anlatım tekniğini kullanırım.	12. I only use direct instruction.
13. ders öncesinde öğrencilerin hazırbulunuşluklarını test ediyorum.	13. Before the lesson, I check the students' readiness.
14. sınıf içi öğretimde eğitim teknolojilerini kullanmaya çalışırım.	14. I try to use educational technologies when teaching in-class.
15. öğretim esnasında yapılandırmacı yaklaşım (örn. bilgiyi araştırma, yorumlama ve analiz etme, düşündürme sürecini geliştirme vs.) tekniklerini	15. I prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.).

TÜRKÇE	ENGLISH
Teaching Methods	
kullanmak tercihimdir.	
16. farklı soru sorma tekniklerini (Neden?, Niçin?, Nasıl? vb.) kullanmaya ihtiyaç duyuyorum.	16. I feel the need to use different questioning techniques (e.g. Why? How? etc.)
17. anlattığım konuya günlük hayattan örnek veririm.	17. I give examples from daily life while teaching a topic.
18. öğrencilere üzerinde çalıştıkları konu hakkında düşünmelerini sağlayan bir öğrenme ortamı tasarlamayı tercih ederim.	18. I prefer designing a learning environment that makes students think about the topic they work on.
19. öğrencilerin kendi öğrenme süreçlerinde sorumluluk aldıkları öğretim tekniklerini (gösteri, soru-cevap, beyin fırtınası, tartışma) kullanırım.	19. I use instructional techniques that require students to take responsibility for their learning (e.g. demonstration, question-answer, brainstorming, discussion).
20. öğrencileri araştırma yapmaya teşvik ediyorum.	20. I encourage students to do research.

TÜRKÇE	ENGLISH
Measurement-Evaluation	
1. önceki uygulamalara göre ölçme-değerlendirme sürecinde değişiklik yapıyorum.	1. I make changes in the measurement and evaluation process compared to the previous implementations.
2. sınavlarımı çevrimiçi kaynaklardan (forum, paylaşım siteleri vb.) indirerek uyguluyorum.	2. I administer examinations based on downloaded online sources (e.g. forums, websites etc.).
3. çoktan seçmeli sorular içeren ölçme araçları kullanıyorum.	3. I use measurement tools that include multiple-choice items.
4. dersin kazanımlarını ölçmek için biçimlendirici değerlendirme (formatif değerlendirme) kullanıyorum.	4. I use formative assessment to measure course learning outcomes.
5. çoktan seçmeli ve kısa cevap gerektiren karma sınav soruları hazırlıyorum.	5. I prepare examinations that include a mixture of multiple-choice and short-answer items.
6.günlük hayattaki problemleri çözmeye yönelik sorular soruyorum.	6. I ask problem-solving items related to real life problems.
7.öğrencilerin dönem sonunda performanslarını gösterecekleri öğrenci portfolyo dosyalarını kullanıyorum.	7. I use portfolio that will enable the students to show their performances at the end of the term.
8. kısa sınavları (quiz) uyguluyorum.	8. I apply quizzes.
9.sınıf içi sınavlarımda açık uçlu soruları kullanıyorum.	9. I use open-ended items in my in-class examinations.

TÜRKÇE	ENGLISH
Teaching Methods	
10. öğretim sırasında soru-cevap tekniğini kullanıyorum.	10. I use the question-answer technique in my teaching.
11. öğrencilere sınavlardaki soru türleri hakkında tercih hakkı tanıyorum.	11. I give students choice to choose which item types they want to be included in their examinations.
12. sınavlarda öğrencilerin işlemsel becerilerini (prosedürel) kullanacakları soru türlerini tercih ediyorum.	12. I prefer item types that require students to use procedural skills in the examinations.
13. soruları öğrencilerin üst düzey bilişsel becerilerini (ne bildiğini bilme, düşüncenin farkındalığı) kullanacakları soru türlerinden (PISA, TIMSS gibi uluslararası sınav soruları gibi) seçiyorum.	13. I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought).
14. sınıf içi sınavlarımda hem açık uçlu, hem de çoktan seçmeli soruları birlikte kullanıyorum.	14. I use open-ended and multiple-choice items together in my in-class examinations.
15. uyguladığım sınıf içi sınavların sayısını öğrencilerle birlikte belirliyorum.	15. I determine the number of in-class examinations to be administered together with the students.

O. CODING PROCESS OF THE TMMESP-Q



ÖĞRETMEN ÖĞRETİM YÖNTEMİ VE ÖLÇME-DEĞERLENDİRME TERCİH ANKETİ

Değerli Meslektaşım,

Bu anket, 2017-2018 eğitim öğretim yılından itibaren okullarda uygulanmakta olan matematik eğitim programı çerçevesinde sınıf içerisinde uygulanan öğretim yöntemleri ve ölçme değerlendirme süreçlerine yapılan değişiklikleri anlamak için hazırlanmıştır.

Anket iki bölümden oluşmaktadır. Birinci bölümde demografik bilgilere ilişkin sorular, ikinci bölümde ise Öğretim Yönteminiz ve Ölçme-Değerlendirme Süreciniz ile ilgili görüş ve düşünceler yer almaktadır. Verdiğiniz bilgiler sadece bilimsel araştırma amaçlı kullanılacak ve yeni matematik programının uygulanabilirliğinin geliştirilmesine katkı sağlaması hedeflendiği kadar sınıf içi uygulamalarına da katkı sağlaması beklenmektedir. Kişisel ve kurumsal bilgileriniz araştırmacılar tarafından gizli tutulacak ve hiçbir şekilde kullanılmayacaktır. Anketin doldurulması yaklaşık 15 dakikadır.

Lütfen anketteki tüm soruları yanıtlayınız.

Zaman ayırdığınız ve emek verdiğiniz için teşekkür ederiz.

Saygılarımızla,

Arş. Gör. Bengi Birgili
MEF Üniversitesi

Doç. Dr. Hanife AKAR
Orta Doğu Teknik Üniversitesi

1. KİŞİSEL BİLGİLER

999 = Missing
9999 = Missing

Lütfen her soruyu cevaplayınız.

1. Cinsiyetiniz: ☐ Kadın ☐ Erkek
2. Yaşınız: ………
3. Çalıştığınız okul: ☐ Devlet ☐ Özel
4. Branşınız: ………………………………………
5. Lisanstan mezun olduğunuz bölüm: ☐ Öğretmenlik ☐ Diğer: ………
6. Beşinci soruya "Diğer" kodladınız ise; Formasyon alınan yer: ……… Süre: ………
9. Bu yıl katıldığınız hizmet içi eğitim süresi: ………………………… (gün/hafta/saat)
7. Kaç yıldır öğretmenlik yapıyorsunuz?: ……… (yıl)
8. Şu anki kurumunuzda geçirdiğiniz süre: ……… (yıl)
10. Haftalık ders saatiniz: ……… (saat)
11. Ders verdiğiniz sınıflar: ☐ 5. sınıf ☐ 6. sınıf ☐ 7. sınıf ☐ 8. sınıf
12. En kalabalık sınıfınızdaki öğrenci sayısı: ………

P. TEACHER-MADE EXAMINATIONS RELATED TO PARTICIPATING SCHOOLS

Table 1. Number of teacher-made examinations related to middle schools

School Name	School Type	Examinations	*f*
School 1	Public School	1. 2016-2017 1st semester 1st exam	14
		2. 2016-2017 2nd semester 1st exam	17
		3. 2017-2018 1st semester 1st exam	16
		4. 2017-2018 1st semester 2nd exam	18
		5. 2017-2018 1st semester 3rd exam	20
		6. 2017-2018 2nd semester 1st exam	12
		7. 2017-2018 2nd semester 2nd exam	20
	Total		11
School 2	Public School	1. 2017-2018 1st semester 1st exam	17
		2. 2017-2018 1st semester 2nd exam	18
		3. 2017-2018 1st semester 3rd exam	16
		4. 2017-2018 2nd semester 1st exam	15
		5. 2017-2018 2nd semester 2nd exam	21
	Total		87
School 3	Public School	1. 2017-2018 1st semester 1st exam	20
		2. 2017-2018 2nd semester 1st exam	25
	Total		45
School 4	Public School	1. 2016-2017 1st semester 1st exam	19
		2. 2016-2017 2nd semester 1st exam	19
		3. 2017-2018 1st semester 1st exam	20
		4. 2017-2018 1st semester 2nd exam	19
		5. 2017-2018 1st semester 3rd exam	20
		6. 2017-2018 2nd semester 1st exam	20
	Total		11
School 5	Private School	1. 2018-2019 1st semester 2nd exam	14
	Total		14
GRAND TOTAL		21	380

Q. ITEM-LEVEL STATISTICS FOR TM AND MES SCALES OF TMMESP-Q

Table Q.1. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (*N* =344 Mathematics Teachers)

Scale Items	*M*	*SD*	*f* (%)				
			Strongly disagree	Disagree	Undecided	Agree	Strongly agree
1. I think there is no change in the purpose of the curriculum.	3.25	1.13	20 (5.8%)	94 (27.3%)	49 (14.2%)	143 (41.6%)	37 (10.8%)
2. I think there is no change in the philosophy of the curriculum.	3.29	1.13	27 (4.9%)	92 (26.7%)	54 (15.7%)	136 (39.5%)	44 (12.8%)
3. I have noticed the decrease in the number of learning outcomes.	4.06	1.07	10 (2.9%)	37 (10.8%)	32 (9.3%)	115 (33.4%)	145 (42.2%)
4. I think the content of the subject has been enriched.	2.84	1.18	35 (10.2%)	129 (37.5%)	71 (20.6%)	70 (20.3%)	34 (9.9%)
5. Using concrete materials (e.g. mathematical objects) during classroom teaching helps me a lot.	4.39	.91	9 (2.6%)	16 (4.7%)	20 (5.8%)	101 (29.4%)	197 (57.3%)
6. I prefer doing activities that provide opportunities for student creativity.	4.36	.82	5 (1.5%)	12 (3.5%)	26 (7.6%)	124 (36%)	173 (50.3%)
7. I think there is no change in the content of the Mathematics resource books of the Ministry of National Education.	3.59	1.22	17 (4.9%)	65 (18.9%)	51 (14.8%)	117 (34%)	93 (27%)

Table Q.1. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (continued)

8. I prefer using the teacher's handbook.	3.17	1.22	26 (7.6%)	90 (26.2%)	69 (20.1%)	92 (26.7%)	56 (16.3%)
9. I change my in-class teaching method to make my students active.	4.48	.71	3 (0.9%)	8 (2.3%)	19 (5.5%)	116 (33.7%)	197 (57.3%)
10. I prefer using group teaching methods (e.g. cooperative learning, think-pair-share etc.).	4.20	.85	3 (0.9%)	20 (5.8%)	36 (10.5%)	142 (41.3%)	141 (41%)
11. I design lessons that enable my students to learn by exploring mathematics effectively.	4.28	.66	2 (0.6%)	7 (2%)	29 (8.4%)	176 (51.2%)	128 (37.2%)
12. I only use direct instruction.	2.10	1.10	96 (27.9%)	147 (42.7%)	27 (7.8%)	58 (16.9%)	10 (2.9%)
13. Before the lesson, I check the students' readiness.	4.27	.80	4 (1.2%)	16 (4.7%)	18 (5.2%)	159 (46.2%)	142 (41.3%)
14. I try to use educational technologies when teaching in-class.	4.33	.79	2 (0.6%)	16 (4.7%)	23 (6.7%)	136 (39.5%)	165 (48%)
15. I prefer using the constructivist approach techniques when teaching (e.g. research, interpret and analyze information, improve the thinking process etc.).	4.41	.68	3 (0.9%)	6 (1.7%)	24 (7%)	140 (40.7%)	171 (49.7%)

Table Q.1. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (continued)

16.	I feel the need to use different questioning techniques (e.g. Why? How? etc.)	4.60	.56	4 (1.2%)	2 (0.6%)	9 (2.6%)	114 (33.1%)	212 (61.6%)
17.	I give examples from daily life while teaching a topic.	4.64	.56	4 (1.2%)	3 (0.9%)	8 (2.3%)	101 (29.4%)	227 (66%)
18.	I prefer designing a learning environment that makes students think about the topic they work on.	4.41	.65	4 (1.2%)	5 (1.5%)	17 (4.9%)	151 (43.9%)	165 (48%)
19.	I use instructional techniques that require students to take responsibility for their learning (e.g. demonstration, question-answer, brainstorming, discussion).	4.32	.76	3 (0.9%)	8 (2.3%)	37 (10.8%)	135 (39.2%)	161 (46.8%)
20.	I encourage students to do research.	4.29	.73	3 (0.9%)	6 (1.7%)	37 (10.8%)	148 (43%)	150 (43.6%)

In addition to TM items, item level analysis were held for MES items.

