A STUDY ON CONSTRUCT VALIDITY OF SCIENCE ITEMS IN PISA 2006: THE CASE OF TURKEY

by

Ruhan Çirci

B.S., Secondary School Science and Mathematics Education, Teaching Chemistry,
Boğaziçi University, 2005.

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Secondary School Science and Mathematics Education
Boğaziçi University
2009

ACKNOWLEDGEMENTS

I would like to thank all people who have helped and inspired me during my thesis study. Completing this thesis was the hardest task that I have ever faced. I made a lot of mistakes during my gradute study and I learnt a lot. Thank to this process that it has made me the person I am today.

I would like to send deeply-felt thanks to my thesis advisor, Prof. Ali Baykal, for his valuable feedback, which made this thesis possible. Also, I have to express my gratefulness for his warm encouragement and thougthful advises.

I want to thank Assoc. Prof Aysenur Yontar Toğrol who is always with me when I cry, smile or give up. She not only gave feedback for my study but also motivated me.

I have to express my gratefulness to Dr. Fatih Çağlayan Mercan who made crucial (no degree to explain it) contribution to my thesis, especiallly for his patience, encouragement, exacting standards and gracious support by spending valuable time to read this thesis.

I want to thank to Assoc. Prof. Esra M. Akgül for her support at all times during my research and for the many affirming and challenging feedback on my thesis.

I have to thank Prof Dilek Ardaç, Assist. Prof Buket Güzel, Assist. Prof Emine Adadan for their unique visions for my thesis and life. I would like to thank Prof. Füsun Akarsu for her magic in my life.

I want to express my gratitude to my friends, Elif, Ayşegül, Fatıma for not only their moral support but also for listening to me whenever I was excited about a new idea. Also, I want to send special thanks to my office buddies, Melike and Hüseyin, for their valuable effort to motivate me.

My deepest gratitude goes to my family for their unflagging love and support throughout my life. Special thanks to Mehmet, my husband, who enriches my life and without your support, not only this thesis but also everyday would not have been completed.

A special word of thanks to Gulşen Sultan, who is my ultimate coordinator, for everything. I will never forget her constant support when I encountered difficulties. Thanks for sharing happiness and sadness.

As a summary, I want to thank everyone that helped me to realize what I want really.

ABSTRACT

A STUDY ON CONSTRUCT VALIDITY OF SCIENCE ITEMS IN PISA 2006: THE CASE OF TURKEY

This study was designed for three major goals. The first goal is to investigate construct validity in the Turkish version of the PISA 2006 science units including stimuli (e.g. text, graphs, and pictures) and items in terms of positive and negative entities embedded in them. The second goal is to understand the effect of the negative entities detected in the science units on total test and at item level by achievement scores of 15 year-old students. The third goal that emerged through the study aimed to explore the 15 year-old students' familiarity with PISA science units in terms of the students' unique learning experiences.

The study consists of three phases. The first phase is the exploration of eight science units including eight stimuli and 25 items. In order to analyze these science units, content analysis was implemented for all of the stimuli and items in the science units. An *Item Rating Form* was developed and used in order to compose categories of the content analysis from the revisions of the teachers (80 secondary school science teachers in total) working at different regions in Istanbul. Content analysis results showed that there were five main categories defined for the negative entities (content, language, typicality, presentation and structure) and four main categories formed for the positive entities (context, content, science process and composition). The number of the thematic units of the negative entities was more than the thematic units of the positive entities.

The second phase of the study based on the results of the content analysis from first phase. For the second phase the revisions on the science units were made for recovery of the negative entities described in the science units. Hence, two tests were present, one of them is PISA-Original Turkish test (PISA-OT test) which includes Turkish science units used in PISA 2006 study and the other is PISA-Revised Turkish (PISA-RT test) which

contains revised versions of science units made at the second phase of the present study. At the second phase the effect of the negative entities on the achievement scores of the students examined by using two instruments *(PISA-OT* and *PISA- RT)*. The results of the second phase showed that there was a statistically significant effect of the negative entities included in the revision process on the total achievement levels of the students. Although the results of analysis at the item level showed that there were significant differences between the two comparison groups for eleven of the items, there was no significant difference for the remaining eleven items. In addition, it is found that the mean scores for the group who answered the *PISA- Revised Turkish test* were higher than the means of the group who took the *PISA-Original Turkish test* on each of the 22 items.

The third phase is the examination for the answer of the third research question related with the familiarity of students with the stimuli of PISA science units in terms of language, lay-out, school knowledge and daily life experiences. It is found that students tend to be moderately familiar with the PISA stimuli.

Based on the results of the three phases, it can be concluded that the aim of selection and formation of the items for the science literacy test is important from the point of construct validity. However, released PISA stimuli and items in Turkish form achieved this aim partially.

ÖZET

PISA 2006 FEN SORULARININ YAPI GEÇERLİLİĞİ ÜZERİNE BİR ARAŞTIRMA: TÜRKİYE ÖRNEĞİ

Bu çalışma üç temel amaç üzerine kurulmuştur. Birinci amaç PISA 2006 çalışmasının mevcut Türkçe fen sorularının yapısal geçerliliğini pozitif ve negatif birimler bağlamında incelemektir. İkinci amaç fen soru ünitelerinde tespit edilen negatif birimlerin bütün test üzerinde ve her bir soru için etkisini 15 yaşındaki öğrencilerin cevaplama başarıları bağlamında araştırmaktır. Üçüncü amaç ise mevcut çalışmanın bir uzantısı olarak 15 yaşındaki öğrencilerin tek ve özel öğrenme deneyimleri çerçevesinde PISA 2006 fen soru üniteleri ile aşinalıklarını araştırmaktır.

Bu çalışma üç safhadan oluşmaktadır. İlk safha sekiz uyarıcı ve 25 sorudan oluşan sekiz adet fen soru ünitesini araştırmaktır. Fen soru ünitelerini analiz ederken, bütün soru ünitelerindeki uyarıcı ve sorular için içerik analizi tekniği kullanılmıştır. İstanbul ilinin değişik bölgelerinde çalışan öğretmenlerden (ortaöğretim kademesinden toplam 80 öğretmen) görüş almak için ve görüşlerin içerik analizinden kategoriler oluşturabilmek için bir Madde Dereceleme Formu (IRF) geliştirilmiş ve uygulanmıştır. İçerik analizi tekniği ile beş ana negatif kategori (dil, tipiklik, taslak, yapı ve içerik) ve dört ana pozitif kategori (bağlam, içerik, komposizyon ve fen süreçleri) oluşmuştur. Kavramsal analiz birim sayısı negatif birimde pozitif birimlerden daha fazla bulunmuştur.

Çalışmanın ikinci safhasını PISA 2006 fen soru üniteleri içindeki uyarıcı ve soruların düzeltmeleri oluşturmaktadır. Bu iyileştirmeler içerik analizi sonucunda ortaya çıkan negatif birimlerin uzmanlar tarafından uygun görülenleri için yapılmıştır. Çalışmanın ikinci safhasında iki test kullanılıştır. Birinci test PISA-Orijinal Türkçe testi (PISA-OT test) adını almıştır ve 2006 PISA çalışmasında kullanılan fen sorularını içermektedir. Diğer test ise değiştirilen fen soru ünitelerini içeren PISA-İyileştirilmiş Türkçe testidir. İkinci safhanın sonucu olarak, negatif birimlerin testin bütününde öğrencilerinin başarı seviyeleri üzerinde

istatistiksel olarak anlamlı bir fark oluşturduğu bulunmuştur. Bununla beraber, soru maddeleri ayrı ayrı incelendiğinde karşılaştırma gruplarının başarıları arasında 11 soru için anlamlı bir fark bulunurken kalan 11 soru için anlamlı bir fark bulunamamıştır. Son olarak, testlerde kullanılan 22 sorunun tamamı için PISA-RT testine cevap veren grubun bütün sorular için ortalama değerlerinin PISA-OT testine cevap veren grubunkinden daha yüksek olduğu bulunmuştur.

Çalışmanın üçüncü safhası için öğrencilerin PISA fen soru ünitelerinin uyarıcılarına dair dil, taslak, okul bilgisi ve günlük tecrübe bakımlarından aşinalık dereceleri araştırılmıştır. Sonuç olarak, Türk öğrencilerin PISA sorularındaki uyarıcılara orta derecede aşina oldukları bulunmuştur.

Çalışmanın her üç safhasının sonuçlarına dayanarak, fen okur yazarlığını ölçmeyi amaçlayan bir test için soruların oluşturulma ve seçilme süreçlerinin yapı geçerliği bakımından önemli olduğu sonucuna varılabilir.Bununla beraber, bu çalışmada Türkçe PISA soru ünitelerindeki uyarıcı ve soruların bu amacı kısmi olarak sağlayabildiği görülmüştür.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS / ABBREVIATIONS

f	Frequency
M	Mean
N	Number
ACID	ACID RAIN Science Unit
CLOTHES	CLOTHES Science Unit
GMC	GENETICALLY MODIFIED CORPS Science Unit
GRAND	GRAND CANYON Science Unit
GREEN	GREENHOUSE Science Unit
IRF	Item Rating Form
MARY	MARY MONTAGU Science Unit
PHYSICAL	PHYSICAL EXERCISE Science Unit
PISA-OT	Test including the Original Turkish version of PISA 2006 science units
PISA-RT	Test including the Revised Turkish version of PISA science units
SOS	Student Opinion Survey
SUN	SUNSCREENS Science Unit

1. INTRODUCTION

In the past decades international studies of educational achievement have drawn great attention in many countries. Educational researchers, policy makers, teachers, parents, and anyone concerned about education in general are all interested in the results of these studies, mainly for the purpose of comparison. Additionally, these international assessments have provided feedback for each country. This feedback is used as the outcomes of education in relation to inputs to education systems like curricular materials and teacher training.

The Programme for International Student Assessment (PISA) is among the most comprehensive international comparative education studies to gather data about the educational systems of the different countries by measuring mathematics, science, and reading literacy of achievement of 15 years-old students. PISA presents an extensive data base for researches including methodologies, data collection tools, achievement scores and so on. However, the results have prompted public debate in some countries, about the educational, institutional, economical, and social reasons for the apparent ranking of some countries over the others. In some countries, like Germany, Norway and Canada, PISA outcomes have gained great importance for the critics of the education systems and national assessments (Lingens, 2005; Knain and Turmo, 2003; Prais, 2007). The results of PISA 2006 showed that Turkish students performed poorly in science literacy. Turkish students' performance was significantly lower than the average (OECD, 2007). The poor performance of Turkish students can be explained by several demographic, school, teacher, and instruction related factors. However, there might be reasons that stem from the test itself rather than these factors.

Mullis, Martin, Gonzales, Gregory, Garden, O'Connor, Chrostowski and Smith (2000) emphasize the necessity of fairness when comparing student achievement across countries. From the measurement perspective, drawing inferences and making comparisons from such large scale, cross national studies is based on a critical assumption that the study itself has adopted proper methodology and has acceptable

validity. The task of measuring students' educational achievement is always challenging. Conducting international assessments add even more challenges to the task.

The present study is an attempt to gain deeper understanding of an international study at the item level. The PISA science items present an opportunity to investigate the possibilities and limitations of large-scale test design to capture and report meaningfulness of the results to be used in researches, educational policy decisions, and classroom applications. Specifically, the researcher investigated construct validity of Turkish version of released PISA 2006 science items in terms of positive and negative entities embedded within the single items and effects of the negative entities on the achievement of the students together with the familiarity of students with the stimuli in terms of their unique learning experiences. Literature review part provides the theoretical rationale for this study. Methodology part of the study describes the procedures involved in the study of positive and negative entities in terms of construct validity. In the results part of the present study, descriptions of the findings for each phase of the study are presented. The discussion and conclusion part discusses the study's findings in the light of existing research together with applications for educational practices and recommendations for further researches.

2. LITERATURE REVIEW

The purpose of this study is to investigate positive and negative entities embedded within the science units of PISA 2006 study together with investigating effect of negative entities on total test and at item level with respect to students' achievement levels. In order to understand PISA study and the concepts of construct validity the literature review is organized around three major parts. These three main parts are named as *(a) background of the programme for international student assessment (PISA)* which aims to introduce PISA as an international assessment study, *(b) PISA 2006 overview* which constitutes base for the present study, *(c) validity concept with a special emphasis on the construct validity* in relation the design of the present study.

The first part of the literature review presents an introduction of the international assessment concept, most common three international assessment studies and the cycles of PISA carried out in the last nine years. The second part of the review concentrates on the PISA 2006 study and its features with detailed descriptions of different elements within study such as scientific literacy concept in comparison with the previous two PISA studies, framework for the context, competencies, knowledge and attitude elements of the assessment. Beside this, the second part covers the related researches. The third part of the literature review first introduces the concept of validity and its development process and then concentrates on construct validity concept by combining the first two parts of the literature review by studying construct validity in international assessment together with research studies on the subject.

2.1. The Programme for International Student Assessment (PISA): Background

The literature review on the background of the PISA study explains the developmental process of the international assessments. Also, the review of this part focuses on the general features of the PISA studies.

Over the few decades the interest in the international comparison of educational systems has led to the growth of large-scale international assessment studies. The popularity of international studies aiming to gather data about the education of different countries has been increasing. As Kellegan and Greaney (2001) point out, there is "global assessment" is being created.

The First International Mathematics Study (FIMS) can be seen to form the origin of international comparative assessment studies. It was conducted between 1961 and 1965 with underlying intention of comparing outcomes of different educational systems. Recently, there are three most influential international studies. These are Trends in International Mathematics and Science Study (TIMSS, formerly known as the Third International Mathematics and Science Study), Progress of International Reading Literacy Study (PIRLS) and Programme of International Student Assessment (PISA). These studies are carried by two organizations, one of them is the International Association for the Evaluation of Educational Achievement (IEA) which conducts TIMSS & PIRLS and the other is the Organization for Economic Cooperation and Development (OECD) which organizes PISA. These assessment studies contributes to gathering comprehensive data on the performance of students and background information about the education systems across the world in developed and developing countries (Mullis and Martin, 2006).

PISA differs from the other two studies in terms of aim, context, task, sampling, and collected information. According to Lange (2006), in a simplified way, TIMSS is a grade-based study, that is, the sample composed of students of Grade 4 and Grade 8, whereas PISA is age-based study, the test is for 15-years-old students. The test items in

TIMSS are more content or standards based, while those in PISA are more literacy orientated. TIMSS proposes to assess how much students have achieved in schools, whereas PISA claims to assess how well students are prepared for the outside world. The remaining part of the literature review is based on the properties of PISA study that these differences are clarified.

According to Stedman (1997), the increasing interest in international assessments, at a time when the link between education and economic well-being is blurred, is puzzling. Robitaille and Garden (1989) express a common belief about this link:

> That the nation's continued economic well being and its ability to compete in the global market place" are strongly linked to how well that countries students do in international tests-particularly in science and mathematics". (p.18)

In a parallel assignation, Organization for Economic Co-operation and Development (OECD) has started a large scale programme of assessing student achievement that aims to improve the quality of education of its member countries as well as to provide information about the educational outcomes of the educational systems of these countries (OECD 2001). The baseline for the study included derivation of economic as well as societal and cultural success from the human capital and linked opportunities for continuous learning.

The relationship, described by OECD, between learning, human capital, and economic well-being of countries constitutes the basics of the Programme for International Student Assessment (PISA), which aims not measure the school curriculum in different countries; but to measure the students' ability to understand concepts and use their knowledge to function in various situations within three main domains (reading, mathematics and science). Prominence is on the mastery of processes, the understanding of concepts, and the ability to function in various situations within each assessment domain. (OECD 1999, 2000, 2003 and 2006)

The Programme for International Students Assessment (PISA) was initiated in 1997 mainly for member countries of the OECD and in years increasing number of non-member partner countries were involved in the study. PISA declares its aim to measure

the essential skills and knowledge extent of students approaching to the end of compulsory education. The focus is on the full participation in a knowledge society and providing empirically qualified information which will accustom policy decisions. It points to regularly observing the outcomes and the progress of education systems in terms of 15 year-old students' achievement. In addition to monitoring student achievement, PISA seeks policy understanding in three ways:

- Bringing to a more advanced or effective better ways of observation of student progress, detecting the comparison between primary education and the age of 15
- Developing a closer look between performance and instruction
- Making use of computer-based assessments (OECD, 2007)

Programme for International Student Assessment (PISA) is an international comparative assessment that is repeated every three years. Scientific literacy is one of the three domains, reading and mathematical literacy is the other two. So far there assessments have been implemented; third PISA is carried out on a three-year cycle. The first PISA study was in 2000 (supplemented in 2002), and this was repeated in 2003 and 2006. The next survey will be in 2009. The survey was undertaken in 43 countries in the first cycle (32 in 2000 and 11 in 2002) and 41 countries in the second cycle (2003). In the third cycle, 57 countries participated, including all 30 OECD members. Each PISA study focuses on one of the three areas of literacy in which knowledge and skills are assessed: reading, mathematics, and science. The main focus for 2000 was reading, for 2003 mathematics, and for 2006 was science. In each three-yearly PISA studies, one subject was chosen as a focus while two other subject areas have been assessed more briefly, for example in 2006 reading and mathematics were the minor domains (see Figure 2.1) As it is claimed by PISA, the countries participated in the study has covered a large area on the world.

2000	2003	2006	2009	2012	2015
Reading	Reading	Reading	Reading	Reading	Reading
Maths	Maths	Maths	Maths	Maths	Maths
Science	Science	Science	Science	Science	Science
	Problem solving				

Figure 2.1. PISA literacy components over years

There is a large amount of documents written and displayed all around the world about the PISA studies. Some of them directly published by the OECD and by the PISA consortium and many others by official centers and researchers in participating countries. There is rich and contrasted information. Some of documents are public like PISA general design, the frameworks, database, and the international reports. However, PISA does not release the entire set of questions in study, the reason for this secrecy is that the items will be used in the next PISA testing round, and therefore they may not be made public (Hopmann, 2006).

Since the beginning of the 1960s, international education studies have been designed to support educators, researchers, policy markers and others with the information about students' achievement and the performance of different educational systems. It can not always easily be seen how to use the results of these studies. However, there is evidence that they often attract a great deal of attention especially when a country's results are poor (Colvin, 1996). Because educational policy markers, researchers, teachers, and parents are interested in the results for comparison, the validity concept for fair inferences from this large-scale, cross national study has increasingly become critical.

As it is stated previously, PISA 2006 survey completes the first cycle of assessment of three major areas - reading (PISA 2000), mathematics (PISA 2003) and science (PISA 2006).

Since the science literacy test in PISA 2006 is selected for reviewing in the present study, more comprehensive revision of the PISA 2006 study is found to be necessary in terms of scientific literacy definition, the contexts in which science items are embedded within, the competencies that students need to handle, the knowledge domains chosen, and students' attitudes.

2.2. An Overview on PISA 2006

This part of the literature review presents an overview about the PISA 2006 study and explanations of its features to clarify the reasons for the present study.

PISA 2006 focused on students' competency in science and emphasized the importance of understanding fundamental scientific concepts and theories and the ability to structure and solve scientific problems in a technology based world. Since some research findings suggest (e.g. Fullilove, 1987; Çavaş, 2004) that student attitudes towards science can have an important role in the decisions of students to study science and technology in university, PISA study contains assessing both students' knowledge and skills and also students' attitudes toward science (OECD, 2006). However, because the present study concentrates on only the features of the items in the science units of the PISA study, it excludes the attitude questions related with the items.

The PISA framework starts with the concept of 'literacy' which is concerned with the capacity of students to deduce from what they have learned, and to apply their knowledge in novel settings, and students' capacity to figure out, draw conclusions and communicate effectively as they pose, solve, and interpret problems in a diversity of situations. A brief coverage with an adaptation from OECD (2007, p. 48) of PISA 2006 features can be seen from Figure 2.2.

Content

- *The main focus of PISA 2006 is science.*
- *The PISA 2006 study also collected data on students' attitudes toward science within the test itself after the cognitive questions.*

Methods

- *There are around 400 000 students in PISA 2006 to represent about 20 million 15-year-olds in the schools of the 57 countries.*
- *PISA is a pencil-and-paper test and students are expected to undertake two hours to complete the assessment.*
- *PISA questions are in the form of units which includes more than one question related to the unit context*

Outcomes

- *A profile of knowledge and skills among 15-year-olds in 2006 with a profile for science, and also for reading and mathematics.*
- *Contextual indicators relating performance results to other characteristics of students and schools.*
- *Data on the students' attitudes to science.*
- *A knowledge base for policy analysis and research.*
- *Trend data on changes in student knowledge and skills in reading and mathematics.*

Future assessments

- *The PISA 2009 survey will return to reading as the major assessment area, while PISA 2012 will focus on mathematics and PISA 2015 once again on*

Figure 2.2. Key features of PISA 2006

The present study focused on construct validity of the released PISA 2006 science items, in order to explain the definition of the scientific literacy construct with its application on the PISA studies and clarify the reason underlying why to choose PISA 2006 science items rather than covering all of the released items from the three PISA

studies, the next section presents the details and structure of the PISA framework for scientific literacy.

2.2.1. Scientific Literacy in PISA 2000, 2003 and 2006 Cycles

The term of scientific literacy appeared in the last century, and nowadays, scientific literacy has been accepted as an internationally well- organized educational slogan, and a contemporary educational goal that education practices such as standardized testing and the selection of content in textbooks, revisions of science curricula are based on the interpretation of this concept.

As Shamos (1995) states, interest in the concept of scientific literacy comes from the beginning of the century. The construct of scientific literacy has developed through a century and subjected various interpretation and debate. Agin (1974) designed a framework of scientific literacy by reviewing the literature and this included six categories: science and society, science and technology, the ethics of science, the nature of science, the concepts of science, and the science of the humanities. Durant (1993) offers a possible short definition of scientific literacy to be "what the general public ought to know about science". Laugksch (2000) claims that it is commonly accepted that these kinds of simple conceptualization of science literacy overshadow different meanings and interpretations associated with the concepts of scientific literacy because of, for example, different views of what the public ought to know about science and who "the public" is. Bybee summarized the definition of scientific literacy by using the principals of different views into four parts. These were;

- scientific knowledge,
- the nature of science,
- the processes of science, and
- the social and cultural implications of science" (1997, p. 56).

Scientific literacy definition of PISA includes several pieces from the literature. In addition to this, the definition changes within the PISA context, which is to mean, the scientific literacy definition of the three PISA studies are not exactly the same.

Compared to the definition of scientific literacy for PISA in 2000 and 2003, the definition for 2006 is more comprehensive. For PISA 2000 and 2003, the framework is based on the knowledge and skills required for adult life, defined as the "ability to undertake a number of fundamental processes in a range of situations, backed by a broad understanding of key concepts" (OECD, 2000, p.7). Scientific knowledge includes understanding facts, fundamental scientific concepts, the limitations of science, and the nature of science as a human activity (OECD, 1999). The initial assertions of the 2000, 2003 and 2006 definitions are mainly the same in that they focus on individuals' uses of scientific knowledge to draw conclusions. In PISA 2006 scientific literacy is defined as:

> An individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen (OECD 2006, p.12).

The definition of scientific literacy in PISA is grounded on three benchmarks. The first is related with the students' understanding of fundamental scientific concepts and theories, as well as the extent to which they can extrapolate from what they have learned in science and apply their knowledge to real-life problems. The other is students' interest in science, the value they place on scientific approaches to understanding the world and their willingness to engage in scientific enquiry. As the last point, students' school contexts including the socio-economic background of school peers and other factors that research suggests are associated with student achievement. In the context of the present study, the focus is on the first benchmark rather than the last two.

In terms of individuals' competencies rather than the policy understanding for the education systems as mentioned before, PISA 2006 defines *scientific literacy* in three main categories in relation with the students' capability. These are;

- Scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues. Understanding of the characteristic features of science as a form of human knowledge and enquiry

- Awareness of how science and technology shape our material, intellectual and cultural environments
- Willingness to engage with science-related issues, and with the ideas of science, as a reflective citizen **(**OECD 2007, p.25).

PISA 2006 develops its science assessment tasks and questions within a framework of four interrelated aspects: the contexts in which tasks are embedded, the competencies that students need to apply, the knowledge domains involved, and student attitudes. Figure 2.3 shows the interrelation between these aspects. In the following pages, the four aspects will be explained separately in more detail.