Table Q.2. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (*N* =344 Mathematics Teachers)

Scale Items	*M*	*SD*	*f*(%)				
			Never	Seldom	Sometimes	Often	Always
1. I make changes in the measurement and evaluation process compared to the previous implementations.	3.46	.95	10 (2.9%)	31 (9%)	144 (41.9%)	107 (31.1 %)	51 (14.8%)
2. I administer examinations based on downloaded online sources (e.g. forums, websites etc.).	2.55	1.22	90 (26.2 %)	82 (23.8%)	82 (23.8%)	72 (20.9 %)	18 (5.2%=
3. I use measurement tools that include multiple-choice items.	3.44	1.03	18 (5.2%)	41 (11.9%)	101 (29.4%)	138 (40.1 %)	43 (12.5%)
4. I use formative assessment to measure course learning outcomes.	3.45	.97	11 (3.2%)	40 (11.6%)	124 (36%)	119 (34.6 %)	48 (14%)
5. I prepare examinations that include a mixture of multiple-choice and short-answer items.	3.70	1.09	16 (4.7%)	38 (11%)	70 (20.3%)	130 (37.8 %)	87 (25.3%)
6. I ask problem-solving items related to real life problems.	3.96	.85	2 (0.6%)	15 (4.4%)	76 (22.1%)	154 (44.8 %)	96 (27.9%)

Table Q.2. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (Continued)

7.	I use portfolio that will enable the students to show their performances at the end of the term.	2.82	1.21	56 (16.3 %)	89 (25.9%)	96 (27.9%)	68 (19.8 %)	35 (10.2%)
8.	I apply quizzes.	3.67	1.07	9 (2.6%)	40 (11.6%)	95 (27.6%)	107 (31.1 %)	91 (26.5%)
9.	I use open-ended items in my in-class examinations.	3.97	.93	8 (2.3%)	21 (6.1%)	61 (17.7%)	146 (42.4 %)	107 (31.1%)
10.	I use the question-answer technique in my teaching.	4.27	.68	1 (0.3%)	3 (0.9%)	33 (9.6%)	173 (50.3 %)	131 (38.1%)
11.	I give students choice to choose which item types they want to be included in their examinations.	2.70	1.18	60 (17.4 %)	99 (28.8%)	99 (28.8%)	55 (16%)	29 (8.4%)
12.	I prefer item types that require students to use procedural skills in the examinations.	3.76	.82	4 (1.2%)	18 (5.2%)	91 (26.5%)	170 (49.4 %)	58 (16.9%)

Table Q.2. Item-Level Descriptive Statistics for the Main Data of the 20 TM Items in the TMMESP-Q (Continued)

13. I choose item types that appear in international examinations (such as PISA, TIMSS) to enable students to use their high-level cognitive skills (e.g. metacognition, awareness of thought).	3.14	1.12	25 (7.3%)	77 (22.4%)	112 (32.6%)	84 (24.4 %)	46 (13.4%)
14. I use open-ended and multiple-choice items together in my in-class examinations.	4.04	1.05	15 (4.4%)	22 (6.4%)	39 (11.3%)	136 (39.5 %)	132 (38.4%)
15. I determine the number of in-class examinations to be administered together with the students.	2.09	1.27	160 (46.5 %)	76 (22.1%)	47 (13.7%)	39 (11.3 %)	22 (6.4%)

R. THINK-ALOUD PROCESS PROTOCOL FOR FIFTH GRADE STUDENTS

1. Öğrenci cevabı doğru mu? Yanlış mı?
2. Zorlanma seviyesi 1 (Kolay), 2(Orta), 3(Zor)
3. Think aloud ve ses kaydı bize neler söylüyor? (Üstbiliş ve Duygu Boyutları)
    a. Bilişsel strateji davranışı vs eye-tracking sistemi verisi
    b. Öz-kontrol davranışı vs eye-tracking sistemi verisi
    c. Sorulara kaç kere dönüş yaptı? Kaç kere odaklandı? Dönüp dönüp bakma ve çözme davranışı var mı?
4. Çözüm yoluna bak.
    a. Bilişsel stratejiyi kullandı mı? Kullanmadı mı?
    b. Öz-kontrol becerisini kullandı mı? Kullanmadı mı?

*Eylemleri-Söylemleri arasındaki benzerlikler ve farklılıklar.

5. Duygular: E4 raporu ne gösteriyor? (Ses, Görsel Alan notu)
    a. Çaba hissi
    b. Endişe hissi
    c. Başlangıç ve bitiş dk sına gidip E4 verileri arasından EDA ve BVP screenshot koy.

Başlangıç dksı-Bitiş dksı (… ses kaydından bakıldı)

Soru 1. Ali kedisinin ağırlığını bulmak istiyor.
Önce kendisi tartılıyor ve **57 kg** geliyor. Daha sonra kedisini kucağına alarak tartıya çıkıyor ve ağırlığı **62 kg** olarak kaydediyor.

Kedinin ağırlığı kaç kilogramdır?

A) 119
B) 15
C) 5
D) 9

1. Doğru Cevap: C
   Öğrenci Cevabı: (puan)
2. Zorlanma seviyesi:

3. Think aloud ve ses kaydı bize neler söylüyor? (Üstbiliş ve Duygu Boyutları)
    a. Bilişsel strateji davranışı vs eye-tracking

S. CODEBOOK FOR NEUROEDUCATION PROCESS

Feature	Code	Code Explanation
Feature-1: Item Type	Code-MC	Multiple-choice items
	Code-OE	Open-ended items
Feature-2: Item Correctness	Code-1	Correctly solved items
	Code-0	Incorrectly solved items
Feature-3: Difficulty Level	Code-1	The participant expressed that an item responded and solved easily. Items must be correctly solved. Also it should be validated by eye-tracking, hand-writing and researcher's note-taking relying on observation. GSR EDA values were also checked.
	Code-2	The participant expressed that an item responded and solved moderately. Items may be solved correctly or the participant made a mistake. Also it should be validated by eye-tracking, hand-writing and researcher's note-taking relying on observation. GSR EDA values were also checked.
	Code-3	The participant expressed that an item either was not solved or responded and solved in difficulty. The participant would give up while responding. Also it should be validated by eye-tracking, hand-writing and researcher's note-taking relying on observation. GSR EDA values were checked.
Feature-4: Number of Solution Process	Code-1	Number of solution process step = 1 (e.g., addition)
	Code-2	Number of solution process step = 2 (e.g., counting and multiplication)
	Code-3	Number of solution process step = 3 (e.g., multiplication, counting and imagination)
	Code-4	Number of solution process step = 4 (e.g., rereading question root, division, subtraction, and transformation)
Feature-5: Electrodermal Activity (EDA)	Code-YES	When the participant responded and solved the item easily and EDA confirmed that or when the participant responded and solved the item in difficulty and EDA confirmed that.
	Code-NO	When the particant's expression conflicted with EDA value (e.g., when the participant responded and solved the item easily. Nevertheless, EDA confirmed that he/she had difficulty.

Feature	Code	Code Explanation
	Code-NA (i.e., Not Applicable)	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-6: Cognitive Strategy (CS1): Organization of knowledge	Code- Qroot	The participant read the question root.
	Code- Subtraction	The participant did subtraction.
	Code- Mental Operation	The participant did a mental operation.
	Code- Multiplication	The participant did multiplication.
	Code- Others (e.g., equation, algebra)	The participant used different equations to solve the item.
	Code- Counting	The participant mentally counted the shapes, boxes, figures.
	Code- Shape	The participant examined the shapes given in the item.
	Code- Alternative	The participant checked the values given in the alternatives of the items to solve it.
	Code- Elimination	The participant could use an alternative elimination testing strategy to find the correct solution.
	Code- Operation	The participant expressed in general how to use operations.
	Code- Transform	The participant transformed metric measurement units.
	Code- Drawing	The participant drew a given shape onto the paper or a different perspective of the shape.
	Code- Addition	The participant did addition.
	Code- Reasoning	The participant explained the meaning of question root, used reasoning skills before operations.
	Code- Sorting	The participant sorted between given value/image/shape in terms of perspective or its features.
	Code- Division	The participant did division.
	Code- İmagination	The participants used imagination skills, perspectives, and pictures of oneself into items.
	Code- Part-whole	The participant used part-whole relationships in fractions.

Feature	Code	Code Explanation
	Code-NA	When the participant was not able to use any codes, the Organization of Knowledge.
Feature-7: Number of Steps	Code- Step_N	The total number of steps from Feature-6 while responding to the item.
Feature-8: CS2: Rereading	Code- YES	The participant reread the item to understand the given values and/or to summarize it.
	Code- NO	When the participant was not able to use rereading skill.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-9: CS3: More than one strategy	Code- YES	The participant declared he/she used more than one strategy to solve the item.
	Code- NO	The participant declared he/she did not use more than one strategy to solve the item.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-10: CS4: Re-expression (i.e., Reexplanation)	Code- YES	The participant indicated to re-express or re-explain the given information with his/her own words.
	Code- NO	The participant indicated not being able to re-express or re-explain the given information with his/her own words.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-11: Number of Looking Back and Forth (LB)	Code- LB_Number	When the participant looked back and forth around the item. His/her Area of Interest was numbers.
	Code- LB_Qroot	When the participant looked back and forth around the item. His/her Area of Interest was question root.
	Code- LB_Alternative	When the participant looked back and forth around the item. His/her Area of Interest was alternatives.
	Code- LB_Shape	When the participant looked back and forth around the item. His/her Area of Interest was shaped(s).
	Code- LB_Other (e.g., between different area)	When the participant looked back and forth around the item. His/her Area of Interest was between different areas.

Feature	Code	Code Explanation
	Code- LB_NumberTotal	The total number of looking back and forth areas.
Feature-12: Self-Checking (SC1): Checking answer	Code- YES	The participant expressed that he/she checked his/her answer while responding.
	Code- NO	The participant expressed that he/she did not check his/her answer while responding.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-13: Self-Checking (SC2): Process control	Code- YES	The participant expressed that he/she controls the overall process of their solution, answer etc.
	Code- NO	The participant expressed that he/she did not control the overall process of their solution, answer etc.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-14: Self-Checking (SC3): Finding Error	Code- YES	The participant expressed that he/she found an error and corrected it in the solution.
	Code- NO	The participant expressed that he/she did not find any error and did not correct it in the solution.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-15: Self-Checking (SC4): Ask question to stay on track	Code- YES	The participant expressed that he/she did self-talk and asked whether he/she stays on track while responding.
	Code- NO	The participant expressed that he/she did not use self-talk and did not ask whether he/she stays on track while responding.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-16: Worry (W1): Performance	Code- YES	The participant expressed that he/she worried about the performance he/she showed while responding.
	Code- NO	The participant expressed that he/she did not worry about the performance he/she showed up while responding.