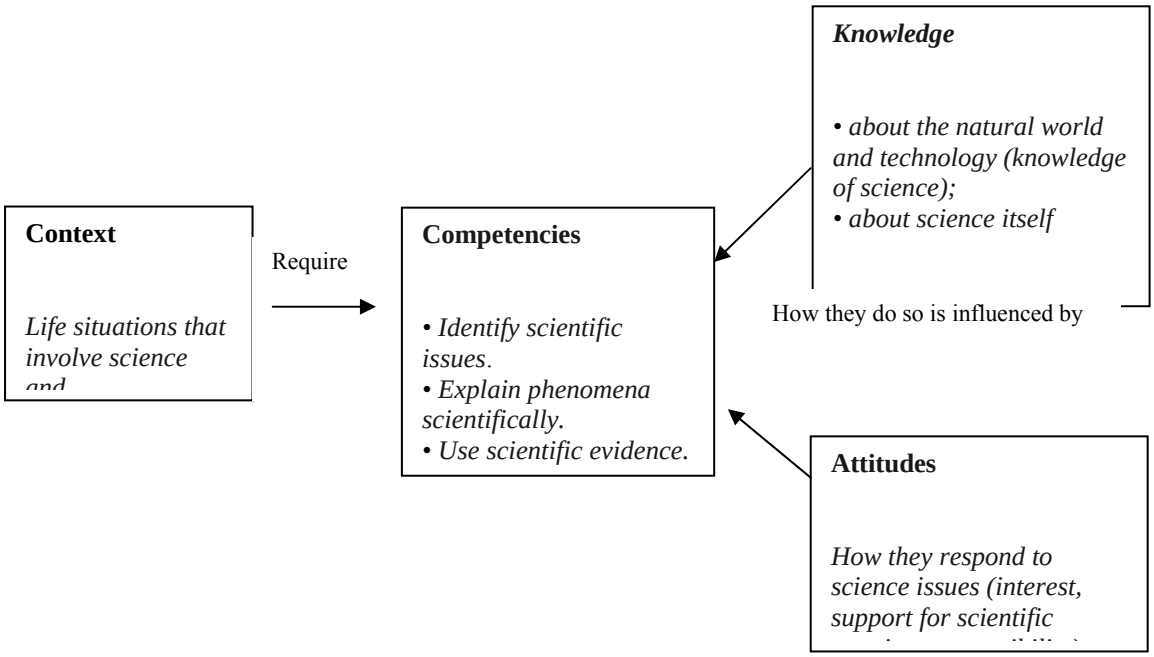


Figure 2.3. Framework for PISA 2006 science assessment (OECD, 2006)

According to Figure 2.3, all of the three aspects (context, knowledge and attitudes) will lead to the competency aspect in different ways. Context aspect requires people to achieve three competencies of identifying scientific issues, explaining phenomena scientifically, and using scientific evidence. Knowledge and attitude are the aspects that people are influenced when they trying to achieve competencies. More detailed definitions of the aspects will be given below as *a) context, b) competencies* and *c)*

knowledge. Since the *attitude* aspect is beyond the concept of the present study, it will not be explained in detail.

The *context* aspect of the PISA study is presented in several ways that the PISA 2006 science questions are designed to form within "health", "natural resources", "environmental quality", "hazards", and "frontiers of science and technology" situations and these situations were related to three major contexts: *personal* (the self, family and peer groups), *social* (community) and *global* (life across the world). As an adapted version of OECD (2007, p.86) it is exemplified in Figure 2.4, there are intersections between the situations and context and the related daily life examples which they shaped the questions. The contexts used for questions are mentioned by OECD (2007) to be chosen in the light of relevance to students' interests and lives, representing science-related situations that adults meet.

	Personal Self, family and peer	**Social** The community	**Global** Life across the world
Health	Maintenance of health, accidents, nutrition	Control of disease, social transmission, food choices	Epidemics, spread of infectious diseases
Natural Resources	Personal consumption of materials and energy	Maintenance of human populations, quality of life, security, production and distribution of food,	Renewable and nonrenewable, natural systems, population growth, sustainable use of species
Environment	Environmentally friendly behavior, use and disposal of materials	Population distribution, disposal of waste, environmental impact, local weather	Biodiversity, ecological sustainability, control of
Hazard	Natural and human	Rapid changes	Climate change,

	induced, decisions about housing	(earthquakes, severe weather)	impact of modern warfare
Frontiers of science and technology	Interest in science's explanations of natural phenomena, science based hobbies, sport and leisure, music and	New materials, devices and processes, genetic modification, transport	Extinction of species, exploration of space, origin and structure of the universe

Figure 2.4. Examples from PISA 2006 Science Context

From the point of *competencies*, the PISA 2006 declares its aim to produce science questions to ask students identifying scientific issues, explaining phenomena scientifically and using scientific evidence. These selected three competencies have relation to the scientific literacy literature in practice of science and connection to key cognitive abilities such as inductive/deductive reasoning, systems-based thinking, critical decision making, transformation of information (*e.g.* creating tables or graphs out of raw data), construction and communication of arguments and explanations based on data, thinking in terms of models, and use of science. OECD (2007) describes the key features of each of the three science competencies as below;

Identifying scientific issues

- Recognizing issues that are possible to investigate scientifically
- Identifying keywords to search for scientific information
- Recognizing the key features of a scientific investigation

Explaining phenomena scientifically

- Applying *knowledge of science* in a given situation
- Describing or interpreting phenomena scientifically and predicting changes
- Identifying appropriate descriptions, explanations, and predictions

Using scientific evidence

- Identifying the assumptions, evidence and reasoning behind conclusions
- Reflecting on the societal implications of science and technological development

The *knowledge* aspect has two minor aspects as *knowledge of science* (knowledge of the different scientific disciplines and the natural world) and *knowledge about science* as a form of human enquiry. *Knowledge of science* includes understanding basic scientific concepts laws, and theories; *knowledge about science* focuses on the understanding the nature of science. PISA 2006 science questions assess knowledge of science and knowledge about science. At the point of priorities to include in the PISA assessment and restrictions of the content to cover for assessment, PISA mentions the relevancy to real-life situations, representativeness of important scientific concepts and thus of enduring utility and appropriateness to the developmental level of 15-year-olds criteria to be most essential. Content areas for both *knowledge of science* and *knowledge about science* domains presented at the next page as they are described by PISA (OECD, 2007).

Content areas for the *knowledge of science* domain are described as below.

Physical systems

- Structure of matter (e.g. particle model, bonds)
- Properties of matter (e.g. changes of state, thermal and electrical conductivity)
- Chemical changes of matter (e.g. reactions, energy transfer, acids/bases)
- Motions and forces (e.g. velocity, friction)
- Energy and its transformation (e.g. conservation, dissipation, chemical reactions)
- Interactions of energy and matter (e.g. light and radio waves, sound)

Living systems

- Cells (e.g. structures and function, DNA, plant and animal)

- Humans (e.g. health, nutrition, disease, reproduction, subsystems)
- Populations (e.g. species, evolution, biodiversity, genetic variation)
- Ecosystems (e.g. food chains, matter, and energy flow)

Earth and space systems

- Structures of the Earth systems (e.g. lithosphere, atmosphere, hydrosphere)
- Energy in the Earth systems (e.g. sources, global climate)
- Change in Earth systems (e.g. plate tectonics, geochemical cycles)
- Earth's history (e.g. fossils, origin and evolution)
- Earth in space (e.g. gravity, solar systems)

Technology systems

- Role of science-based technology (e.g. solve problems)
- Relationships between science and technology (e.g. technologies)
- Concepts (e.g. optimization, trade-offs, cost, risk, benefit)
- Important principles (e.g. criteria, constraints, cost, problem solving)

Categories for the *knowledge about science* domain can be summarized as below.

Scientific enquiry

- Origin (e.g. curiosity, scientific questions)
- Purpose (e.g. to produce evidence that helps answer scientific questions)
- Experiments (e.g. different questions suggest different scientific investigations)
- Data (e.g. quantitative [measurements], qualitative [observations])
- Measurement (e.g. inherent uncertainty, replicability)
- Characteristics of results (e.g. empirical, tentative, testable)

Scientific explanations

- Types (e.g. hypothesis, theory, model, scientific law)
- Formation (e.g. existing knowledge and new evidence)
- Rules (e.g. logically consistent, based on evidence)

According to the aspects defined above, the items included in the PISA studies are developed. In order to clarify properties of the items based on these aspects and to understand the structure of the PISA test, a closer look will be taken of the PISA test design, science units and science items.

Additionally, the PISA 2006 science units and test design is worthy to mention about that construction of PISA 2006 science units are carried out with the guidance of an international expert panel based on input and expertise from the participating countries to cover the various aspects of the framework described above: contexts, competencies, knowledge, and attitudes. The science questions used in the assessment were developed based on material submitted by the participating countries. Questions were presented in the form of science units. *Science unit* involves a group that included some type of stimulus, which is then followed by a number of questions. Each PISA test question is characterized by its context, the competencies it brings out, and the knowledge domain it represents. In each unit, the context is represented by the stimulus material – usually a brief written passage or text accompanying a table, chart, graph, photographs, or diagram. As it is claimed, each question requires students to use one or more of the science competencies as well as knowledge of science and/or knowledge about science (OECD, 2007).

In order to explain the extent of this study, it is noteworthy that only a small number (25 questions) of the questions will be used in the study. The main reason for this restriction is the policy of PISA to measure achievement trend of the countries. PISA does not release all of the questions used in the assessments. It means a number of questions are kept to be used in the other surveys to gather data on the trends of countries. The remaining questions are released after the survey to illustrate the ways in which performance was measured. So as to clarify the meaning of the scientific literacy as major area of the PISA 2006 in terms of the number of questions, and indicate the ratio of the questions used in the present thesis study to the whole, the Table 2.1 is given on the following page.

Table 2.1. Distributions of science items in 2000, 2003 and 2006 PISA

	PISA 2000			PISA 2003			PISA 2006		
Item format / Competency	Closed items	Open items	Total	Closed items	Open items	Total	Closed items	Open items	Total
Explaining phenomena scientifically	12	4	16	11	5	16	37	15	52
Identifying scientific issues	6	3	9	5	3	8	20	5	25
Using scientific evidence	5	5	10	6	5	11	16	15	31
Total	23	12	33	21	13	35	73	36	108

As it is seen from the Table 2.1, there is an increase in the number of science questions in years. Because the major domain of the PISA 2006 was science literacy, the number of the questions was nearly triple that of PISA 2003. There were 108 science items used in PISA 2006, compared with 35 in PISA 2003 and 33 in PISA 2000.

Table 2.2 shows a map of these released PISA 2006 science questions. For each of the three science competencies, the released questions and scores (shown in parentheses after each question) have been ordered according to difficulty, with the most difficult at the top and the least difficult at the bottom. The difficulty levels described in detailed in Appendix A.

Table 2.2. A map of released science questions in PISA

Level /Score limit	Identifying scientific issues	Explaining phenomena scientifically	Using scientific evidence
6 /707,9	Acid Rain Q 5.2 (717)	Greenhouse Q 5 (709)*	
5/633,3			Greenhouse Q 4.2 (659)*
4/558,7	Sunscreens Q 4 (574)* Q 2 (588)* Clothes Q 1 (567)*	Physical Exercise Q 5 (583)*	Sunscreens Q 5.2 (629)* Q 5.1 (616) * Greenhouse Q 4.1 (568)*
3/484,1	Acid rain Q 5.1 (513)* Sunscreens Q 3 (499)* Grand Canyon Q 7 (486)*	Physical Exercise Q 1 (545)* Acid Rain Q 2 (506)* Mary Montagu Q 4 (507)*	Greenhouse Q 3 (529)*
2/409,5	Genetically Modified Corps Q 3 (421)*	Grand Canyon Q 3 (451)* Mary Montagu Q 2 (436)* Q 3 (431)* Grand Canyon Q 5 (411)*	Greenhouse Q 3 (460)*
1/334,9		Physical Exercise Q 3 (386)* Clothes Q 2 (399)*	

*Numbers in brackets refer to the difficulty level of the question where students may receive full or partial credit is also indicated.

So as it is seen that there are 25 science questions released and they are classified under the six proficiency levels. All of these 25 items included first phase of the study. Three of the items were not used at the second phase of the study because two of them are not categorized under any of the proficiency levels by PISA, they are below the Level 1, and one of them was announced not to be included in score calculation processes.

2.2.2. The situation of Turkey in PISA 2006

This section gives a brief view on the Turkey's participation in PISA studies and specifically, results were taken from PISA 2006.

Because the pre-application of PISA 2000 was missed, Turkey could not participate PISA 2000 that PISA 2003 was the first time for Turkey to be included in PISA study. According to results of PISA 2003, Turkey ranked thirty sixth among 41 participating countries in the science major. The National Center for Educational Research and Development Directorate (EARGED) was responsible for the implementation of PISA in the Turkey. PISA 2006 was administered in May 2006. The Turkey sample included both public and private schools, randomly selected and weighted to be representative of the several types of school. In total, 160 schools and 4942 students from 51 cites participated in PISA 2006 in Turkey (EARGED, 2007). Distribution of student numbers according to school types can be seen from Table 2.3.

Table 2.3. Distribution of Turkish 15-year-old students, by school type: 2006

School Type	No. of Students	Percentage(%)of Students
Primary School	116	2.3
General High School*	2266	45.9
Anatolian High S.	549	11.1
Foreign Language Weighed High S.	9	0.2
Science Lice	35	0.7
Vocational High S.	1510	30.6
Anatolian Vocational High S.	179	3.6
Multiple Programmed High S.	278	5.6
TOTAL	4942	100.0

* Private schools are supposed to be in General High School

In PISA 2006, combined science literacy scores are reported on a scale from 0 to 1,000 with a mean set at 500 and a standard deviation of 100[1] Fifteen-year-old students in Turkey had an average score of 424 on the combined science literacy scale, lower than the OECD average score of 500. As it is seen from the Table 2.4, Turkish students scored lower in science literacy than their peers in 28 of the other 29 OECD jurisdictions and 15 of the 27 non-OECD jurisdictions.