Feature	Code	Code Explanation
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-17: Worry (W2): How Confident	Code- 1	The participant expressed feelings of low self-confidence (0-39%).
	Code- 2	The participant expressed a feeling of moderate self-confidence (40-69%).
	Code- 3	The participant expressed a feeling of high self-confidence (70-100%).
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-18: Worry (W3): Type of feeling	Code- 1	The participant expressed and showed a positive feeling toward the item (e.g., calm, easy, comfort).
	Code- 2	The participant expressed and showed negative feelings toward the item (e.g., worry, anxiety, stress, confusion).
	Code- 3	The participant did not express or shows positive or negative feeling (e.g., Nötr)
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-19: Effort (E1): Concentration	Code- 1	The participant expressed he/she had low concentration while responding.
	Code- 2	The participant expressed he/she had moderate concentration while responding.
	Code- 3	The participant expressed he/she had high concentration while responding.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.
Feature-19: Effort (E2): Amount of Effort	Code- 1	The participant expressed he/she spent low effort while responding to the item.
	Code- 2	The participant expressed he/she spent middle effort while responding to the item.
	Code- 3	The participant expressed he/she spent high effort while responding to the item.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol

Feature	**Code**	**Code Explanation**
		interview questions.
Feature-20: Effort (E3): Give up	Code- YES	The participant asked to give up or expressed he/she thought about giving up while responding.
	Code- NO	The participant expressed he/she did not think about giving up while responding.
	Code- NA	When the participant was not able to respond to the item or did not answer the think-aloud protocol interview questions.

T. CURRICULUM VITAE

Surname, Name: Birgili, Bengi

EDUCATION

Degree	Institution	Year of Graduation
Visiting Scholar	University of Vienna/ Universität Wien, Doctoral Program in Philosophy, Education	Spring 2019
M.Sc.	METU Educational Sciences, Curriculum & Instruction	2014
B.Sc.	Boğaziçi University (Mathematics Education)	2011
High School	Çapa Anatolian Teacher Training High School	2006

FOREIGN LANGUAGES

Advanced English, Beginner German

PUBLICATIONS

1. Aydın, U., & B. Birgili "Assessing Mathematical Higher-Order Thinking Skills: An Analysis of Turkish University Entrance Examinations" Educational Assessment. (In Press, 2022).

2. Birgili, B., & Ö. Demir. "An Explanatory Sequential Mixed-Method Research on the Full-Scale Implementation of Flipped Learning in The First Years of the World's First Fully Flipped University: Departmental Differences" Computers & Education. 176, 104352 (2022).

3. Taylan R. D., Z. Tunç-Pekkan, U. Aydın, & B. Birgili. "Teacher Educators in K-12 Classrooms: How to Nurture Professional Development and Research" Journal of Higher Education Theory and Practice. 22(1), (2022).

4. Birgili, B., F. N. Seggie, & E. Oğuz. "The Trends and Outcomes of Flipped Learning Research Between 2012 and 2018: A Descriptive Content Analysis" Journal of Computers in Education. 8(1), (2021).

5. Birgili, B. "Comparative Reflection on Best Known Design Models: Notes from the Field" Current Issues in Emerging eLearning. the Center for Innovation and Excellence in eLearning, and sponsored by the College of Advancing and Professional Studies at the University of Massachusetts Boston, USA. 6(1), Article 5. Available at: https://scholarworks.umb.edu/ciee/vol6/iss1/5 (2019).

6. Birgili, B., F. N. Seggie, & Z. Kızıltepe "Investigating the Relationship between Volitional Strategies and Academic Achievement in a Flipped Learning Environment" Croatian Journal of Education. 21(1), 345-375 (2019).

7. Aydın U., Z. Tunç-Pekkan, R. D. Taylan, B. Birgili, & M. Özcan. "Impacts of a University-School Partnership on Middle School Students' Fractional Knowledge." The Journal of Educational Research. 111(2), 151-162 (2018).

8. Birgili, B., Z. Kızıltepe & F. N. Seggie. "Teaching Method Preferences of Teachers: The Cooperative Teaching Method" World Studies in Education. 17(2), 37-52 (2016).

U. TURKISH SUMMARY/TÜRKÇE ÖZET

ARAŞTIRMANIN TÜRKÇE KISA ÖZETİ

1. GİRİŞ

Okullar çocuklarımızı hayata hazırlar. Çocuklar ilk kez ailelerinden farklı bir ortama girer ve bunun anlamı ilk kez kuralları olan bir sosyal gruba dahil olmaya çalışması, bu kuralları öğrenmesi, ondan beklenen bilgi, beceri ve tutumlarla donatılması demektedir. Tahmin edileceği gibi hazır bulunuşluğa sahip çocuklar olabileceği gibi okula hazır olmayan çocuklar da bu sürece dahil olacak demektir. Çocukların hazır bulunuşluk düzeyleri birbirinden farklı, temel beceri kazanım düzeyleri, ilgi ve tutumları da birbirinden farklı olacaktır. Farklı sosyal ve ekonomik düzeylere sahip olması onların okul ortamına başlangıç seviyesinde farklı şekillerde uyum sağlamaya çalışması demektir. Öğrenim hedefine ilişkin yetenekleri farklıdır. Öğrenmeye gösterdikleri istek dereceleri farklıdır. Yeterli zaman ayrıldığında, onlara tam öğrenme fırsatı verildiğinde sadece bilgi düzeyinde değil, uygulama yaparak beceri düzeyinde kavrandığı zaman öğrenmeleri gerçekleştiği bilinmektedir (APPG, 2017; Hansen vd., 2003; Nunn, 2014).

Bu kadar farklılıklar içerisinde önemli olan ise nitelikli bir eğitim öğretim ortamı sağlanarak çocuklarımızın mutlu bireyler olarak yetişmelerine imkân sağlamaktır. İşte okulun en büyük rolü budur. Eğitim sisteminin hedefi, sağlıklı ve mutlu bireyler yetişmesine en uygun okul ortamını, sınıf iklimi, öğretmen ve öğrenme süreçlerinin tasarlanmasına imkân sağlamaktır. Böylelikle çocuklar dünyayı ve dünyadaki gelişmeleri anlayan, bilgileri özümseyen, yorumlayan, dünya ile yarışacak düzeyde bilgi ve donanıma sahip bireyler olabilecektir. Ülke olarak gelecek nesiller diye tabir ettiğimiz çocuklarımız için kurduğumuz hayal budur. İyi yetişmiş insan gücünün de temeli okullarda atılır. Çünkü eğitim öğretim sürecinde öğrenciler bir bütün olarak ele alınır. Onlar sadece geleceğin bireyleri değil, bilim insanlarıdır.

Okulu okul yapan en hayati noktalardan biri de öğretmenlerdir (Darling-Hammond ve Lieberman, 2012; Ticha ve Hospesova, 2006). Özellikle Türkiye bağlamında çocuğun kişiliğini şekillendiren, yeni bilgi, beceri ve değerler kazanmasına vesile olan aileden sonra ilk kişi öğretmendir. Öğretmenlerin niteliği, deneyimi, yeterlilikleri öğretimin niteliği ile ilişkilidir (Darling-Hammond ve Lieberman, 2012). Öğretimde nitelik, öğrenme ve öğretme süreçlerinin hedefine ulaşmak için en uygun deneyim durumlarını belirleme ve öğrencinin gereksinimine göre (onun ön bilgi, beceri ve tutumlarını dikkate alarak) düzenleme yaparak sağlanabilir. Okulun sahip olduğu donanım ve yeterlilikleri çocuğun yararına kullanabilen, imkansızlıklar bağlamında yine kendi donanım ve araçlarını öğretme öğrenme sürecinin yararına tasarlamayı bilen öğretmendir, onun niteliğidir. Nitelikli bir öğretim süreci için meslek öncesi fakültelerde ya da mesleki gelişim aşamasında (hizmetiçi eğitim süreci gibi) öğretmenlerimize ihtiyaçları doğrultusunda profesyonel gelişim desteği verilmelidir (Darling-Hammond ve Lieberman, 2012; Darling-Hammond, 2017; Darling-Hammond ve Oaks, 2019). Çünkü iyi yetişmiş *nitelikli bir öğretmen* öğrencilerini de aynı şekilde nitelikli, dünyayı takip eden, ufukları açık minik bilim insanları olarak yetiştirecektir. Okulun bir amacı da yüzde yetmiş oranda çocuğu bir sonraki kademeye, bir sonraki sınıf seviyesine hazırlamaktır. Bu hazırlama süreci planlı, hedefleri belli sistematik bir şekilde planlandığı sürece işe yarar. Hedeflerin belirlenmiş olması, süreçteki geri dönütler ve kontroller sayesinde öğrenme hedeflerine ne kadar ulaşıldığını, öğrencilerin bir beceriyi onu bir üst sınıfa taşıyacak şekilde kazanıp kazanamadığını kontrol etmemize yarar.

Eğitim programı ve öğretim kavramları ayrı olarak incelense de birbirine bağımlı kavramlardır. Eğitim programını planlanmış öğrenme yaşantıları olarak düşünürken öğretimi ise yöntem bilgisi, öğretme stratejisi, yöntem ve teknikleri, kısacası öğretme rolü olarak düşünebiliriz. Bu ilişkide verilere dayalı olarak öğrencilerin gelişim özelliklerine göre uygun program ve öğretim hazırlamak gerekmektedir. Öğretim öğrenme sürecinin bir parçasıdır. Öğretim yaklaşımı eğitim programlarının içerik boyutunun dağıtım sistemini temsil eden bir süreç olmakla birlikte öğretimin kaynağı, yürütücüsü bir de öğretim yönteminin ögeleri bu süreç ile ilgili temel yapıları kapsamaktadır. Bu iki kavram birbiriyle yaklaşım ve tasarım olarak uyumlu olmak zorundadır. Bu yapıların test edildiği, öğretmene yöntemi konusunda, öğrenciye de hedeflerine ne kadar ulaştığı konusunda dönüt veren en önemli öge de öğretimin ölçme-değerlendirme sürecidir.

Ölçme-değerlendirme sürecinde sınıf içi sınavların öğrenci başarısının izlenmesinde ve kanıta dayalı kararlar alınmasında payı büyüktür. Çünkü öğrencinin öğrenme kazanımlarından ve günlük planlarda hedeflediğimiz kazanımlardan kaçına ne düzeyde ulaşıldığını sınıf içi sınavlarda görebiliriz. Özellikle ülkemizde korkulan bir ders olan, ulusal ve uluslararası sınavlarda en düşük başarı sergilenen branşlardan biri matematiktir (OECD, 2016, 2018b; PISA, 2015). Matematik konusunda özellikle sınıf içi sınavların yansıması bize öğretmenin öğretim tasarım süreci, hedefleri, ölçme-değerlendirme kalitesi hakkında geri bildirim sağlamaktadır. Ne yazık ki alan yazına baktığımızda öğretmenlerin sınıfiçi sınav sorularını kendi bilgi becerileriyle hazırlamadığı genellikle çeşitli kaynaklardan kopyala yapıştır yaptıkları, çoğunlukla internet kaynak olarak kullandıkları belirlenmiştir. Hatta otantik öğretmen yapımı sınavların kalitesi ile ilişkili olduğu bilinen bilgi ve bilişsel boyutları öğretmenlerin yoğunluklu olarak düşünmeden sınıfiçi sınavları hazırladıkları da vurgulanmıştır. Özetle, sınıfiçi sınavlar üzerine yapılan araştırmalar (Aldım, 2010; Birgili vd., 2021; Çevik, 2009; Çağatay ve Kılıç, 2019; Delil ve Özcan, 2019; Guvendir ve Ozkan, 2021; Hartell ve Strimel, 2019; İnci, 2014) ve sınavların psikometrik özellikleri göstermektedir ki matematik öğretmenlerinin sınıfiçi sınavlarını detaylı, ulusal ve uluslararası düzeyde incelemeye hala ihtiyaç vardır. Matematik öğretim programı kazanımlarını sınıf içi sınavlar çok kapsamlı şekilde ölçememekte ve ileriye yönelik sınav puanları ile ilişkili olması önemini bir kez daha açığa çıkarmaktadır.

Öğrencilere planlı olarak kazandırılacak bilgi, beceri, tutum ve davranışlar bütünsel bir stratejinin parçası olarak öğretim programları ile düzenlenmektedir. Türkiye'de bu kapsamda eğitim programlarının hazırlanması, geliştirilmesi ya da değiştirilmesi sürecinin eğitim sisteminin diğer alt bileşenleri ile birlikte düşünülmesi gerekmektedir. Milli Eğitim Bakanlığının taslak öğretim programlarını temel beceri ve yeterlikler çerçevesinde oluşturmuş olduğu gözlenmekte, bu kapsamda Avrupa Yeterlilikler Çerçevesi (2008), Milli Eğitim Kalite Çerçevesi ve resmi gazetede yürürlüğe giren Türkiye Yeterlilikler Çerçevesi (2016) dikkate alınarak hazırlanmış olduğu belirtilmiştir. Ancak öğretim programı geliştirme sürecinin sürekliliği açısından, önerilen programların önceki programlardan hangi noktalarda farklılaştığının; felsefe, amaç, öğretim süreci ve ölçme-değerlendirme boyutlarında yapılan değişikliklerin nedenlerinin açıklanmasına ihtiyaç vardır. İlgili alanyazın incelendiğinde, öğretmenlerin bu değişimin bir parçası olarak öğretim programı sürecinin çeşitli boyutlarında kısıtlı oynadıkları (Priestley, Edwards, Priestley & Miller, 2012) ortaya çıkmıştır. Programın amacını, içeriğini, kazanımları bilmekte ancak bunları geliştirme konusunda düşük yetkinlik gösterdikleri belirlenmiştir (Eriş ve Kılıçoğlu, 2019). Dolayısıyla 2017 yılındaki öğretim programı değişikliği sürecinden sonra öğretmenlerin sınıf içi yenilikçi eğitim yaklaşımlarını uygulamada ve ölçme-değerlendirme süreçlerinde ne tür değişiklikler yaptıklarına dair somut kanıtlar ve öğretmenlerin deneyimlerine ait güncel veriler sınırlıdır (Kerkez, 2018). Ayrıca PISA, TIMSS gibi uluslararası sınavlarda öğrencilerimiz diğer ülkeler arasında çok düşük başarı göstermektedir (MEB PISA Ön Raporu, 2019). Bu tür sınavlarda açık uçlu sorular ile de karşılaşmaktalar ancak açık uçlu soru türlerinin çözüm sürecine alışkın olmadıkları için bu tür sorularda düşünme süreçlerini yeterince yansıtamadıkları problem durumu olarak karşımıza çıkmaktadır.

Nitelikli bir öğretim ve ölçme-değerlendirme süreci için öğretmenlerin, ülkemizdeki öğretim programı değişikliğinden sonra öğretim programı değişikliğinden ne derece haberdar oldukları, adaptasyon olabilme süreçleri, öğretim programı felsefesi ve yaklaşımına uygun olarak sınıf içi öğretim yöntemlerinde ve ölçme-değerlendirme stratejilerinde ne tür değişiklikler yapmayı tercih ettikleri, sınıfiçi sınavlarını hangi bilişsel düzeylerde hazırlayabildikleri keşfedilmeyi bekleyen olgulardan birkaçıdır. Öğretmenlerin halihazırdaki niteliği ile birlikte ortaokul öğrencilerinin de üst düzey bilişsel becerileri ölçebilen yenilikçi soru türlerine hazırbulunuşlukları, bilgilerini ne düzeyde yansıtabildikleri, üt düzey bilişsel becerilerini ne düzeyde kullanabildikleri ve sorularla başbaşa kaldıklarında duygularını nasıl yansıttıkları çeşitli araştırmalarla başlamıştır. Öğretmenlerin ne bildiğini ve ne öğrettiklerini derinlemesine araştıran fakat aynı bağlamda öğrencilerin ne anladığını ve konu kapsamında yapabildiklerini anlatıp büyük resmi gösteren çok boyutlu çalışmalara ihtiyaç giderek artmaktadır. Ancak, Türkiye'deki disiplinler arası çalışmaları içeren ilgili literatürde çok az kanıt bulunmaktadır (Ünal vd., 2020; Azevedo ve Aleven, 2013; Van Gog ve Jarodzka, 2013).

Insanların iç dünyasında meydana gelen üstbilişsel bilgi ve beceriler, bu becerilerin kullanımı ve deneyimi sanıldığı kadar kolay ölçülememekte, çeşitli çalışmalara konu olmaya devam etmektedir. Psikolojik araç ölçümleriyle, fizyolojik araç ölçümleriyle duygu ve biliş içerisinde olan beceriler ölçülüp değerlendirilmekte hatta katılımcı grupları (yetişkin veya çocuk) sesli düşünme sürecine (think-aloud process) maruz bırakılarak içsel konuşmalarını yansıtması beklenmektedir. Çünkü bilgi ve becerileri bilmek her zaman bilinçli kullanmak demek olmayabilir. Bu nedenle geçerli ve güvenilir araçlarla ölçülmesi, araştırma katılımcısının araştırmacının sorusuna verdiği yanıtların ve yansıttığı cevapların doğruluğunu test etme açısından önemlidir. Karma yöntemlerin artmasıyla birlikte veri çeşitlemesini kullanma en güçlü stratejilerden biri olmaktadır (Toraman, 2021). Öğrenci ve öğretmenler üzerinde göz izleme teknolojileri, galvanik cilt tepkisi (GSR) araçları, akıllı saat araçları, grafik/bambu tabletler, GoPro yüksek teknoloji kameraları, günlük dosyaları, fizyolojik laboratuvar verileri, ekran kayıtları, sesli düşünme süreçleri, kendi kendine düşünmeyi ölçen teknolojilerini kullanmaya artan bir ilgi vardır. Kendini açıklama (re-expression) oturumları, duyguların ve yüz ifadelerinin ve söylemlerin dilsel analizi (Azevedo, 2002; Azevedo vd., 2017; Azevedo ve

Gašević, 2019; Jarodzka vd., 2017; van Gog ve Jarodzka, 2013) eğitim araştırma çalışmalarında teknolojinin ilerlemesiyle çokça kullanılmaktadır. Öğrencilerin bu yöntemlerle analizi de araştırmanın diğer bir boyutunu belirlemede yönlendirici olmuştur.

1.1.Çalışmanın Amacı

Bu çalışmanın öncelikli amacı, öğretmenleri değişimin bir parçası sayan ekolojik yaklaşım bağlamında, Türkiye'de 2017-2018 eğitim öğretim yılından beri uygulanmakta olan matematik öğretim programına göre ortaokul matematik öğretmenlerinin öğretim yöntemi tercihlerine ve ölçme-değerlendirme süreçleri ile onların öğrencilerinin açık uçlu sorularla ilgili deneyimlerine bakmaktır. Diğer bir ifadeyle, Türkiye'de eğitim politikası değişikliği sonrası ortaokul matematik öğretmenlerinin öğretim yöntemi ve ölçme-değerlendirme süreçlerinin kalitesini ekolojik bir yaklaşımla ve ortaokul öğrencilerinin farklı soru maddelerine verdikleri yanıtların kalitesini onların üstbiliş ve duygusal tepkileri bağlamında çoklubiçimli karma yöntem eşzamanlı baskın durum tasarımı (multimodal mixed methods concurrent dominant status design) kullanarak incelemektir.

1.2.Tanımlar

Öğretim Programı Değişikliği, öğretim programını bazı yönlerden farklı kılmak amacıyla, amaç ve hedefler açısından felsefesini değiştirmek, içeriğini gözden geçirmek, öğretim yöntemlerini gözden geçirmek ve değerlendirme prosedürlerini yeniden düşünmektir (Priestley vd., 2015). **Değerlendirme Değişimi** ise, ölçme, değerlendirme ve değerleme süreçlerinde ve türlerinde yeni uyarlamalar yapabilmek, öğretmenin otonomisiyle sınıf içinde farklı değerleme türleri kullanabilmesi, yeni bir teknolojinin benimsenmesini, kilit bir süreci değiştirmeyi veya değerlendirme sistemini yeniden yapılandırmayı dahil ederek temel öğretme ve öğrenme hedeflere ulaşmaktır (Carless ve Zhou, 2015).

Üstbiliş (Metacognition), en geniş çerçevede ne bildiğini bilme ve kendi düşünce sisteminin farkında olma olarak tanımlanırken (Flavell, 1979); bireylerin problemleri çözmek için stratejiler geliştirmesini ve kendi düşünce süreçleri hakkında düşünmelerini içermektedir. Bu süreç planlama, izleme, kendi kendini kontrol etme, bilişsel strateji, farkındalık gibi çeşitli alt kategorilere ayrılmıştır (O'Neil ve Abedi, 1996; O'Neil ve Brown, 1998). Üst bilişin alt becerilerinden olan **bilişsel strateji (cognitive strategy) ise** öğrencilerin istenen becerileri gerçekleştirmelerini sağlayan, içsel prosedürler geliştirdikçe performansını kolaylaştıran veya destekleyen, hedefe yönelik ve bilinçli olarak kontrol edilebilir bir süreçtir (Mcewen, Huijbregts, Ryan ve Polatajko, 2009). Diğer bir alt beceri olan **öz kontrol (self-checking),** bir göreve başlarken veya çözüm süreci içerisinde performansını kendi kendine izleme ve hatalarını farkedip düzeltme olarak tanımlanmaktadır (Shaughnessy vd., 2008).

Duygu (affect), duyguların ve duyguların deneyimini ifade eden psikolojik bir terimdir. Duygular, ruh halleri veya benlik saygısı gibi benliği rahatsız eden duygular ve diğer zihinsel durumlar için genel bir terimdir (Forgas, 1994). Ayrıca öğrencilerin korku, endişe, fiziksel rahatsızlık veya sinirlilik gibi test durumlarına fiziksel bir tepkisidir (Lufi vd., 2004). Duyguyu yansıtan **endişe (worry);** öğrencilerin değerlendirme türünün bilişsel yeteneklerini ve kaygı deneyiminin bilişsel bileşenlerini ölçmek için uygun olup olmadığına dair öz değerlendirmesidir (Awang-Hashim vd., 2010). Test durumlarına bağlanan bilişsel bir sıkıntı olarak da tanımlanmaktadır (Lufi vd., 2004)**. Çaba** ise denemeye devam etme isteği ve bir görevi tamamlamak için zihinsel güç veya devam etme isteği olarak tanımlanmaktadır (Awang-Hashim vd., 2010).

Göz İzleme, görsel dikkate erişmek için göz konumunu ve hareketini izlemektir. Araştırmacıların katılımcının göz hareketlerini gerçek zamanlı olarak ölçmelerine ve herhangi bir zamanda kullanıcının odak noktasının nerede olduğunu bilmelerine yardımcı olur (IMOTIONS, 2021).

1.3.Araştırma Soruları

1) Ortaokul öğrencilerini öğrenme çıktılarına hazırlamak için yürürlüğe giren ortaokul matematik öğretim programı (2017-2018 akademik dönemde geçerli olan) önerilen değerlendirme prosedürleriyle ne ölçüde uyumludur?,

2) Ortaokul matematik öğretmenlerinin matematik öğretim programı değişikliği sonrasında sınıfta kullandıkları öğretim yöntemleri ve ölçme-değerlendirme stratejileri önceki programa (2016-2017 akademik yarıyıl sonuna kadar geçerli olan) göre nasıl farklılık göstermektedir?,

3) Ortaokul öğrencileri üstbilişsel becerilerini (bilişsel strateji ve kendi kendini kontrol etme) ve duyuşsal süreç (çaba ve endişe) düzeylerini farklı madde türlerine verdikleri yanıtlara göre nasıl yansıtmaktadırlar? Öğrencilerin üstbilişsel beceri ve duyuşsal düzeylerinin çoktan seçmeli ve açık uçlu maddelere verdikleri yanıtlara yansıma miktarları arasında anlamlı bir fark var mıdır?,

4) Öğrencilerin farklı türdeki sorulara verdikleri tepkileri ve sorulara yanıtları galvanik cilt tepkisi (GSR) ve kalp atış hızı (HR) dahil olmak üzere göz izleyici ve biyometrik sensörlerin kullanımıyla farklı bilişsel stratejiler ve duyuşsal açısından (aktif kullanım) açılardan değerlendirmek nasıl mümkün olabilir?,

5) Öğrencilerin üstbilişsel (bilişsel strateji ve kendi kendini kontrol etme) ve duyuşsal süreç (endişe ve çaba) tepkilerini bir derin veri model tasarımı aracılığıyla ölçüp değerlendirebilecek hangi nöro/biyobelirteçlere ihtiyaç vardır?