Table 2.4. Range of rank of the countries on the different science scales

Science scale						
	Mean score	S.E.	Range of rank			
			OECD countries		All countries	
			Upper Rank	Lower Rank	Upper Rank	Lower Rank
Finland	563	(2,0)	1	1	1	1
Hong Kong-China	542	(2,5)			2	2
Canada	534	(2,0)	2	3	3	6
Chinese Taipei	532	(3,6)			3	8
Estonia	531	(2,5)			3	8
Austria	511	(3,9)	8	15	12	21
Belgium	510	(2,5)	9	14	14	20
Ireland	508	(3,2)	10	16	15	22

[1] The combined science literacy scale is made up of all items in the three subscales. However, the combined science scale and the three subscales are each computed separately through Item Response Theory (IRT) models. Therefore, the combined science scale score is not the average of the three subscale scores.

Hungary	504	(2,7)	13	17	19	23
Sweden	503	(2,4)	14	17	20	23
Poland	498	(2,3)	16	19	22	26
Denmark	496	(3,1)	16	21	22	28
France	495	(3,4)	16	21	22	29
Croatia	493	(2,4)			23	30
Iceland	491	(1,6)	19	23	25	31
Israel	454	(3,7)			39	39
Chile	438	(4,3)			40	42
Serbia	436	(3,0)			40	42
Bulgaria	434	(6,1)			40	44
Uruguay	428	(2,7)			42	45
Turkey	424	(3,8)	29	29	43	47
Jordan	422	(2,8)			43	47
Thailand	421	(2,1)			44	47
Romania	418	(4,2)			44	48
Montenegro	412	(1,1)			47	49
Mexico	410	(2,7)	30	30	48	49
Indonesia	393	(5,7)			50	54
Argentina	391	(6,1)			50	55
Brazil	390	(2,8)			50	54
Colombia	388	(3,4)			50	55
Tunisia	386	(3,0)			52	55
Azerbaijan	382	(2,8)			53	55
Qatar	349	(0,9)			56	56
Kyrgyzstan	322	(2,9)			57	57

Along with scale scores, PISA 2006 also uses six proficiency levels (Levels 1 through 6, with Level 6 being the highest level of proficiency) to describe student performance in science literacy. An additional level (below Level 1) encompasses students whose skills cannot be described using these proficiency levels. The proficiency levels describe what students at each level should be able to do and allow comparisons of the percentages of students in each jurisdiction who perform at different levels of science literacy (OECD, 2007).

Table 2.5. Percentage distribution of 15-year-old students in Turkey: PISA 2006

PISA 2006	Average	Below Level 1	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Turkey	424	12.9	33.7	31.3	15.1	6.2	0.9	0.0
OECD Overall	491	6.9	16.3	24.2	25.1	18.7	7.4	1.4
OECD Average	500	5.2	14.1	24.0	27.4	20.3	7.7	1.3

According to results of PISA 2006, Turkey has a greater percentage of students at or below Level 1 and Level 2 than the OECD average percentages on the combined science literacy scale. Turkey also has lower percentages of students at Levels 3, 4 and 5 than the OECD average percentages. The percentages of Turkey students performing at Level 6 are rounded to be zero. Turkey is announced to be at Level 2 on average and Turkish students are described by OECD (2007) as:

> Have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They should be capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem solving". (p.43)

The reasons underling for this classification for Turkish students can be explained in terms of several variables such as student demographics, attitudes, parents' socio-economic situation etc. In addition to these, researcher as a science teacher tries to investigate the item level reasons if they present.

2.2.3. PISA in Turkey and some other countries

Not only the research studies (e.g. Sjoberg, 2007; Acar, 2008) but also popular media coverage of PISA results create the public perception of the quality of a country's overall school system. In the absence of meaningful critics, the rankings of countries

play the most important role in the decisions related with the education systems of the countries.

According to the Nóvoa and Mashal (2003), research like PISA has critical conclusions that they give rise to the definitions of 'good' or 'bad' educational systems, and lead to seeking solutions to results. Moreover, the mass media are keen to diffuse the results of these studies, in such a manner that reinforces a need for urgent decisions, following lines of action that seem to be carried out without arguing the process because they have been internationally asserted.

PISA study shapes international educational policies and also influences national policies in most of the participating countries. Moreover, the PISA results provide media and the public with convincing images and perceptions about the quality of the school system, the quality of their teachers' work and the characteristics of both the school population and future citizen (e.g. Ziman, 2000; Sjoberg, 2007)

As Sjoberg (2007) mentions, PISA results cause hot debates in Norway and Germany and have great effect on the educational policy. In Norway, PISA results have been presented in the media with war-like headings, shaping public perception about the national school system and PISA results presented in the leading Norwegian newspaper Dagbladet with the heading "Norway is a school loser". Similarly, in Germany, the political agenda and the public image of the quality of the entire school system have been formed by the PISA results. There is evidence that PISA has provided - and will continue to provide - results, ideologies, concepts, analysis, advice and recommendations that will shape future of educational debates and reforms, nationally as well as internationally.

In Turkey, there is similar public attention on the results of the PISA study that the results from PISA 2000, PISA 2003 as well as from PISA2006 provided big headings in most national newspapers.

Eğitimin hali harap

OECD raporu kırıklarla dolu: Türk eğitim sistemi, bilimsel ve ekonomik gerçeklere duyarlı değil. Müfredatta eksik çok. Eğitim altyapısı da zayıf

08/12/2004 (1198 kişi okudu)

RADİKAL - ANKARA - Kalkınma ve Ekonomik İşbirliği Topluluğu'nun (OECD) kısa adı PISA olan Uluslararası Öğrenci Başarısını Belirleme Programı, Türkiye'deki eğitim sisteminin eksiklerini gözler önüne serdi. Milli Eğitim Bakanlığı'ndan (MEB) itiraf: Türk eğitim sistemi, bilimsel ve ekonomik gerçeklere duyarlı değil. Müfredatı eksikliklerle dolu olan Türkiye, öğrenci sayısı, eğitimin bütçedeki payı ve araştırmaya ayrılan payın azlığının yanı sıra kişi başına düşen milli gelir açısından da dezavantajlı durumda.
PISA'nın 2000-2003'ü kapsayan araştırmasında bir kısmı OECD üyesi 41 ülkedeki 15 yaş grubu öğrencileri karşılaştırdı. Öğrencilerin zorunlu eğitimin sonunda, gerçek hayatta karşılaşabilecekleri durumlarda sahip oldukları bilgi ve becerileri kullanabilme yetenekleri, düşüncelerini analiz edebilme, akıl yürütme ve okulda öğrendikleri fen ve matematik kavramlarını kullanarak etkin iletişim kurma becerisine sahip olup olmadıkları değerlendirildi.

OECD'nin PISA Programı kapsamında 41 ülkede öğrencilerin matematik, fen bilimleri, okuma ve problem çözme alanlarında bilgi ve becerileri ölçüldü.

Uluslararası Öğrenci Değerlendirme Programı sonuçlarına göre

MİLLİ EĞİTİM sınıfta kaldı

30'u OECD ülkesi olmak üzere toplam 41 ülkede 15 yaş grubu öğrenciler iki saatlik bir teste tabi tutuldu. Türkiye matematik, fen ve okuma-yazmada kötü bir performans sergiledi.

Türkiye'nin eğitim alanında Avrupa ülkelerine erişmek için uzun bir yola ihtiyacı olduğunu söyleyenlerin doğru bir tespit yaptıkları. Uluslararası Öğrenci Değerlendirme Programı (PISA) sonuçlarıyla ortaya çıktı. 30'u OECD ülkesi olmak üzere 41 ülkede, 15 yaş grubundaki 250 binden fazla öğrenci üzerinde

PISA eğitim raporunda Türkiye yine sınıfta kaldı

Türkiye, Ekonomik İşbirliği ve Kalkınma Teşkilatı (OECD) tarafından üç yılda bir yapılan Uluslararası Öğrenci Değerlendirme Programı`nın (PISA) yeni araştırmasında 44. oldu.

OECD tarafından bugün açıklanacak Zaman`ın ele geçirdiği PISA 2006 raporunda, Türkiye 424 ülke ortalamasıyla; Hırvatistan(493), Sırbistan(436), Bulgaristan(434) ve Uruguay (428) gibi ülkelerin bile gerisinde kalarak OECD ortalamasının çok altında yer aldı. 57 ülkede, 15 yaş grubunda yirmi milyon öğrenciyi temsil eden dört yüz bin öğrenci üzerinde yapılan araştırmanın sonucunu yansıtan PISA 2006, öğrencilerin fen bilimleri alanlarındaki bilgi ve becerilerinin ölçümü üzerinde yoğunlaşıyor. Rapor, araştırmanın ana gövdesini oluşturan fen bilimlerinin yanında metin anlama ve matematik değerlendirmelerini de kapsıyor. Türkiye, her iki alanda OECD ortalamasının altında bulunuyor. Metin anlamada, Türkiye 447 ortalamayla İspanya(461) ve Yunanistan`ın (460) ardından

GÜNÜN MANŞETLERİ

- Şok iddia: Mehmetçik Gazze'ye gidiyor
- Alkollü şahıstan güldüren anons: Merkez be...
- Motorlu Taşıt Vergisi kredi kartı ile öden...
- Selçuklular aspirini 800 yıl önce keşfetmiş
- İşte Gökçek'in yeni hedefi !
- Hamas liderlerinden Rayyan öldürüldü
- Başkent, Filistin için ayaklanacak!
- Soros'un Açık Toplum Enstitüsü kapanıyor
- Demir-çelikte dünya 11'incisiyiz
- Trafik Ceza Puanlarına AB Standardı
- Yeni Yılda Sıkı Takip Geliyor
- Yerel seçimlere katılacak 19 parti!
- Çalışan, 2009'da kaç gün tatil yapacak?
- 7 üniversite öğrencisi sobadan öldü

Source http://www.tumgazeteler.com/?a=2405810

Figure 2.5. Examples from the Turkish media on the PISA studies results

In addition to these, PISA study is one of the important reasons underlining the changes in educational policy and curricula in Turkey that Ministry of Education mentions about the PISA results, especially PISA 2003, to form baseline for necessity to make a comprehensive reform in Turkish educational system (MEB, 2005).

From the point of educational research, PISA presents comprehensive data for Turkey and there are several researches related to the PISA results. Acar (2008) evaluates the competitive power of Turkey under the light of PISA results and focuses on the requirement of a new education reform by analyzing the PISA top performers'

reasons underling the achievement and necessity of rising achievement level in PISA to develop of human capital that will be participated industrial sector in the future. Berberoğlu (2004) presents the results form the PISA 2003 and concludes the school types to be the most effective factor for the difference of Turkish students to be in different competency levels. In a different study, Berberoğlu and Kalender (2005) investigates school and regional difference by using national students selection examination and PISA 2003 data, the study shows that school differences are greater than the regional ones and there is no improvement across the years. In another study, Savran (2004) examines the creation of PISA question in terms of their practicality, validity, reliability, linguistic aspects and content for the administration by comparing the kinds of questions given to Turkish students in some exams like Lycee Entrance Exams and suggest some advises like developing self-esteem from the beginning of preprimary education, need for the students-centered education, change of the concept of lesson text books. In a study carried out by Aşkar and Olkun (2005), the researchers compare computer usage in Turkish schools and OECD countries and conclude that computer usage positively related with the mathematics and science achievement of the students.

Applications and research studies show that performance data from such assessments are used to make high-stakes decisions regarding program development, evaluation, and curriculum. Since PISA presents a large amount of data, most of the studies base on the usage of these data that allow comparisons of educational input, process, and achievement in participating countries and lead to a different perspective for evaluating and improving a country's education. However, the complicated nature of this type of assessments makes them very sensitive to the methodologies used and the validity of such comparisons depends on these methodologies. Such a massive effect leads to questions of the suitability and validity of the PISA study for Turkish student group.

The validity of such decisions critically depends on the meaningfulness of scores from these assessments. Previous research has demonstrated that multilingual versions of assessments cannot be assumed to provide comparable scores (Allalouf, Hambleton, & Sireci, 1999).

The important point is to understand what PISA does and does not measure. What their results means for countries and to what degree the questions to be answered, such as: the independence of curriculum content and comparability of PISA outcome variables across countries, usage of performance scales and the validity concept from the creation of test items to the predictive power.

2.2. Overview on Validity Concept

In this part, particular emphasis will be given to review the changing view of validity by providing a brief historical context on the concept of validity in testing and on definitions of validity from traditional to contemporary, with its emphasis on construct validity.