2. YÖNTEM

Bu karma yöntem çalışması üç temel araştırma boyutu ve beş alt aşamadan oluşmaktadır. 1) *Doküman Analizi*: Özgün öğretmen yapımı soru maddelerinin incelenmesi; 2) *Nicel Tarama Aşaması*: Öğretmenlerin öğretim yöntemi ve ölçme-değerlendirme stratejisi tercihlerinin incelenmesi; 3) *Çoklubiçimli Aşama (Multimodal Phase)*: Öğrencilerin üstbiliş ve duyuşsal süreçlerinin farklı soru türlerine yansıması; 4) *Çoklubiçimli Aşama*: Göz izleme ve biyometrik sensörler kullanılarak öğrencilerin farklı türdeki sorulara tepkilerinin ve yanıtlarının değerlendirilmesi (Aşama 3 ve Aşama 4 Çoklubiçimli Aşama: Eğitimsel Nöreobilim olarak adlandırılacak) 5) Birleşme (Entegrasyon): Üstbiliş ve duygusal süreçler için bir Derin Veri Sisteminin modellenmesi.

2.1.Araştırma Tasarımı

Bu çalışmanın amacı ve araştırma sorularının doğası gereği, araştırmanın deseni karma yöntem araştırma desenidir. Karma yöntemli bir çalışma, verilerin eşzamanlı veya sıralı olarak toplandığı, öncelik verildiği ve verilerin sürecin bir veya daha fazla aşamasında entegrasyonunu içeren tek bir çalışmada hem nicel hem de nitel verilerin toplanmasını veya analizini içerir (Creswell vd., 2003). Creswell'e (2018) göre, karma yöntem araştırmaları arasında birçok tasarım mevcuttur. Creswell'in Kısmen Karma Eşzamanlı Baskın Durum tasarımından adapte edilerek (detay için sf. 84'deki görseli inceleyiniz) Çoklubiçimli Karma Eşzamanlı Baskın Durum deseni özgün olarak tasarlanmıştır.

2.2.Araştırma Katılımcıları

Kısmen karma eş zamanlı baskın durum araştırma tasarımına dayanan bu çalışma, 2017-2018 akademik yılının güz döneminde başlamış, 2020-2021 akademik dönemlerinde de devam etmiştir. Orta Doğu Teknik Üniversitesi (ODTÜ) ve Milli Eğitim Bakanlığı (MEB) etik kurul izinlerinin ardından (bkz. Ekler D-I) araştırmacı, aşağıdaki süreçleri gerçekleştirmiştir. İstanbul ili Sarıyer ilçesindeki resmi ve özel ortaokullardan araştırmaya katkıda bulunmak isteyen toplam 8 okul (2 özel ve 6 devlet) bulunmaktaydı.

2.2.1. Doküman Analizi: Otantik öğretmen yapımı soru maddelerinin incelenmesi

Bu araştırmanın ilk boyutunun çalışma grubu amaçlı örnekleme olarak seçilmiştir. Analiz birimi İstanbul ili Sarıyer İlçesi mahallesinde bulunan her bir okuldur. 5 ortaokul (1 özel ve 4 devlet) çalışmanın ilk aşamasında katkı sağlama davetini kabul etmiş, otantik öğretmen yapımı sınavlar toplam 10 ortaokul matematik öğretmeninden toplanmıştır.

2.2.2. Nicel Tarama Aşaması: Öğretmenlerin öğretim yöntemlerinin ve ölçme-değerlendirme strateji tercihlerinin değerlendirilmesi

Bu araştırmanın ikinci boyutunun ilk çalışma grubu amaçlı örnekleme yöntemine göre seçilmiştir. Sarıyer ilçesindeki 8 ortaokuldan (2 özel ve 6 devlet) toplam 14 ortaokul matematik öğretmeni katılmıştır. Katılımcılar 28-42 yaşları arasında 11 kadın (%79) ve 3 erkek (%21) olmak üzere mesleki kıdemleri 2 ile 20 yıl arasında değişmektedir. Öğretmenlerin çoğu (n = 11) ilköğretim matematik öğretmenliği (eğitim fakültesi) mezunudur. Bunlardan sadece üçü matematik (fen fakültesi) bölümünden mezun olmuş ve pedagojik formasyon sertifikası almışlardır. Öğretmenlerin çoğu iki sınıf düzeyinde (örneğin, 5. ve 7. sınıf veya 5. ve 6. sınıf), çok azı ikiden fazla sınıf düzeyinde (örneğin, 5. sınıf, 7. sınıf ve 8. sınıf) matematik dersi vermekteydi.

Yukarıdaki pilot çalışmaya katılan öğretmen grubundan sonra, araştırmanın ikinci çalışma grubu (ana katılımcı grubu) 350 ortaokul matematik öğretmeninden oluşmaktadır. Bu katılımcılar elverişli örnekleme (convenient sampling) yöntemine göre seçilmiştir. Araştırmacı İstanbul ilindeki ilçeleri gezerken bulduğu okullarda uygun olan ve araştırmaya gönüllü katkı sağlamak isteyen öğretmenlerden veri toplayabilmiştir. Hatta okulların tatil olduğu seminer dönemlerinde İlçe Milli Eğitim'den seminer okullarını öğrenmiş; böylece o okulları ziyaret ederek seminer öncesi veya çıkışlarında yüksek sayıda ortaokul matematik öğretmenine ölçeği uygulamak üzere ulaşabilmiştir.

2.2.3. Çoklubiçimli Aşama: Nöroeğitim

(Bu bölüm Aşama 3: Öğrencilerin üstbiliş ve duyuşsal süreçlerinin farklı soru türlerine yansıtılması ve Aşama 4: Göz-izleme ve biyometrik sensörler kullanılarak öğrencilerin farklı soru türlerine fizyolojik tepkilerinin ve yanıtlarının değerlendirilmesi adımlarını içeren araştırmanın üçüncü temel boyutudur.)

Nöroeğitim çalışmasına (araştırma bağlamı ve ortamı için bkz. sf. 104) katılan ortaokul öğrencileri önce amaçlı örnekleme sonra da kartopu örneklemine göre seçilmiştir. 5. sınıf öğrencileri 10 yaşında, 17'si kız (%53) ve 15'i erkek (%47)'tir. Toplamda yaklaşık beşte biri (%22) özel okullarda okurken çoğu Beşiktaş, Sarıyer, Fatih ilçelerinde devlet okullarında okumaktadır. Öğrencilerin eğitim gördüğü ilçeler üç bölgeye ayrılmış; beş öğrenci Beşiktaş'ta (%16), 13'ü Sarıyer'de (%41) ve 14'ü Fatih ilçelerinde (%44) öğrenim görmektedir. Öğrencilerin, laboratuvar araştırması sürecinde matematik sorularını çözerken kullandıkları maksimum toplam süre 31 dakika, minimum süre 9 dakikadır. Test verilerine göre, süre açısından, maddelere yanıt verirken en fazla zaman ayırdıkları soru 2,9 dakika ile 3. soru (Revize edilmiş Bloom Taksonomisine göre *İşlemsel* bilgi, *Değerlendirme* bilişsel süreci gerektiren bir soru) olmuştur.

2.3. Veri Toplama Araçları

2.3.1. Doküman Analizi: Otantik öğretmen yapımı soru maddelerinin incelenmesi

Türkiye'deki ortaokul matematik öğretmenleri, öğrencilerine bir yarıyılda iki temel sınıfiçi sınav yapmakla yükümlüdür. Araştırma sürecinin ilk sorusunu takiben otantik öğretmen yapımı sınıfiçi sınavların kendisi ortaokul matematik öğretmenleri tarafından bizzat araştırmacı ile paylaşılmıştır.

2.3.2. Nicel Tarama Aşaması: Öğretmenlerin öğretim yöntemlerinin ve ölçme-değerlendirme strateji tercihlerinin değerlendirilmesi

Araştırmacı tarafından geliştirilen, sekiz soru içeren yarı yapılandırılmış bir görüşme protokolü (bkz. Ek K) hazırlanmıştır. Formun geçerli ve güvenilir olması için ayrıca iki matematik öğretmeni, matematik eğitimi alanında bir doçent ve araştırmacı danışmanından uzman görüşüne tabi tutulmuştur. Matematik öğretmenlerinden 2017 sonrası Türkiye'deki eğitim politikası değişikliği sonrası öğretim programı değişikliği (öğretim programı bakışları, gözlemledikleri benzerlikler, matematik öğretimine yönelik öğretim yöntemi tercihleri ve öğrenci değerlendirme ve sınav hazırlama tercihleri, yapılandırmacı yaklaşım) hakkında konuşmaları istenmiştir. Araştırmacı kültürel özel bağlama göre pilot görüşmeden çıkarımlara ve teorik literatüre dayanarak, Öğretim Yöntemleri [ÖY/TM] ve Ölçme-Değerlendirme Strateji [ÖDS/MES] Tercihleri Anketi (ÖYÖDST-A) geliştirilmiştir.

ÖYÖDST-A için maddelerin yazılması. ÖYÖDST-A (İngilizcesi TMMESP-Q), matematik öğretmenlerinin sırasıyla öğretim yöntemi ve ölçme-değerlendirme stratejileri tercihlerini ölçen iki bölüme dayanmaktadır. Ölçek "kesinlikle katılıyorum" ile "kesinlikle katılmıyorum" arasında 5'li Likert tipi bir ölçekte derecelendirilen toplam 35 maddeden oluşmaktadır. TMMESP-Q maddelerinin Türkçe versiyonunun çevirisi Ek M'de, çeviri süreci ise Ek N'de tablolaştırılmıştır (Yapı geçerliği sonuçları için bkz. sf. 180).

2.3.3. Çoklubiçimli Aşama: Nöroeğitim

Pilot hazırlık aşamasından (detay için bkz. sf. 137) sonra TIMSS, PISA, MEB ulusal sınavları ve matematik öğretmenlerinin sınıf içi otantik sınavlarına benzer şekilde ortaokul öğrencileri için 10 matematik maddesi hazırlanmış ve uzmanlarında görüşleri alınarak soru havuzundan maddeler seçilmiştir. Bu soru havuzu çoktan seçmeli ve açık uçlu olarak 2 gruba ayrılmıştır. ME, CI ve MFE departmanlarının uzman görüşünden sonra, Sesli Düşünme Süreci protokolü (bkz. Ek R) tasarlandı. Bu protokolün amacı öğrencilerle yapılan ana çalışmadaki görüşmelerde öğrenciler soru maddelerini çözerken üstbilişsel ve duyuşsal süreç ölçümlerinin yapıldığı performansları sözlü olarak değerlendirmelerine yardımcı olmaktır.

2.4.Veri Toplama Süreçleri

2.4.1. Doküman Analizi: Otantik öğretmen yapımı soru maddelerinin incelenmesi

Otantik (özgün) öğretmen yapımı sınıfiçi sınavlar, Türkiye'de alt-orta ve orta SES ilçelerinde bulunan 5 farklı devlet (%62) ve özel okuldan (%37), iki set olarak, 10 matematik öğretmeni tarafından araştırmacıya teslim edilmiştir. Toplam 21 özgün öğretmen yapımı sınav kağıdı toplanmış, soru maddesi toplamının 380 olduğu görülmüştür.