Validity is one of the most central features in the field of measurement in the social and behavioral sciences. It has been discussed for many years but the concept changed dramatically in last decades. Sireci (2004) mentions about the dynamic nature of validity that it is a concept has that evolved and is still evolving. The question of validity has evolved from the question of whether one measures what one intends to measure, to the question of whether the empirical relations between test scores match theoretical relations in a nomological network, and finally, to the question of whether interpretations and actions based on test scores are justified—not only in the light of scientific evidence but with respect to social and ethical consequences of test use. These intellections are not opposite of each other but mainly differ in their focus and scope. Related literature will be presented around the answers these three main questions in a historical sequence.

As a concept, validity emerged at the turn of the twentieth century. In the beginning, it was a rather a theoretical and narrow concept that a test suggested either to be valid for anything which it correlates or to describe the representativeness of items chosen for a test (e.g. Bingham 1937; Guilford 1954). In a similar view, Cureton and Gulliksen (as it is cited by Goodwin and Leech, 2003) considered a test to be either valid

or not as evidenced by the correlations between the test and some other "external" criterion measure.

With the publication of the 1966 Standards (APA, 1966), evidence of validity has been gathered in three separate categories: content validity, criterion-related validity, and construct validity. Content validity is about the relevance and representativeness of contents included in a measurement instrument. Ideally, items chosen for a test should be a representative sample from the universe of all possible items referring to the domain of interest. The typical method for evaluating content validity has been expert judgment. Secondly, criterion-related validity has been defined as the association between test scores and some criterion or criteria of interest external to the test. The purpose of the test is often predictive and the method used for validation is often correlation or regression. Lastly, construct validity as a concept was initially introduced as an alternative to the other types of validity in cases where neither content validity nor criterion related validity could be applied and/or evaluated. Construct validity as originally conceived refers to the extent to which the contents of a measurement instrument are able to measure a theoretical construct.

In the following years, psychometricians and measurement experts started to point out the importance of the interpretations and decisions made from test scores and criticize separate meanings of validity. It is taken from the Standards for Educational and Psychological Testing (1985, p. 9) usually referred to as the APA Standards:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences from the test scores. Test validation is the process of accumulating evidence to support such inferences. (p.9)

The definition emphases validity both as a property of tests and a property of test score interpretations. A test and its scores seemed to be aimless and useless without evidences of validity.

Also gaining attention during the 1980s and 1990s was the need for evidence about the social consequences of test use. As a current and influential definition of validity, Samuel Messick (1995) is one of the most prominent modern validity theorists and his

model of construct validity as an all-inclusive concept has been very influential on the discourse about validity. For Messick, a *unified* construct validity framework was necessary not only from a scientific point of view but also for the applied use of test scores. The definition of validity changed by Messick (1989) as:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment. (p.13)

According to modern conceptions of validity, validity is about the appropriateness, meaningfulness, and usefulness of score based inferences (NCME, 1999). The *Standards* (1999) succinctly defined validity as;

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. (p. 9)

Simply put, validity is about what a test score means (Gregory, 2004) and validation is the process by which test scores take on meaning (Benson, 1998).

Although few would dispute this definition or the importance of considering validity as a unified concept (Messick, 1989), actual criteria for examining validity vary widely. During time, AERA (1999) studies discussed five sources of validity (based in part on Messick, 1989): evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing.

The main point about the discussions on validity is the role of consequences that studying both intended and unintended consequences of test use. The advantages and disadvantages of including investigations about consequences as part of validation broadly discussed at the national and international conferences well as at the variety of theoretical papers and research studies. (e.g. Linn, 1997; Mehrens, 1997; Popham, 1997; Shepard, 1997). Some measurement experts (e.g. Kane, 2001; Linn, 1997; Shepard, 1997) have argued for the broader conceptualization of validity (one that includes the consequences of using tests and other measures), whereas others (e.g. Popham, 1997) have advocated for a more limited and definition of validity that focuses primarily on the descriptive interpretation of scores.

In addition to these, the quality of the measurement instrument has never lost its value in the validation process. It does imply, however, that a sound measurement instrument is a necessary but not sufficient condition for the valid interpretation and/or use of test scores. Further, in the modern validity framework it is recognized that evaluations of validity are dependent on context, culture, scientific paradigm, prevailing values, and so forth. Validity is further seen as a matter of degree, validity evidence as always incomplete and validation as a continuing process (Benson, 1998).

2.3.1. Construct Validity

Within the ongoing change of validity concept, construct validity has a distinguishing place. Construct validity, the last type of the validity that entered literature by "trinity" view of validity (content, criterion-related, and construct) introduced by Cronbach and Meehl in 1955 that they claimed the most important step in validation is to define the construct and called this process as *construct formulation.* Based on the concept, construct validity was defined in the APA Standards as "the degree to which the individual possesses some hypothetical trait or quality [construct] presumed to be reflected in the test performance." (1966)

As a different ascertain, Schwab (1980) defines the construct validity to be ''representing the correspondence between a construct (conceptual definition of a variable) and the operational procedure to measure or manipulate that construct'' (1980, p.5). Also, from the point of construct validation, construct validity is an important part that requires a multi-step process for assessing the adequacy of measures.

Construct validity has been a dominant aspect of modern validity theory which is mainly based on the concept of construct validity. Messick (1995) builds construct validity by all other types of validities (e.g., content-related, criterion-related, face-validity related) traditionally used for validation. According to Messick, there can be no validity without construct-referenced measurement, as no score interpretation is possible without construct-referencing (Messick, 1988).

In addition to the differences in the description of construct validity, there is another disparity related to its nature that some authors (e.g. Angoff, 1988; Cronbach & Quirk, 1976) argue that construct validity cannot be expressed in a single coefficient; there is no mathematical index of construct validity. Rather the nature of construct validity is qualitative. In contrast, according to Hunter and Schmidt (1990), construct validity is a quantitative question rather than a qualitative distinction such as "valid" or "invalid"; it is a matter of degree. Construct validity can be measured by the correlation between the intended independent variable (construct) and the proxy independent variable (mark, notice) that is actually used.

While the importance of construct validity is self-evident, it takes on special importance in the context of international assessments. Since large-scale assessment measures are used in order to inform curriculum, program development and evaluation and decisions concerning educational policies, and to make comparisons of student achievement across countries, one of the major assumptions made in these assessments is that constructs being measured are the same for all participants.

In the assessment of construct validity, there are two distinct, but equally important, components. The first component is related with the test itself and the extent to which a test measures what it was designed to measure and the second component of validity is about the results and interpretations in the extent to which it is appropriate to use the results of a test for a specific purpose.

From the view of test itself, Dohn (2007) argues methodology of PISA's operationalisation of 'real life' situations in items and discuss PISA to assess 'knowledge and skills in assessment situations'. Moreover, the study criticizes the severely biased or ambiguous items to be in study that the validity of results and evaluation displayed by the PISA reports to be weakened. From a similar perspective, Wuttke (2007) concludes that significant differences does not mean reliable, valid, or relevant because the large numbers to cause statistical significance. From the item view, Wuttke gives an example to ask how the vaccination compares to alternative or complementary means of protection and the answer to be sought in the reading text but the preference on alternatives seems to be shaped by reliance on technology, and belief

in nature by different cultures. Wuttke claims reliability and validity of the study to remain limited because of uncertainty and bias.

As another example Prais (2003) points out that one explanation for the differences between PISA and the IEA studies is that the PISA questions were not designed to reflect curriculum content. As Prais (2003) notes, "the stated focus was ostensibly distinct from details of the school curriculum, and was intended to elucidate how pupils might cope in real life with the help of what they have learnt." It is not clear, however, that the resulting set of questions is any more or less 'real life' than the school curricula. Moreover, the selection of an arbitrary set of "international" questions biased the results against countries which pursued different curricular objectives.

Bodin (2005) focuses on the PISA external validity in connection with the construct validity, limited to its mathematical part, and that, from a French point of view (French, as related to the French mathematics curriculum, French customary assessment settings, etc.). PISA mathematical items seem to have epistemological and didactical validity issues. He further states that some precaution have to be taken and also justify the idea that some complementary studies should be undertaken.

Reviewing the results of the study, Psalidas et al. (2007) investigate the extent to which PISA science items validly assess the knowledge and skills of 15 year-old Greek students and examine the effect of factors: student's gender, scientific processes and contexts (situations) on the students' performance in these PISA items. The basic findings show that the paper-and-pencil test with the PISA Science items does not tend, unlike the interview, to effectively record the Greek students' Science knowledge and skills.

Beside these, in most of the studies (Ercikan, 2002; Oliver, 2005) conducted in relation to the construct validity of developed items, the focus is on the construct comparability that looks for the variance in factor structures being measured by different language versions of tests administered in different countries.

Sireci (1997) emphasizes the importance of ensuring that constructs being measured by the international tests to be equal for making fair and valid interpretations based on large-scale assessment data. In other words, the risk of making invalid interpretations by practitioners must be taken into consideration when factors outside of the construct measured are interpreted as real differences. Additionally, construct equivalence is said to be ascertained before scores from such assessments can be used meaningfully to inform decisions that require comparison of scores across these language groups or countries (Ercikan and McCreith, 2002; Ercikan, 2006). In their study, Kirsch, Long, Lafontaine and McQueen (2002) mentions about the extent of PISA reading items content and familiarity and the specificity of familiarity and content making comparability highly problematic.

The factors generating threats to construct equivalence are listed by Hambleton et.al (2005) and proposed to examine whether two measures have construct equivalence and are comparable. Effects of test translation and adaptation, measurement equivalence, sample representation and cultural and linguistic loadings of tests are some of the factors that may lead to assessment bias and may pose additional threats to test validity and comparability of scores. Study of Bonnet (2002) makes this point and discusses the difficulties with translation across diverse systems. A particular concern among some Francophone commentators on PISA (e.g Romainville, 2002) is the bias that may be induced by the Anglo-Saxon composition of the research, the technical advisors and the origins of the test materials. Additionally, American Psychological Association [APA], Standards for Educational and Psychological Testing (American Educational Research Association [AERA], and National Council on Measurement in Education [NCME] declares the critical value of fairness in testing and eliminating bias to make comparisons of individuals from different linguistic and cultural backgrounds in a fair and reliable way.

There is no single method for determining the construct validity of a test. Usually many different methods and approaches are combined to present an overall picture of the construct validity of a test. Besides the correlational approach described earlier, other frequently used methods are statistical or qualitative basing on the judgmental reviews. Evidence of construct validity involves making hypotheses and collecting information

over a period of time, using many sources and methods (Sartory and Pasini 2007). Methodologies for conducting construct comparability studies include using statistical procedures to examine the structural equivalence at the test level and differential item functioning (DIF)[2] analysis to examine comparability at the item level. The use of judgmental reviews to identify potential sources of DIF was also outlined. Some of the sources of DIF found were due to adaptation and curricular differences including differences in vocabulary, sentence structure, differential levels of information given by the test, content or language that is differentially familiar to one group of examinees versus another. If the construct being assessed is not consistent across all groups of interest, inferences based on the assessment results may be biased (Benson, 1987).

Golstein (2004) addresses the restricted nature of the data modelling and analysis, and the resulting interpretations. It points to certain features of the results that raise questions about the adequacy of the data and it stresses the failure to introduce a longitudinal component.

PISA assessment use the IRT, that difficulty or discriminative value of each item is independent from the context (OECD, 2008), to minimize the linguistics and cultural bias. As pointed out by Murat and Rocher (2004), such an aim is disputable that item rankings are depending on students' success is not same from one culture to another. In the other words, the level of difficulty is not independent of context and there is a cultural, linguistic and pedagogical impact.

As part of Rochex (2006) secondary analysis of PISA 2000 literacy test, analysis of three science questions shows that choice of expected answer can hide highly different methods for different students and for same student from one question to another and question format can overshadow levels of difficulty considered to be equivalent by PISA interpreters.

[2] DIF refers to "differences in item functioning after groups have been matched with respect to ability or attribute that the item purportedly measures… DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to attribute measured by the item and the test on which it appears" (Dorans & Holland, 1993,p37)

In her study, Olivery (2007) uses statistical and qualitative linguistic reviews to examine construct comparability and potential sources of incomparability. Exploratory factor analysis is used to evaluate the structural equivalence of the problem-solving measure for the English and French-speaking groups and translator reviews examine the sources in the different functioning items.

As it is seen from the literature, research is diverse that some of them focus on the differences in constructs which are assumed for students from different countries who take international assessments in different languages due to cultural, curricular, and linguistic differences and some others address the technical issues for the comparisons of constructs. Cultural diversity among the countries can affect intrinsic interest and familiarity of the content of items and differences coming from the curriculum issues can result in varying degrees of student exposure to the domain of items. In addition to these differences, the examinees in different countries respond to different language versions of the test items. Differences created by the adaptation process as well as linguistic differences that might affect examinee performance can affect the equivalence of constructs assessed in different countries.