2.4.2. Nicel Tarama Aşaması: Öğretmenlerin öğretim yöntemlerinin ve ölçme-değerlendirme strateji tercihlerinin değerlendirilmesi

Araştırmacı etik kurul izinleri kapsamında İstanbul’da bulunan çeşitli ilçelere kendi imkanları ile ulaşmış, okul yönetimi ve çalışmaya gönüllü katılmak isteyen matematik öğretmenleri ile tanışmıştır. Ders aralarında öğretmenler ile birebir görüşerek TMMESP-Q anketini doldurtmuştur. Bu şekilde metropol bir şehir olan İstanbul’da ilçeleri gezerek 20’den fazla okuldan veri toplamıştır. Daha önce de belirtildiği gibi araştırmacı İstanbul ilindeki ilçeleri gezerken okulların tatil olduğu seminer dönemlerinde İlçe Milli Eğitim’den seminer okullarını öğrenmiş; böylece o okulları ziyaret ederek seminer öncesi veya çıkışlarında yüksek sayıda ortaokul matematik öğretmenine ölçeği uygulamak üzere ulaşabilmiştir.

2.4.3. Çoklubiçimli Aşama: Nöroeğitim

RQ3 ve RQ4'ü araştırmak için (ayrıntılar için sayfa 5'e bakın), uzun Covid-19 salgını sırasında Beyin Dinamikleri Laboratuvarı'ndaki yardımcı danışmanımla işbirliği yaptım. Pilot hazırlık aşamasından sonra ortaokul öğrencileri için TIMSS, PISA ve MEB ulusal sınavlarına ve otantik matematik öğretmenlerinin sınıf içi sınavlarına benzer 10 matematik maddesi hazırlanarak seçilmiştir. Bu madde havuzu çoktan seçmeli ve açık uçlu maddeler olmak üzere iki gruba ayrılmıştır. Ölçme ve Değerlendirme, Eğitim Programları ve Öğretim, ve Matematik ve Fen Bilimleri Eğitimi Bölümlerinden uzman görüşüne danışıldıktan sonra, Think-aloud Process (sesli düşünme süreci) protokolü tasarlandı. Ancak Kovid-19 koşulları nedeniyle MEB'den izin yine ertelendi ve neredeyse altı ay beklemek zorunda kaldım. Laboratuvar ortamı öğrencilerin iyiliği için özel olarak hazırlanmıştır. Otuz iki 5. sınıf öğrencisi aileleri ile birlikte davet edildi ve bazen özel okul servisi ile taşındı. Deneysel sürece gönüllü olarak katılmışlardır ve laboratuvarda yürütülen ve yaklaşık 40 dakika süren üstbilişsel ve duyuşsal süreçlerle ölçülen performansları değerlendirilmiştir.

Bu çalışmalar 27 Ocak 2021 ile 26 Mart 2021 tarihleri arasında iki ay sürmüştür. Öğrenciler maddeleri cevaplarken sesli düşünme süreci gerçekleştirilmiştir. Dört ay boyunca, farklı biyometrik araçlar, görüşmeler ve alan notlarımdan ayrıntılı veriler içeren derinlemesine analiz raporları yazıldı. Analizin kodlayıcılar arası güvenilirliğini sağlamak için bir ortak kodlayıcı da kullanıldı. Her bir veri analiz süreci, fikir alışverişi ve uzman görüşleri danışman ve yardımcı danışman tarafından kayıt altına alınmıştır. Veriler bütüncül olarak analiz edildi ve sonuçlar elde edildi. Adım 4: Yüksek Sesle Düşünme Protokolünden ve Bilişsel-Duygusal Ölçüm Araçlarından Veri Derleme Süreci Sesli düşünme sürecini kullanan öğrencilerin ses kayıtları kelimesi kelimesine yazıya dökülmüştür. Google Drive'da her öğrenci için bireysel dosyalar açılarak göz takip sistemi, ses kayıtları ve GSR Empatica 4 duygu ölçümlerinden elde edilen veriler saklandı. Bu aşamada her bir birey için veri biriminin kaydedilmesi bir saat sürmüştür.

2.5.Veri Analizi

2.5.1. Doküman Analizi: Otantik öğretmen yapımı soru maddelerinin incelenmesi

Otantik öğretmen yapımı sınavlardan gelen 380 sınıf içi sınav maddelerinin içeriğini 5 ana temada belirlemek için doküman analizine (Patton, 2002) tabi tutulmuştur. Böylece sınav maddelerine ilişkin doküman incelemesi 5 ana temaya ayrılmıştır: 1) Madde türü, 2) Ortaokul Türkiye Matematik Öğretim Programının Öğrenim Ünitesi (MEB, 2018), 3) Egitim programından elde edilen kazanımlar (MEB, 2018), 4) Revize Edilmiş Bloom'un Taksonomisi (Bloom ve diğerleri, 1956; Anderson ve Krathwohl, 2001); 4.1.) Bilgi Düzeyi ve 4.2) Bilişsel Süreç Boyutu, 5) Uluslararası Matematik ve Fen Eğilimleri Araştırması [TIMSS] Bilişsel Alan ve alt alanları.

2.5.2. Nicel Tarama Aşaması: Öğretmenlerin öğretim yöntemlerinin ve ölçme-değerlendirme strateji tercihlerinin değerlendirilmesi

Tarama süreci sonunda elde edilen veriler öncelikle kontrol edilmiş, her bir öğretmenin ankete verdikleri yanıtlar önce excel.xls ardından SPSS istatistiksel paket programa aktarılmıştır. Eksik veri (missing data) olup olmadığı kontrol edilmiştir. Veriler betimsel olarak analiz edilmiş, aykırı değerler danışman görüşleri de dikkate alınarak ayıklanmıştır. 350 katılımcı grubundan 6 adet ayrıkı değer atılmış, 344 veri analiz edilmek üzere saklanmıştır. Veriler araştırma sorularını takiben betimsel ve yordamsal açıdan istatistiksel olarak analiz edilmiştir.

2.5.3. Çoklubiçimli Aşama: Nöroeğitim

Çalışmanın üçüncü büyük boyutundaki veriler, sesli düşünme süreci protokolü kullanılarak 10 yaşındaki 32 5. çocuktan toplanmıştır. Cinsiyet, okul bölgesi, okul büyüklüğü, okuldaki öğretmen sayısı ve bölgedeki okul sayısı ile ilgili bilgiler elde edildi. Araştırmanın amacı, çocukların 10 matematik maddesini çözerken kullandıkları üstbilişsel ve bilişsel becerileri belirlemektir. Görüşmeler yazıya dökülmüş ve çeşitli cihazlar kullanılarak fizyolojik veriler toplanmıştır. İki farklı şablon hazırlandı. Görüşmeler bu şablonlar kullanılarak her çocuk için ayrı ayrı kodlanmıştır. Farklı biyometrik araçlar, görüşmeler ve araştırmacının alan notlarından detaylı veriler olduğu için yaklaşık 4 ay boyunca derinlemesine analiz raporları yazılmıştır. Analizin kodlayıcılar arası güvenilirliğini sağlamak için bir ortak kodlayıcı da kullanıldı. Çünkü veri birleştirme (data aggregation), karma yöntemden gelen araştırma verilerinin *birleşme* fonksiyonu dikkate alınarak bir araya getirildiği ve özet olarak aktarıldığı süreçtir. Tipik olarak istatistiksel analizin performansından önce kullanılır.

RQ 3'ü ve RQ 4'ü araştırmak üzere, her adımın hazırlanmasında danışmanım ve yardımcı danışmanımdan uzman görüşü aldım ve istatistiksel analizler yaptım. Verilerin her aşamada anlamlı bir şekilde derlendiğinden, toplandığından ve entegre edildiğinden emin oldum. Haftalık 2 saatlik toplantılarda danışmanımdan ve yardımcı danışmanımdan uzman görüşü ve geri bildirim aldım. Daha sonra, temel istatistik kitaplarından araştırma sorularına cevap arayacak analizleri kontrol ederek istatistiksel analizleri çalıştırdım (örn., Field, 2013; Tabachnick ve Fidell, 2013). Çalışmanın sonuçlarını detaylandırdıkça, nöroeğitim araştırma sürecinde topladığım ve analiz ettiğim verilerin büyük bir yüzdesinin kategorik olduğunu, yani nominal bir ölçüm düzeyine sahip olduğunu fark ettim; Nöroeğitim verileri ile istatistiksel olarak çıkarımsal bir analizini yapmak için Ki-Kare Testi ile Binomial Lojistik Regresyon analizleri seçilmiştir (Tabachnick ve Fidell, 2013). Veriler bütüncül olarak analiz edilmiş ve sonuçlar raporlanmıştır.

3. BULGULAR

3.1.1. Doküman Analizi: Otantik öğretmen yapımı soru maddelerinin incelenmesi

380 otantik öğretmen yapımı sınıfiçi matematik sınavı maddesinin tamamı, ulusal ortaokul matematik müfredatı öğrenme çıktıları ve konu alanı (yani matematik üniteleri) ile ilgili olarak analiz edildi. Özellikle, maddelerden elde edilen bulgular, ortaokul matematik öğretmenlerinin 5. sınıfta sıklıkla Sayılar ve İşlemler temel ünitesine dayalı test maddeleri hazırlama eğiliminde olduklarını ve madde türleri açısından kendilerini açık-uçlu (AU) ve çoktan seçmeli (ÇS) maddeleri ile kısıtladıklarını ortaya koymuştur. Öğretmenler matematik öğretim programı öğrenme çıktılarına uygun maddeler hazırlayabilmekteler. Bununla birlikte, bu öğrenme çıktılarının, maddelerin geliştirildiği 5. sınıf seviyesinin üzerinde (yani 6. veya 7. sınıf) veya altında (3. veya 4. sınıf) olduğu bulundu. 5. sınıf ulusal matematik öğretim programında yer alan kazanımların beşte birinin (f = 70) üst düzey düşünme becerileri (HoTs) ile ilişkili olduğu, beşte dördünün (f = 310) ise alt düzey düşünme becerileri (LoTs) (f = 310) ile ilişkili olduğu tespit edilmiştir. Beş farklı okulda 10 matematik öğretmeni tarafından yapılan tüm otantik öğretmen yapımı maddeler (N = 380) analizi, 13 sınavın 1. yarıyıl kazanımlarına göre, sekiz sınavın ise 2. yarıyıl kazanımlarına göre hazırlandığını göstermiştir. Politika değişikliği öncesi ve sonrası ile ilgili olarak, diğer bir deyişle politika değişikliği öncesi 69, politika değişikliği sonrası 311 soru maddesi hazırlanmıştır.

10 matematik öğretmeninin sınıfiçi sınav analizinden elde edilen verilere göre (Revize edilmiş Bloom Taksonomisinin Bilgi boyutunda) 55 madde Olgusal (%14.47), 96 madde Kavramsal (%25.26), 228 madde İşlemsel (%60), 1 madde Üstbilişsel (%0.26) düzeyde; (Revize edilmiş Bloom Taksonomisinin Bilişsel Süreç boyutunda) 44 madde Hatırlama (%11.58), 91 madde (%23.97), Uygulamada 217 madde (%57.11), Analiz etmede 25 madde (%6.58), Değerlendirmede 3 madde (%0.79) olduğu görülmüştür. Yaratma (creating)

düzeyinde herhangi bir soru maddesi eşlenememiştir. Özetle, sınıfiçi sınav maddesi analizleri Revize edilmiş Bloom Taksonomisine göre, çoğunlukla Prosedürel bilgi düzeyi boyutuna (*f* = 228, %60) ve Uygulama bilişsel süreci boyutuna (*f* = 217, %57.11) dayandığını ortaya koymuştur.