The present study focuses on the construct validity in terms of the entities (something that exists as a particular and discrete unit) of the Turkish version of PISA science items. In order to address the issue, guidance comes from the field of test development and literature based on the construct validity and construct comparability of the translated tests.

Banerjee and Luoma (1997) emphasize that validity requires thorough understanding of the test from the early phases of test design to the conclusion and results.

Osterlind (1990) defines the *item* as:

> A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form of answering; and is intended to yield a response from an examinee from which performance in some psychological construct (such as knowledge, ability, predisposition, or trait) may be inferred. (p. 24)

Item development is a complex and sensitive process that a test is not better than the total of items. Downing and Haladyna (1997) state item writing as an art and mentions about items which are inaccurate to confuse to examinees or erring in anyway, so such kinds of items to threat the validity. Downing and Haladyna (2006) comprehend this framework about the multiple-choice, matching and alternate-choice (e.g. true-false) items with evidence on the validity issue. The framework includes four parts which are content, formatting concerns, style concerns and options including 31 items. Frey et al. (2007) broaden this guideline to 41 items and group validity concern to the five topics which are;

- covering important concepts and objectives,
- confusing wording or ambiguous requirements,
- guessing,
- rules addressing test-taking efficiency,
- rules designed to control for test wiseness.

As Nardi (2008) mentions learning is an inner pursuit which is not directly perceptible, but can only be assessed through a test which for the purposes of this paper can be described as a first 'intervention' (i.e. between the student and the test). This test is then analyzed by a marker (the second intervention) who in turn gives an assessment. The intervention between the student's skills and the test used to assess such skills implies the concept of construct validity. Among other conditions, the construct of a test is valid when it measures what we intend to assess. In other words, construct validity the degree to which the test allows us to assess the skills attained in relation to the objectives of the educational proposal.

Test bias is a major threat against construct validity, and therefore test bias analyses should be employed to examine the test items (Osterlind, 1983). The presence

of test bias definitely affects the measurement of the psychological construct. Additionally, Shepard (1987, p.179) made the point as clearly as possible "Bias is defined as invalidity" Two of the major threats to validity are construct under-representation and construct-irrelevant variance. Construct under-representation is present when the empirical domain is defined too narrowly, and thereby fails to adequately represent the theoretical domain of the construct (Benson, 1998). More simply put; the measurement captures only part of the construct one is interested in measuring. Construct-irrelevant variance is present when the empirical domain contains reliable variance that is unrelated to the construct of interest. That is, one unintentionally measures things that are unrelated to the construct of interest. Both these sources of error can distort test interpretation and use.

Birenbaum (2007) lists the ten sources of evidence for the construct irrelevant variance. These are;

- Culture and ethnicity
- Language
- Context
- Format and mode
- Test wiseness
- Test anxiety
- Perception of assessment
- Learning disabilities
- Gender
- Opportunity to learn

The issues of item bias and construct validity are interrelated in the number of skills tried to be measured and the degree to which comparisons between groups are appropriate are the issues of construct validity. If the test lacks construct validity, it contains items that are measuring skills other than those wanted to be measured and so the potential for the items bias occur.

However, obeying the item writing rules or the absence of test bias or construct irrelevant variance does not guarantee that the test possesses construct validity. In other words, the absence of these are a necessary, but isn't a sufficient condition. Construct validity includes all of them but means more.

As it is stated by Messick (1990), “tests are imperfect measures of constructs because they either leave out something that should be included…or else include something that should be left out, or both” (p.34)

In a similar way, Baykal (2008) emphasized the importance of construct validity among the other validity types and mentions the validity of a test to be directly related to the degree of construct it aims to measure. The definition is presented in a visual way in the Figure 2.6.

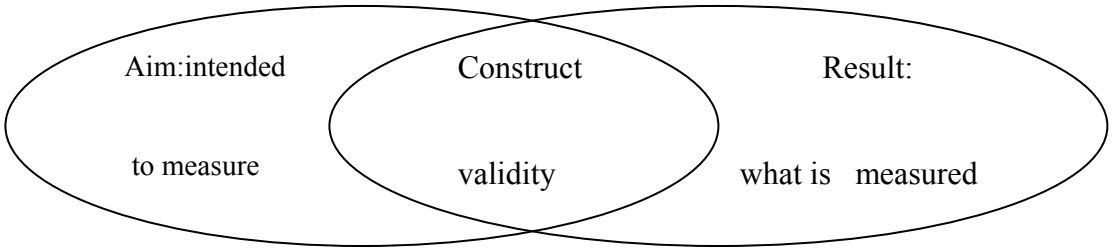


Figure 2.6. Construct validity concept in measurement

Beyond this implicit illustration, Baykal (2008) gives a two dimensional representation of the construct validity concept with example of national high-stake Student Selection Exam (OSS) in Figure 2.7.

	Could not be measured	Could be measured
Construct intended to be measured	Synthesis Creativity Scientific attitude Moral values	Knowledge Comprehension Analysis Evaluation
Construct intended not to be measured	Eye color Social habits Musical talent Weight	Chance Natural ability Reading ability Anxiety

Figure 2.7. Two dimensional representation of construct validity

There is similarity on the description of the validity at the items between Baykal (2008) and Messick (1990). However they differ about the unity of the validity concept that Messick (1990) emphasize the construct validity to cover all of the other validity types defined while Baykal (personal communication, February 19, 2008) reports construct validity to be related with the item although the other validity types connected to the whole test.

As it is mentioned before there is no single way for determining the construct validity of a test and numerous other strategies can be used to study the construct validity of a test. In most cases, construct validity tried to be demonstrated from a number of perspectives. Hence, the more strategies used to demonstrate the validity of a test, the more confidence test users have in the construct validity of that test, but only if the evidence provided by those strategies is convincing. Messick 1990) claims the construct validity of a test to be demonstrated by an accumulation of evidence. That can include using content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pretest-posttest intervention studies, factor analysis, multi-trait/multi-method studies, etc. Naturally, doing all of the above would be tremendous amount of work will take a long time and effort. Baykal

(2008) mentions construct validity to be mainly a quality and most of the entities qualitative in nature.

It is clear that efforts to create and use measures in a way that have adequate construct validity by minimizing bias are important in order to make valid decisions and comparisons across language groups or countries. Under the light of literature, the present study concentrates on investigation of construct validity aspect at the item level mainly as described by the Messick (1990) and Baykal (2008). Also design of the study composed of mixed qualitative quantitative method to achieve a comprehensive understanding on the usage of international assessments, in special PISA, from the point of what can be measured and not for the Turkish student population.

3. SIGNIFICANCE OF THE STUDY

International large scale studies have an undeniable popularity among the countries to see the status of their educational systems. National policy makers, curriculum designers etc. use the results to monitor their education systems in order to improve the quality of education. Whether international studies can help to see a complete picture of the skills of students have been a matter of debate. Some of the researchers found them to provide valuable data (e.g. Beaton *et al.*, 1999; Owen, 2001; Linn, 2002) while some other criticize them in several ways (e.g. Pollit and Ahmed, 2002; Egelund, 2008).

In last ten years, Turkey has participated several large-scale assessments as TIMSS, PIRLS and PISA. Data coming from these assessments are used to make education policy related decisions regarding curriculum development and curriculum evaluation. In most of the researches, the international studies are taken as only references to support the necessity of change in the educational systems (Savran, 2004; Şahin, 2007). However, it is obvious that the validity of these decisions depend on the relevance of scores taken from these assessments that this is depend on the degree of the overlap between what is aim and what is measured in various ways.

This study was conducted to investigate construct validity support at the item level for the Turkish version of PISA 2006 science items and questions its effect on the students' achievement change that can be guide at the formation of the further Turkish version of international assessments, especially PISA, and further studies on the different perspectives for international assessments.

4. STATEMENT OF THE PROBLEM

The existing studies (e.g. Ercikan, 2002; Simola, 2005;Golstein, 2004) in the literature did not only indicate the importance of gathering data from the countries to make comparisons and lead to educational policy decisions but also emphasized the necessity of creating valid tests to be confident on the fairness of the results. As Sireci (1997) states when factors outside of the construct measured are interpreted as real differences, there is a great risk of making invalid interpretations of assessment results.

The overall purpose of instruments used in PISA is to produce reliable and valid items to measure the competencies, e.g. scientific literacy. Development and evaluation of items are based on this purpose. As it is seen from the literature review, most of the studies describe different test structures stemming from methodological differences of the whole test that it leaves paucity for research at the item level (Sipps and DiCaudo, 1988). Due to the fact that PISA is an international study, it requires translation of the items to different languages. At that point the validity of the items in translated languages becomes crucial for understanding and interpreting the results of the PISA study not only in a global manner but also in a local manner. The literature review on the validity concept shows that assessment of construct validity is a complex process. Since assessment of construct validity is important in order to see the degree to which items measure unintended constructs and the degree which items do not measure intended construct (Messick, 1990; Baykal, 2008), this study focused on the construct validity of items of the PISA 2006, scientific measure for the Turkish population. The purpose of this study is to search for construct validity of single items and effects of negative construct validity entities on the students' achievement. This study aims to collect the item specific information by reviews of science teachers in Turkish sample of the items and search for the empirical evidence about the effect of these reviews on the achievement scores of the students. The problem statement of the present study is "What are the entities affecting the construct validity of the Turkish version of PISA 2006 science items and is there any significant effects of the negative entities on the achievement levels of the subjects?" It is hoped to make a contribution in the

clarification of validity of the Turkish version of the PISA items, specifically PISA 2006 science items, and gain a deeper understanding about extend of PISA study.

4.1. Research Questions and Hypotheses

There are three main research questions under the light of goals mentioned above examined in this study, these are as follows:

Research Question 1: What are the entities that impact on the construct validity of Turkish version of released PISA science items? The research question, in general, intended to explore teachers' judgments about the various components of PISA science items in relation to the construct validity. There are three sub questions:

i. What are the positive entities embedded within the items that affect measuring and evaluating of a person's science mindness?

ii. What are the negative entities embedded within the items that affect measuring and evaluating of a person's science mindness?

iii. Science items are how valid, how necessary and how important in terms of measuring and evaluating of a person's science mindness?

Research Question 2: Do the entities affecting PISA science items' construct validity negatively have an effect on the achievement scores of the students? The second question is intended to determine if there any significant differences between achievements scores of students on the original Turkish version items and revised ones. It is hypothesized:

i. There will be significantly higher scores on the PISA-Revised Turkish test (PISA-RT) than scores on PISA-Original Turkish test (PISA-OT) for the whole tests.

ii. There will be significantly higher scores on the PISA-Revised Turkish test (PISA-RT) than scores on PISA-Original Turkish test (PISA-OT) for each item.

Research Question 3: What are students' ideas about appropriateness of PISA stimuli with their learning experiences?

4.2. Variables and Operational Definitions

This study aims to investigate construct validity of items in terms of the positive and negative entities and their effects on the achievement of students. Construct validity is designated to be a qualitative property (e.g. Angoff, 1988; Cronbach and Quirk, 1976). It is defined to be a property of the test and defined to the extent to which the test or instrument is a measure of the particular psychological construct it was designed to measure. Because the phases in the study included qualitative and quantitative methods, the variables are defined as qualitative and quantitative. As Foussier (2006) states introduction of qualitative variables are a bit disturbing that a qualitative variable does not imply a numerical ordering and they are simply categories. The qualitative variable in the present study **'entity'** that is defined to be something that exists as a particular and discrete unit detected in the items.

For the quantitative parts of the study, the dependent and independent variables were defined separately as below.

4.2.1. Dependent variables

At the first phase of the study, the research question is described in three sub questions. The dependent variables in relation to the third sub question are **validity, importance** and **necessity**. The description of these variables was left to the common sense and intuition of the teachers who rated on them. Two examples from the teachers interpretations for these attributes were given in the below quotations.

> I used importance and necessity with their word meanings when I rated the questions in terms of improving science mindness of a person, especially my students. Validity is a school concept that I learnt in the university first and I remembered that it means to be able to measure truly what you want to measure. (T34)

> As I understood importance means 'is this topic or question critical in the learning cycle, is it a cornerstone for school and life' and for the necessity 'is this topic really required to improve a student science literacy, science mindness in you writings'. The validity is the hardest one to answer and rate, because it needs to understand what you are trying to measure first and usually I looked for is there any point which can effect your question like clarity, curriculum relation etc. (T75)

At the second phase of the study, the dependent variable is **students' science achievement scores.**

In two student groups **science achievement scores** were measured with two separate instruments:

i. *PISA Original Turkish Test (PISA-OT)*, was used to measure science achievement scores of the students by using released items as presented in the Turkish version of PISA 2006 study (EARGED, 2008).

ii. *PISA Revised Turkish Test (PISA-RT)*, was used to measure science achievement scores of the students by using revised items as examined at end of the first phase of the present study.