Bulgular, matematik öğretmenlerinin en çok geleneksel nesnel testleri kullanma eğiliminde olduğunu ortaya koymuştur. Revize edilmiş Bloom Taksonomisinin bilgi düzeyi ve bilişsel süreç boyutları ile ilgili olarak, matematik maddelerinin çoğu Prosedürel (*f* = 228, %60), dörtte biri Kavramsal (*f* = 96, %25.3) ve bazıları Olgusal (*f* = 55, %14.5) düzeyi yansıtıyordu. Neredeyse hiç Üstbilişsel (*f* = 1, %0.3) bilgi düzeyi boyutunda soru hazırlamadıkları fark edilmiştir. Bilişsel süreç boyutu ile ilgili olarak öğretmenlerin yarısı Uygulama (*f* = 217, %571), yaklaşık beşte biri Anlama (*f* = 91, % 23.9), yüzde onbir Hatırlama (*f* = 44) ve biraz Analiz (f = 25, %6.6) düzeyinde soru hazırlama eğiliminde olmuşlardır. Ancak birkaçı Değerlendirme (*f* = 3, %0.8) seviyesinde soru hazırlayabilmiştir. TIMSS Çerçevesinde tamamlayıcı bulgulara baktığımızda, öğretmen tarafından hazırlanan otantik sınıfiçi sınav maddelerinin çoğunlukla ana alanın Bilgi düzeyine (*f* = 331, %87) ve alt alanın Bilgi işlem düzeyine (*f* = 164, %43) dayandığını ortaya konulmuştur.

Özetle, bulgular öğretim programı değişikliğinin öğretmenlerin sınıfiçi sınav hazırlama konusunda tam olarak kendilerini yenilemediğini ortaya koydu. Ayrıca, öğretmen yapımı soru maddeleri genellikle düşük bilişsel düzey becerileri ölçecek düzeyde hazırlanmakta, uluslararası standartlarda bile üst düzey bilişsel becerileri ölçecek şekilde karşılamadığı tespit edilmiştir.

3.1.2. Nicel Tarama Aşaması: Öğretmenlerin öğretim yöntemlerinin ve ölçme-değerlendirme strateji tercihlerinin değerlendirilmesi

TMMESP-Q Analizi Betimsel İstatistik Sonuçları. Betimsel istatistik sonuçlara göre öğretmenler, öğretim programının felsefesinde bir değişiklik olmadığı konusunda hemfikirdir (%52.50); öğrenme çıktılarının sayısındaki düşüşü (%76.70) fark ettiklerini, aksine konunun içeriğinin zenginleştirildiği (%47.50) konusunda hemfikir olmadıklarını belirtti. Milli Eğitim Bakanlığı Matematik kaynak kitaplarının içeriğinde (%61) herhangi bir değişiklik olmadığını, öğretmen el kitabını kullanmayı tercih ettikleri konusunda düşüncelerini belirtti (%42.70). Genel öğretim boyutunda öğrencilerin yaratıcılığına olanak sağlayan etkinlikler yapmayı tercih ettiklerini (%87.90); öğrencilerini aktif hale getirmek için sınıf içi öğretim yöntemlerini değiştirdiklerini (%92.60); dersten önce öğrencilerin hazır bulunuşluklarını kontrol ettiklerini belirtti (%88.50). Yalnızca doğrudan öğretim yöntemi kullanmamayı (%72.60) tercih ettikleri ortaya çıktı. Ortaokul matematik öğretmenleri, eğitim politikası değişikliğinden sonra öğretim programının temellerinde (yani felsefe, amaç, içerik) belirli bir değişiklik olmadığını düşünmekte, öğrencilerin hazır bulunuşluklarını kontrol ederken öğrencilerini aktif hale getirmek için sınıf içi öğretim yöntemlerini değiştirmeyi, öğrenci yaratıcılığına olanak sağlayan etkinlikler yapmayı tercih etti. Öğretim tekniklerine bakıldığında ise sınıf içi öğretimde somut materyal kullanımı, grup öğretim yöntemleri, eğitim teknolojileri ve farklı sorgulama teknikleri gibi daha yapılandırmacı yaklaşımları tercih etme eğiliminde oldu.

Ölçme-değerlendirme açısından genel ve alt boyutlar değerlendirildiğinde, öğretmenlerin genel ölçme-değerlendirme süreci boyutunda, önceki uygulamalara göre değişiklik yapmayı “bazen” tercih ettikleri, sınavları nerdeyse hiç veya nadiren indirilen çevrimiçi kaynaklara dayalı olarak hazırlama eğiliminde oldukları (örneğin forumlar, web siteleri vb.) belirlendi. Öğrenme kazanımlarını ölçmek için “bazen” biçimlendirici değerlendirmeyi kullandıkları, genellikle öğrencilerin sınavlarda prosedürel becerileri kullanmasını gerektiren madde türlerini tercih ettikleri belirlendi. Öğretmenlerin %53'ü çoktan seçmeli maddeler içeren ölçme araçlarını kullanmayı, %63.60'ı çoktan seçmeli ve kısa yanıtlı maddelerin karışımını içeren sınavlar hazırlamayı, %74.40'ı sınıf içi sınavlarımda açık uçlu maddeler kullanmayı tercih ettiğini belirtirken, %42.40'ı öğrencilerin dönem sonunda performanslarını göstermelerini sağlayacak portfolyo kullanmayı tercih ettiğini belirtti.

TMMESP-Q Analizi Çıkarımsal İstatistik Sonuçları. Sonuçlar çıkarımsal istatistiklerle de değerlendiridi. Çünkü ilgili literatür öğretmenlerin cinsiyeti, okul türü, kıdem yılı (meslek yılı) ve eğitim düzeyi (mezuniyet programı) gibi bazı bağımsız değişkenlerin öğretmenlerin öğretimsel strateji ve yöntemlerini (bağımlı değişken) tercihlerini belirlediğini belirtmektedir.

Öncelikle t testi istatistiklerinin varsayımları kontrol edildi. Rastgele örnekleme, bağımsız gözlem, normallik varsayımları kontrol edildi. Levene Test sonucu ile homojenlik varsayımı takip edilmiş, anlamlılık değeri anlamlı bulunmadı. Bağımsız örneklemler T testi ile devam edildi.

Öğretmenlerin *cinsiyet*lerine göre öğretim yöntemi tercihlerini karşılaştırmak için bağımsız örneklemler t testi kullanılmıştır (Tablo 4.15). Sonuçlar, kadın öğretmenler (*M* = 79.36, *SD* = 7.98) ve erkek öğretmenler (*M* = 77.07, *SD* = 7.71) arasında öğretim yöntemi tercihleri toplam puanlarında istatistiksel olarak anlamlı bir fark olduğunu; t (297) = 2.32, $p < .05$, $r^2 = .02$, iki kuyruklu test. Kadın öğretmenlerin öğretim programı değişikliği sonrası öğretim yöntemlerini benimsemeleri erkek öğretmenlere göre daha yüksek bulunmuştur. Ortalama farkı okursak, Cohen's d = (Ortalama Fark/SD) = (2.82)/ $\sqrt{15.69}$= 0.71. Cohen's Standard'a (1988) göre 0 < d < .20 = küçük etki; .20 < d < .80 = orta etki ve d > .80 = büyük etki, 0.71 > .20 olduğundan orta etkidir. Ayrıca, Eta kare $\eta^2 = t^2/(t^2+df) = (2.32)^2/ (2.32^2+297) = 0.02$. Cohen'in Standardına (1988) göre .01 = küçük etki, .06 = orta etki ve .15 = büyük etki; .02 > .01 olması nedeniyle küçük ile orta düzeyde bir etkidir. Yani, *cinsiyet değişkeni* öğretmelerin öğretim yöntemi tercihlerindeki varyansın %2'sini açıklamaktadır. Ancak, öğretmenlerin ölçme-değerlendirme stratejisi tercihine göre; erkek öğretmenler (*M* = 51.34, *SD* = 7.70) ve kadın öğretmenler (*M* = 50.87, *SD* = 6.90) arasında ölçme-değerlendirme stratejisi tercihi toplam puanları arasında istatistiksel olarak anlamlı bir fark bulunmamıştır.

Öğretmenlerin *kıdem yıllar*ına göre öğretim yöntemi tercihlerini karşılaştırmak için başka bir bağımsız örneklem t testi kullanılmıştır (Tablo 4.16). Sonuçlar, kıdem yılı 15'ten küçük olan öğretmenler (*M* = 78.11, *SD* = 8.19) ile kıdem yılı 15'ten büyük ve eşit olan (*M* = 80.27, *SD* = 6.84); t (296) = 2.05, $p < .05$, $r^2 = .01$, iki kuyruklu test. Meslekte 15 yıl ve üzeri deneyime sahip öğretmenlerin öğretim programı değişikliği sonrası öğretim yöntemi tercih toplam puanının 15 yıldan az meslekte deneyimli öğretmenlere göre daha yüksek olduğu bulunmuştur. Meslekte en az 15 yıl görev yapmış öğretmenlerin öğretim programı değişikliği sonrasında farklı yapılandırmacı öğretim yöntemlerini tercih etme eğiliminde oldukları ve öğretim yöntemi tercihleri açısından değişimin sınıf içerisindeki değişim öncüsü (i.e., agent of change) olmaya eğilimli olduğu söylenebilir. Ortalama farkı okuyarak, Cohen's d = (Ortalama Fark/SD) = (2.16)/ $\sqrt{15,03}$= 0.56. Cohen's Standard'a (1988) göre $0 < d < .20$ = küçük etki; $.20 < d < .80$ = orta etki ve $d > .80$ = büyük etki, $0.56 > .20$ olduğundan orta etkidir. Ayrıca, Eta kare $\eta^2 = t^2/(t^2+df) = (2.05)^2/ (2.05^2+296) = 0.01$. Cohen'in Standardına (1988) göre .01 = küçük etki, .06 = orta etki ve .15 = büyük etki; .01 = .01 olduğundan küçük bir etkidir. Ayrıca öğretim yöntemi tercihlerindeki varyansın %1'inin *kıdem yılı* ile açıklandığı şeklinde yorumlanabilir. Ancak, öğretmenlerin ölçme-değerlendirme stratejisi tercihi bağlamında; kıdem yılı değişkenine göre ölçme-değerlendirme stratejisi tercihi toplam puanları arasında istatistiksel olarak anlamlı bir fark bulunmamıştır.

Öte yandan eğitim düzeyi ve okul türüne göre, öğretmenlerin öğretim yöntemi tercihi toplam puanları arasında istatistiksel olarak anlamlı bir farklılık bulunmamıştır. Benzer şekilde devlet okullarında görev yapan öğretmenler (*M* = 51.32, *SD* = 7.43) ile özel okullarda görev yapanlar (*M* = 50.48, *SD* = 6.60) arasında ölçme-değerlendirme stratejisi tercihi toplam puanları arasında istatistiksel olarak anlamlı bir fark bulunmamıştır.

Çoklubiçimli Aşama: Nöroeğitim. Öğrencilerin çoktan seçmeli maddelere verdikleri yanıtlarda bilişsel strateji (cognitive strategy) becerisini yansıtma düzeyi (*f* = 129, %26), yansıtamayanlara (*f* = 166, %34) göre sayıca biraz daha düşüktü. Ayrıca, öğrencilerin çoktan seçmeli maddelere verdikleri yanıtlarda öz kontrol (self-checking) becerisini yansıtma düzeyleri (*f* = 190, %29), yansıtamayanlara (*f* = 67, %10) göre 3 kat daha fazlaydı. Diğer bir deyişle, ortaokul öğrencilerinin üçte birinin öz kontrol becerilerini yansıtabilme düzeyi (*f* = 190, %29), bilişsel strateji becerilerini (*f* = 129, %26) yansıtabilmelerinden biraz daha yüksekti. Çözüme yönelik kendine güven düzeyleri düşükse, kendi kendini kontrol etme becerilerini kullanma eğilimindeydiler. Üçte birinden fazlası açık uçlu maddelere verdikleri yanıtlarda bilişsel strateji kullanımını yansıtırken (*f* = 162, %33), üçte birinden daha azı çoktan seçmeli maddelere verdikleri yanıtlarda (*f* = 129, %26) yansıttı.