At the third phase of the study, the dependent variable is **appropriateness of the stimuli with students learning experiences** that appropriateness is defined as in its

word meaning of suitability, typicality and learning experiences are defined language usage, school knowledge, daily life knowledge and lay-out.

4.2.2. Independent variables

At the second phase of the study, the independent variable is the **negative entity.** The negative entities were explored and categorized at the first phase of the study for each stimuli and every item in science units. Negative entity will be defined as sub category (ies) which they are formed by the thematic units droven from the reviews of teachers as existing negative properties which they are particular and discrete unit(s) in the science units. Besides, since some of negative entities were used in the revision process of the items as described at the part eight of the present study, at the second phase of the study the **negative entity** will be used for one or more negative entity (ies) used in the revision process of stimuli or items.

5. METHODOLOGY

Methodology and research design direct the researcher in planning and implementing the study in a way that is most likely to achieve the intended goal. It is a blueprint for conducting the study (Burns & Grove, 1998). This section introduces the methods used in the study, population, and sample by examining the selection of sample for the purpose of the study. Also, the instruments used as data collection tools are introduced, the development of instruments and data collection are explained. Lastly, analysis of data is described.

5.1. Sample

This study was conducted in three phases. At the first phase of the study, sample included two groups (item and teacher). One of groups, item group, consisted of 25 Turkish version of released PISA 2006 science items in eight science units. The second group included 80 secondary school science teachers as judgmental experts. At the second phase of the study, the sample included two groups (item and student). Item group included original and revised version of the 25 science items. Student group included 60 students who are 15 years old. At the third phase of the study, sample consists of 30 students who are 15 years old. The summary of the sample selection process for each three phases of the study is given at Figure 5.1. Then, detailed demographic information of the groups in the sample for three phase of the study is presented.

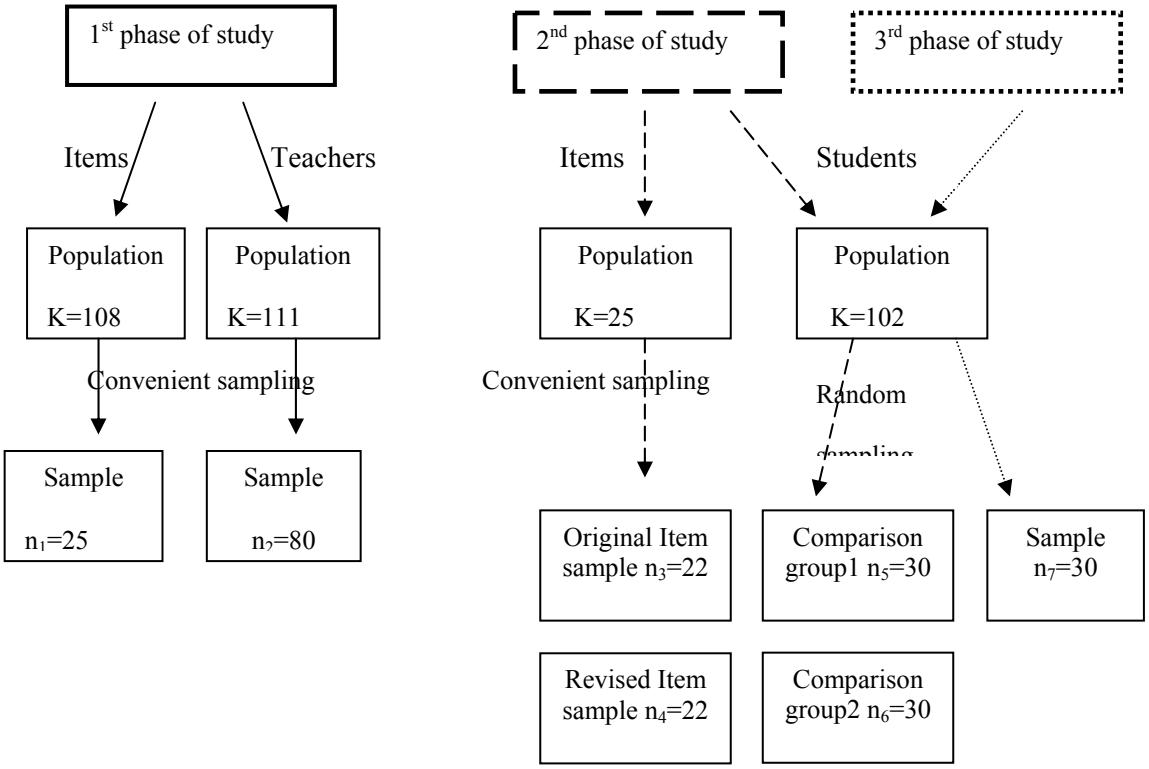


Figure 5.1. Process of sample selection

As it is seen from the Figure 5.1, at the first phase of the study, there are item and teachers groups that samples selected through convenient sampling. Since the revision process of the science units completed at the end of the first phase, the second phase of the study included original and revised item samples, and also selection process of two comparison groups. The reason for the reduction of the item number for the second phase is based on the absence of answer keys for three of the items. Hence, these three items excluded from the tests used in the second phase of the study. Lastly, the sample in the third phase of the study was composed of 30 students.

At the first phase of the study, the item sample consists of eight science units containing eight science stimuli texts and 25 science items distributed publicly for PISA 2006. The teacher sample consists of 80 secondary school science teachers.

Firstly, the properties of the science units will be given. The science items were chosen on purposeful criterion basis (Panton, 1999). The reason to select purposeful sampling was because this sampling type is as a dominant strategy in qualitative research and purposeful sampling roots information rich situations which can be studied in depth (Patton, 1990). The PISA test items are strictly confidential since they are candidates for reuse in future assessment cycles and most of the question kept for the next assessment to prevent time and money consuming procedures. Despite this fact, the PISA authorities have released some examples of test items. Therefore, the sample of the present study involves the stimulus texts and 25 publicized science items in eight science units. These were identified thorough investigation in PISA 2006 publications and matched the publicized items in Turkish (see Appendix B). Since these items are released so as to be read by all the interested parties as exemplars, it could reasonably be argued that they are representative and reflective of the entire PISA. As it is mentioned in part 2.2.1, in this study only PISA 2006 questions are used because the definition of 'scientific literacy' differs for the PISA 2000, PISA 2003 and PISA 2006 cycles. In other words, the differences in the constructs intended to be measured and the changing number of the questions in the three PISA studies leads this study to use only PISA 2006 science units which includes a greater number of the released questions. Sample included a total of 25 items were placed within the eight PISA science units. The titles of the science units, the number of items included, the scientific process that examines each item and the context in which each item is incorporated are presented briefly in the following Table 6. Abbreviations for the science-units presented in the parenthesis next to the titles of the science units. At the remaining parts of the study the abbreviations of the science units will be used rather than the whole names of the science units.

Table 5.1. Descriptions of PISA 2006 released the science units

Title of Science unit/Abbreviation	Items	Competency	Item format*
GREENHOUSE (GREEN)	1st	Using scientific evidence	OC
	2nd	Using scientific evidence	OC
	3rd	Explaining phenomena scientifically	OC
CLOTHES (CLOTHES)	1st	Identifying scientific issues	CMC
	2nd	Explaining phenomena scientifically	MC
GRAND CANYON (GRAND)	1st	not defined	**(OC)
	2nd	Identifying scientific issues	CMC
	3rd	Explaining phenomena scientifically	MC
	4rd	Explaining phenomena scientifically	MC
SUNSCREENS (SUN)	1st	Identifying scientific issues	MC
	2nd	Identifying scientific issues	MC
	3rd	Identifying scientific issues	MC
	4rd	Using scientific evidence	OC
MARY MONTAGU (MARY)	1st	Explaining phenomena scientifically	MC
	2nd	Explaining phenomena scientifically	MC
	3rd	Explaining phenomena scientifically	OC
ACID RAIN(ACID)	1st	Explaining phenomena scientifically	OC
	2nd	Using scientific evidence	MC
	3rd	Identifying scientific issues	OC
PHYSICAL EXERCISE (PHYSICAL)	1st	Explaining phenomena scientifically	CMC
	2nd	Explaining phenomena scientifically	CMC
	3rd	Explaining phenomena scientifically	OC
GENETICALLY MODIFIED CORPS (GMC)	1st	not defined	**(CMC)
	2nd	Identifying scientific issues	MC
	3rd	not defined	**(OC)

* The abbreviations used for item format OP=open constructed, CMC=complex multiple choice MC=multiple choice

** the detailed explanations were not given by OECD (2007) because these items were below the Level 1 or they were dummy items that they were not used in the measurement of countries means scores. However, the information in parenthesis were taken from the source of released Turkish version (EARGED,2007)

As it is seen from the Table 5.1, the sample of the items used in the first phase of the study varied in format, nine of the questions are 'open ended' that ask to construct responses, five of the questions are in 'yes-no' format which are called as complex multiple choice and they ask to partial of full credit and the remaining nine question in 'multiple choice' format that requires the selection of true alternative among four (see Appendix B).

The teacher sample of the first phase of the study includes 80 secondary school science teachers. Demographic information about the teachers will be presented below. Teachers participated first phase of the study were examined by gender, age, experience and school type they work, the results presented on Table 5.2.

Table 5.2. Teachers' Demographic Information

	Frequency	Percent	Valid Percent	Cumulative Percent
Gender				
Male	38.0	47.5	47.5	47.5
Female	42.0	52.5	52.5	100.0
Total	80.0	100.0	100.0	
School type				
State	46.0	57.5	57.5	57.5
Private	34.0	42.5	42.5	100.0
Total	80.0	100.0	100.0	

As it is seen form the Table 5.2, over half 52 per cent (n=42) of the teachers were female. The remaining 48 per cent (n=38) were male. The majority 60 per cent (n=48) were between the ages of 23-30. Approximately, 18 per cent (n=14) were between ages 31 and 33. Ten of them, (12.5 per cent) were between the ages of 34-36. Few 7.5 per cent (n=6) were between the ages of 34-40. The remaining 2.5 per cent (n=2) were above the age of forty. Most 57.5 per cent (n=46) of the teachers were working in a state school while 42.5 per cent in a private school.

At the second phase, there are item and student groups. The study conducted in a private 'dershane', which is a preparatory center for high school entrance exam which is called Level Determination Exam (SBS) and university entrance exam which is called

Student Selection Exam (ÖSS). The 'dershane' is located in Besiktaş region. All ninth grade students (K=102) in this center are considered as the target population and 95% of the students live in this region. 60 students were selected for this study due to some practical reasons such as time and place restrictions. Researcher selected 60 students randomly and 30 of them assigned to one of the comparison groups and remaining 30 students included in the other comparison group. The demographics of the students participating in the second phase of the study presented at the Table 5.3. It is seen that more than 50 % of the students were male and most of the students (n=34) attended to private schools. In the comparison group, the distribution of gender was similar with the values that there were 18 males in one comparison group while there were 17 male students in the other. However, there were more students (n=28) attending private schools in the comparison group which filled the PISA- RT test than the students (n=6) in comparison groups which answered PISA-OT test. There was random assignment of the students to the comparison groups. However, the 'dersane' located in Besiktaş region which is one the wealthy region in Istanbul, most of the students attend to the private schools.

Table 5.3. Students' Demographic Information for Second Phase of Study

	Frequency	Percent	Valid Percent	Cumulative Percent
Gender				
Male	37	61.7	61.7	61.7
Female	23	38.3	38.3	100
Total	60	100.0	100.0	
School type				
State	26	43.3	43.3	43.3
Private	34	56.7	56.7	100
Total	60	100.0	100.0	

In order to check whether there is a significant difference between the science achievements of these groups, two criteria were defined. Student's score gained on the 'Genel tarama testi' (GTT), which is a test covering subjects until the test date at the 'dersane' administered to whole population, it was the first criterion. The science subject score from the same test was the second criterion. Students' school science grades were collected as the third criterion but they were not used because of the unstandardized scores resulting from the different school types on which students attending. To show

that there is not significant difference between these two groups in terms of their science achievement, independent sample t-tests were conducted. The first one is carried out between GTT scores of the two groups, and it is found that there is no significant difference between these two groups' GST scores (t= 1.044, p= 0.391). Secondly, independent sample t-test was also conducted between science subject scores of these two groups of students. Also, no significant difference is found for these scores (t= 1.883, p= 0.065). All of the 60 students took part in the second phase of the study; it means there were no students missed the application of second phase of the study. As a result, sample of the second part of the study is composed of 60 students.

At the second phase of the study, the item sample composed of items in two versions of tests. PISA-OT test composed of eight science units which included eight stimuli and 22 items examined at the first phase of the study. There were 25 items at the first phase but three of the items excluded from the science units at the second phase. There were not marking keys for two of the items and one is assigned to be dummy item that not used in the original PISA 2006 study. PISA-RT test included revised version of 22 items in PISA-OT test.