Üstbilişsel alt beceriler ile çoktan seçmeli ve açık uçlu madde türleri arasındaki ilişkiyi görmek için Ki-Kare Bağımsızlık Testi kullanıldı. Bağımsız gözlem ve beklenen frekansların boyutu karşılandı ve varsayımlar ihlal edilmedi. *Çoktan seçmeli maddelerde öğrencilerin bilişsel strateji becerisini kullanma eğilimleri azalırken, öz kontrol becerilerini kullanma olasılıklarının arttığı bulundu. Açık uçlu maddelerde öğrencilerin bilişsel strateji becerisini kullanma eğilimleri artarken, öz kontrol becerilerini kullanma olasılıkları da artmaktadır*. Betimsel sonuçlar, öğrencilerin çoktan seçmeli maddelere karşı olumlu, açık uçlu maddelere karşı ise olumsuz duygular uyandırdığını ortaya koydu. Betimsel analize ek olarak, madde türleri ile duyuşsal süreçler arasında anlamlı bir ilişki olup olmadığını araştırmak için Ki-Kare istatistik testi uygulandı. Duyuşsal süreçler ile madde türleri arasında istatistiki olarak anlamlı bir ilişki olduğunu göstermiştir, X^2 (1, N = 32) = 54.92, $p < .05$. *Öğrencilerin çoktan seçmeli* *maddelere karşı olumlu uyarılma olasılıkları artarken, açık uçlu maddelere karşı olumsuz uyarılma olasılıkları artmaktadır.*

Bir başka araştırma alt hipotezi ise "H_0: ortaokul 5. sınıf öğrencilerinin yeniden okuma alt becerisinin (rereading subskill of cognitive strategy), cinsiyetleri (gender), soru çözümünde geçirdikleri toplam süre (total time) ve bakışlardaki kaymalar (gaze shifts) ile ilişkili değildir." Bu bağlamda, verilere binomial lojistik regresyon modeli yapıldı.

The predicted logit of (REREADING) = .93 + (-.27)*TOTAL TIME + (-.35)*GAZESHIFT + (.70)*GENDER

Soru çözümünde geçirdikleri toplam süre ve bakış kaymalarındaki her bir nokta artışı, yeniden okuma olasılığında sırasıyla.76 ve.70'lik bir azalma ile ilişkilendirildi. Bir erkek çocuğunun yeniden okuma alt becerisini kullanma olasılığı, bir kız çocuğuna göre 2.01 kat daha fazlaydı. Genel olarak, bu modelde olayların %91'i (1) doğru tahmin edilirken, gerçekleşmeyenlerin olasılıkları %49.2'si (0) doğru tahmin edilmiştir.

Birleşme (Entegrasyon): Üstbiliş ve duygusal süreçler için bir Derin Veri Sisteminin modellenmesi. Bu çalışmadan ortaya çıkan derin veri model tasarımına göre öğrenciler için insan-bilgisayar etkileşimini artıran sistemlerin tasarlanabileceğinin ilk adımları atılmıştır. İyi bir derin veri model tasarımı için, 1) insan davranışlarını anlamlandırabilmek, insanlardan veri toplamak önemlidir. 2) Öğrencileri için hazırlanan soru maddelerinin uzmanlar tarafından belirlenip sisteme öğretilmesine ihtiyaç vardır. 3) Soru çözme sırasında (özellikle açık uçlu soruları çözerken) öğrencilerin stresini fark edip ve stres düzeyini hızlı bir şekilde azaltmak için önerilerde bulunan akıllı sistemler tasarlanabilir. 4) Öğrenci yanlış cevaplar verdiğinde soru düzeyi düşürülebilir. 5) Sistem öğrenciden geri bildirim alır ve içeriğini revize edebilir. 6) Kişiselleştirilmiş otomatik raporlar verebilir. 7) Bu tür modellerin tasarımında problemlerin önceden tanımlanmış olması işleri kolaylaştıracaktır. 8) Bu çalışmada kullanılmayan/henüz ölçülemeyen diğer veriler de öğrencilerden örnekleme alınarak, EEG, fNIRS, 3D beyin tarama verileri, genetik biyobelirteçler (Ahmad vd., 2011) kullanılarak ölçülebilir ve katılım üzerine derinlemesine araştırma olarak toplanabilir. Bu çalışma kapsamında kısıtlı zaman aralıklarında derin verilere ulaşamadığım bilişsel yük, bekleme süresi, göz bebeği genişlemesi, öz açıklama becerisi ve diğer duygusal yapılar hakkında veri toplanması sistemin modellenmesine ışık tutması tutabilir.

4. TARTIŞMA

Bu karma yöntem çalışması kapsamında ortaya çıkan bulgular ve bunların alan yazınla ilişkisi tezde oldukça yoğun şekilde açıklanmaya çalışılmıştır. Bu bölümde en temel tartışma noktalarına değinilecektir.

Eğitim politikası değişikliğinden sonra matematik öğretmenlerinin öğretim yöntemi ve değerlendirme tercihi değişikliklerine ilişkin bu çalışmanın bulguları, ortaokul matematiğinin öğretim programının temellerinde (yani felsefe, amaç, içerik) belirli bir değişiklik olmadığını düşündüğünü gösteren daha önceki gözlemlerle de uyumludur. Literatürde olduğu gibi Türkiye bağlamında da öğretim programlarının yenilenmesi, Bümen vd. (2014), Öztürk (2012) ve Yaşar (2012) tarafından kabul edildiği gibi, sınıftaki öğretmen davranışlarının yenilenmesini garanti etmemektedir. Öğretmen tercihleri farklı değişkenlerle ilişkilendirilir. Örneğin, öğretmen inançları, yeni reformlara karşı olumlu inançlara sahip olma, mesleki gelişime açıklık, mesleki planlarını yapıp yapmama (Zhang ve Shen, 2012), öğretim programı geliştirme sürecinde yer alma, deneyim yılı, bunların tümünün ağı etkilediği görülmektedir. Öğretim programı ve öğretim arasındaki ilişki hakkında yetersiz bilgiye sahip olmak, Kerkez'in (2018) (ayrıntı için bkz. s. 52) öğretmenlerin öğretim programı geliştirme sürecindeki rolüne dayanan çalışmasıyla uyumludur. Ölçme ve değerlendirme süreç tercihlerinin sadece program geliştirme sürecinde bulunmalarıyla ilgili olmadığı, öğretmenlerin zaman zaman geri bildirime ve profesyonel gelişime ihtiyaç duyduğu bir durumdur. Öğretim süreçlerindeki kadar esnek olamamaları öğretmenlerin özelliği veya hatası değildir.

Öğretmenlerin sınav sorularını yazma ilkelerine ilişkin bilgi ve deneyime sahip olamamaları en büyük problemdir; ölçme ve değerlendirme tercihlerinin öğretmenlerin inançlarıyla değil *mesleki gelişim*leriyle ilgili olabileceği çeşitli araştırmalardan bilinmektedir. Eğitim sistemimizde öğretmenlerden hem öğretim hem de ölçme-değerlendirme süreci açısından beklentiler çok yüksek olduğundan, alandaki performanslarını ve sürelerini yeterince otantik değerlendirme türlerine ayıramamaktadırlar. İlk öncelikleri öğretime odaklanmak ve öğretim programındaki konu içeriğini sunmaktır. Dolayısıyla öğretmenler eğitim politikası değişikliğinden sonra öğretim yöntemlerinde inanışları çerçevesinde otonomilerine göre değişiklik yapabilirken; soru hazırlama kalitelerinde, sınıfiçi ölçme-değerlendirme davranışlarında yeterli düzenlemeler yapamamakta, halihazırda bildikleri ve yapabildikleri ile düzeyde sınıfiçi sınavlarını uygulamaya devam etmektedir.

Ancak öğrenciler, onların üst düzey bilişsel becerilerini ölçebilecek soru kalitesi ve yenilikçi soru maddeleriyle karşı karşıya kalmaya hazırdır. Öğrenciler açık uçlu sorulara yanıt verirken bilişsel strateji gibi üstbilişsel becerileri kullanmaya çoktan seçmeli sorulara nazaran daha fazla eğilimlidir. Sonuçlar, O'Neil ve Brown'ın (1998) sonuçlarıyla tutarlıdır. Öğrencilerin çoktan seçmeli maddelere karşı olumlu duygular ve açık uçlu maddelere karşı olumsuz duygular hissetmeleri daha önceki çalışmalarla uyumludur (O’Neil ve Brown, 1998). Her ne kadar üstbilişsel becerilerin değerlendirilmesinde çoktan seçmeli maddenin kullanılması spekülatif bir şekilde görülse de (örn. Silva Soares vd., 2021; Frenken, 2021; O'Neill ve Brown; Chick, 2013; Stillman, 2020; Vuorre ve Metcalfe, 2021) bu çalışma öncü olmaktadır. Unutulmamalıdır ki, soru türünden ziyade (örn., Bassett, 2016; Dutke ve Barenberg, 2015; Scully, 2017) soruların üst düzey düşünme becerilerini ölçecek şekilde ve düzeyde geliştirilmesi, çok daha önem kazanmaktadır.

5. SONUÇ

Eğitim politikaları, sınıfiçi öğretimlerin yenilenmesinde ve eğitim felsefesine yönelik öğretim strateji, yöntem ve tekniklerinin kullanılmasında yetersiz kalmaktadır. Çünkü Türkiye'de ne yazık ki öğretim programı değişikliği çalışmaları yukarıdan aşağıya (top-down) yaklaşıma göre devam etmektedir. Matematik öğretmenlerinin ölçme-değerlendirme tercihlerinde yenilikçi yaklaşımları kullanabilmeleri mesleki gelişim olarak aldıkları eğitimin düzeyine bağlıdır. Matematik öğretmenleri bu çalışmanın sonuçlarından faydalanarak 5. sınıf öğrencilerinin düşünme sürecini anlayabilir, üstbiliş ve duyuşsal tepkilerini fark edebilir, soruları farklı şekillerde çözebileceklerinin farkında olabilir ve içgörü geliştirebilir. Özellikle, bu sonuçları öğretim yöntemlerine ve ölçme-değerlendirme stratejilerine entegre etmede ilerleme kaydetmeleri beklenmektedir. Öğrencilerimiz için geliştirilen ve tasarlanan Türk eğitim sistemine uyarlanmış akıllı bir derin veri modelinin tasarlanması, araştırılması ve uygulanması ancak ortak hazır bulunuşlukta birleşerek yapılabilir, bu da gelecek öngörülerinden biridir.

V. THESIS PERMISSION FORM/TEZ İZİN FORMU

<u>**ENSTİTÜ /** INSTITUTE</u>

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☒

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

<u>**YAZARIN /** AUTHOR</u>

Soyadı / Surname : Birgili
Adı / Name : Bengi
Bölümü / Department : Eğitim Bilimleri, Eğitim Programları ve Öğretim /

<u>**TEZİN ADI /** TITLE OF THE THESIS</u> (**İngilizce** / English): **MIXED METHODS RESEARCH AT THE INTERSECTION OF MATHEMATICS TEACHERS' ADOPTION OF CURRICULUM CHANGE AND ASSESSMENT OF STUDENTS' READINESS LEVEL THROUGH EDUCATIONAL NEUROSCIENCE METHODS**

<u>**TEZİN TÜRÜ**</u> **/** <u>DEGREE</u>**:** **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☒

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☐
2. **Tez <u>iki yıl</u> süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **<u>two years</u>. *** ☐
3. **Tez <u>altı ay</u> süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **<u>six months</u>**. ***** ☒

***** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. /*
A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature .... **Tarih** / Date ....
(Kütüphaneye teslim ettiğiniz tarih. Elle doldurulacaktır.)
(Library submission date. Please fill out by hand.)

Tezin son sayfasıdır. / This is the last page of the thesis/dissertation.