An additional third phase emerged at the middle of the study which is carried out with the 30 students selected among the students not assigned in the comparison groups. At the Table 5.4 demographics of the students participated in the third phase will be given. Demographic data collected for students answered the SOS in the third phase of the study included gender, and school type they currently enrolled.

Table 5.4. Students' Demographic Information for Third Phase of Study

	Frequency	Percent	Valid Percent	Cumulative Percent
Gender				
Male	20	66.7	66.7	66.7
Female	10	33.3	33.3	100.0
Total	30	100.0	100.0	
School type				
State	16	53.3	53.3	53.3
Private	14	46.7	46.7	100.0
Total	30	100.0	100.0	

Of the students, about 53 per cent (n=16) were from state schools compared with the 47 per cent (n=14) of the private school students. With regard to gender, about 67 per cent (n=20) of students were male while 33 per cent (n=10) were female. Students answering the questionnaire were from 10 different schools. The schools were located in Beşiktaş region.

In summary, for the first phase of the study eight Turkish version of PISA 2006 science units were used. Also 80 secondary school science teachers participated examination of these science units. For the second phase of the study the 60 students who were selected by the researcher composed of the student sample of the study. Also items in PISA-OT and PISA-RT tests consist of item sample of the second phase. There is no significant difference between the science achievement scores of the students and students assigned to the two comparison groups. For the third phase, there were 30 students selected through random sampling from the remaining students at the second phase of the study.

5.2. Design and Procedure

The overall study is a composition of two related studies that can be characterized as a work of mixed method research, in which qualitative and quantitative data collection and analysis used in the same study (Teddlie and Tashakkori, 2003). In the

mixed methodology the most appropriate and purposeful methods for data collection are used to answer the research questions to meet the demands of the context of study.

For the first phase of the study, the sample is formed by eight text stimuli and 25 science items released after PISA 2006 study that they are examined by 80 secondary school science teachers with 10 teachers per science-unit. The Item Rating Form (IRF) was prepared for teachers to comment on positive and negative features of the tests and items together with the ratings on the validity, necessity and importance values. Based on the reviews of teachers on the text stimuli and items, positive and negative categories embedded in the texts and items are formed at the first phase of the study, then, these items were revised based on the negative categories. For the second phase, the target population of the study was 102 students who are nine graders of a private exam preparation center in Istanbul. Then, the researcher selected 60 students randomly and students were divided into two comparison groups. One of the comparison groups received the publicized Turkish version of items used in PISA 2006 and other comparison group received the revised version of the items which are prepared after the first phase of the study. Afterwards, the researcher checked for the difference between science achievements of two comparison groups in order to evaluate students' science level. Besides, after the revision of the items under the light of teachers' comments, it became apparent that some entities could not be revised. Because of this reason, an additional questionnaire was prepared and given to the 30 students that these students are not included in the comparison groups. These 30 students took Student Opinion Survey (SOS) prepared based on the entities that these mainly could not be included in the revision items. The accompanying figure indicates the summary of aim of the each phase, design and method of analysis carried out.

Table 5.5. Design of the study and method of analysis

Aim of each phase of the Study	**Phase of the study**	**Method of analysis**
1. to discover positive and negative entities embedded in the science units via written comments of teachers on IRF	First phase; Review of PISA text stimuli and items with ratings and written explanations of the teachers.	Content analysis: reading and classification of positive and negative entities coming from data and calculating means of ratings as indices.
2. to find out how revised items affect Turkish 15-year-old students achievement with special focus on individual items	Second phase; Presenting original Turkish (PISA-OT) version and revised version(PISA-RT) tests to a selected group of 15-years old students	A t-test was used to compare the means to see whether the differences where statistically significant
3. to study the extent to which the relation of the unrevised entities on the science-unit in the second phase of the study with students' familiarity	Additional part; Studying the views of the students on the stimuli of the science-units	Means of the students' ratings on the SFQ

As it is seen from the Table 5.5, for the first phase of the study the quantitative data comes from the Likert-type scale questions on the IRF and qualitative data was collected in the form of open-ended questions on the survey. It is deductive in nature, as specific reactions to the elements of PISA science items are explored. To detect the entities affecting construct validity of items, judgmental analyses by teachers are used. As it is mentioned by Airasian and Jones (1993), classroom teachers are the ultimate purveyors of applied measurement, and they rely on measurement and assessment-based processes to help them make decisions every hour of every school day. Moreover, teachers spend at least one third of their professional time on assessment activities that inform a wide variety of decisions made daily and directly influence students' learning experiences (Stiggins and Conklin, 1992). Teachers have various properties such as;

- reviewing results of standardized tests,
- creating tests of their own using various formats,

- evaluating completed student projects they developed or obtained from resource guides or textbooks,
- assigning work to be done outside of school,
- asking questions, listening, watching, interviewing students,
- posing questions for solution by individuals or groups of students.
- communicating their findings for evaluation of students,

All of these have a crucial impact on the learning process that assessments affect students by communicating learning goals, including the subject matter content and thinking processes valued by their teachers. As a result, teacher judgments used for the analysis and ten teachers for each one of the PISA items asked to filled the IRF.

Second phase of the study includes mainly quantitative data based on the achievement scores of students in comparison groups that each of the groups answered 22 items in the original and revised versions of the items. The data collected are naturalistic as there was no treatment given and there was no interference with the participants' natural behaviors. The scoring process is carried out by researcher and one other expert on the scaling processes as described in original PISA study (see Appendix D).

As an additional part of the study, the third phase included ratings of the students on a Likert-type questionnaire including four items to collect data about the familiarity of the students with stimuli of science-units in terms of content, language and lay-out.

As a summary, the main processes of the study to reach the aims of the study can be seen below,

- Theoretical review
- Development of data collection tools
- Application of IRF as data collection tool
- Analysis of data from first phase of the study
  - Qualitative data
  - Quantitative data

- Revision of the items
- Application of the revised and original Turkish version of items
- Analysis of data from second phase of the study
  - Quantitative data
- Analysis of data from third part of the study
  - Quantitative data
- Results from the analysis of the data
- Conclusion and discussion based on the results

5.3. Instruments

There are four instruments used in the study. First of them is IRF which was designed to collect information based on the judgmental reviews of teachers about the entities of construct validity in items at the first phase of the study. Second instrument is SOS which was developed to collect information about the familiarity of students with the topic, language and layout of the stimuli texts of PISA science units. Third instrument was the PISA Original Turkish test that is composed of the eight science units used in the PISA 2006 study (see Appendix B). The fourth instrument was the PISA Revised Turkish test that it is prepared through the revision of the items at the end of the first phase of the study by recovering negative entities (see Appendix F).

5.3.1. Item Rating Form (IRF)

In order to collect data about the entities in items, teachers were given Item Rating Form (IRF) which is developed by researcher and one measurement and evaluation expert. The instrument is composed of three parts. In part A, teachers are asked for the demographic information of age, gender, experience and type of the school in which they are working. In part B, there are two open-ended questions asking about the

negative and positive entities can be found in PISA science-units. Part C contains three likert - scale response items which were scored on a 9-point scale ranging from 1 to 9.

Part B includes an explanation for the judgments to lead them evaluate the PISA science-units by pointing positive and negative entities in items with their own words. These explanations based on the description of *science mindness* construct and the ways teachers intended to take care by making their reviews. Describing the construct of science mindness was seen as a requirement by the researcher since it was not possible to give science literacy description of the PISA 2006 to each of the teachers because of some restrictions such as time, place, and detailed explanations given by PISA. Part C of the instrument asks for the revision of items in terms of validity, appropriateness and importance for evaluating the *science mindness* of the students. *Science mindness* is described beyond the scientific literacy definition of PISA. It is described to include the properties of 'scientific disposition' and 'science proneness' in addition to the scientific literacy description of the PISA. In order to clarify definition of science mindness, it will be taken a closer look into the nature of 'scientific disposition' and 'science proneness'. Visser (2007) mentions about developing scientific disposition as;

- Functions effectively in unpredictable situations
- Recognizes question-asking as central element of scientific pursuit.
- May apply "the scientific method"
- Sees understanding of science fundamentals as situated.
- Represents "a high level of aesthetic and moral conscience".
- Equally pertinent for continual human development in all areas of the world.
- Comprised of attitudes, beliefs, cognitive and metacognitive strategies

Brandwein (2007) describes *science proneness* to be related with creativity, interest in science, curiosity about what things make work, strong imagination in things scientific, unwillingness to accept the explanations without proof and self understanding in science related situations for gifted children. In the present study, components of science proneness are not thought only in relation to gifted children but the all pupils.

In the IRF, the teachers were asked to evaluate the content, competency, questioning style, verbal expression, visual elements etc. of stimuli and items included in science-units in terms of improving *science mindness* of one person.

IRF was distributed to five judges (one measurement and evaluation expert, one physics teacher, two chemistry teachers and one biology teacher). With the help of the feedback that came from these judges, the instrument was given the original form (see Appendix C).

5.3.1.1. Validity and Reliability Analysis of the Instrument. The validity analysis of the instrument was done qualitatively. One measurement and evaluation expert, one graduate student and four teachers examined the test for the content validity. Because the test composed of two open ended items and three Likert type items, the interrater reliability analysis could not calculated to see the consistency of the prepared form within the aim of the first part of the study. However, the aim for the preparation of IRF explained to all experts and asked to rate the appropriateness of the form to collect data by using a 9 point scale. The mean of the expert ratings were calculated to be 8.4.

5.3.2. Student Opinion Survey

The instrument designed to collect data for the third phase of the study about the familiarity levels of the students with the stimuli of science units in terms of their content, language, and lay-out. The duration for administering this instrument to the students was 45 minutes, one lesson hour.

It is a paper pencil test which contains four Likert-scale response items (see Appendix G). Items were scaled on a five-point scale ranging 1 (very unfamiliar) to 5 (very familiar). Reason to choose five-point scale rather than a nine-point is because of possibility of students' unfamiliarity with a nine-point scale. Reliability analysis of the scale was conducted in the same place with the present study. Sample for the reliability study was 30 fifteen-year-old students and Cronbach's alpha coefficient for this sample

was calculated for the instrument (4 items) from the test resulting in an overall reliability of .78.

Additionally, reliability of ratings was calculated in terms of the interclass correlation coefficient (ICC). For this purpose the aim for the preparation of SOS explained to six experts and asked to rate the appropriateness of the each of the four items prepared to collect data by using a 9 point scale. As it is seen from the Table 5.4, the average measure interclass correlation is found to be 0.84 and values above the 0.70 are considered acceptable (Vincent, 1999). As result, based on these correlation coefficients it can be said that the six judges are very consistent with their ratings about the appropriateness of four items in the SOS.

Table 5.6. Intraclass Correlation Coefficient for SOS

	Intraclass Correlation
Single Measures	0.480
Average Measures	0.847

5.4. Procedure

At the beginning of the study, the IRF was developed by researcher and a measurement and evaluation expert. Then, eight science units and IRF forms were distributed to the 90 secondary school teachers via e-mail and 21 of science units together with IRF forms were given 21 teachers in the form of hard copies. Teachers were asked to complete the IRF forms for each stimulus and items included in one the science-units in three weeks. However, the response rate at the end of the three weeks below the 50%, because of low rate the duration was extended to two more weeks. At the end of the five weeks, there were 80 completed forms turned back. The return ratio of the distributed units was calculated to be 71.4 %. Based on the results of the comments of teachers, the original Turkish versions of the items were revised in some of the categories together with one measurement and evaluation and one language expert.

For the second phase of the study, students participated in 30-minute sessions to complete the tests, PISA-OT and PISA-RT, the time is decided by making an analogy with the given time needed for completing the original whole PISA science test including 108 science items. The PISA-OT and PISA-RT tests included 22 items. It is noteworthy that there were 25 items at the first phase of the study, but three of the items were omitted from the tests used in the second phase. Two of these omitted items were assigned to be below the cut point of the Level 1 (OECD, 2007, pp.38-69) and the answer keys were not present for the items and the third one of the items was assigned to be *dummy* item by PISA (OECD 2007, p. 83). Therefore, three of the items were not included in the PISA-OT and PISA-RT tests given to comparison groups.

As it is mentioned in part 5.2 and 5.4, there is an additional part for the present study which it is called as third phase, an additional questionnaire developed by the researcher based on the feedbacks of the teachers at the first part of the study and the categories could not be included in the revision process of the items. For this additional questionnaire, the students were asked to rate the familiarity on the PISA science units' text stimuli in the school and daily life, and also familiarity with the Turkish language usage on the texts and layout of the texts. In analyzing the data, familiarity induce was

calculated for each text and as a whole. This need mainly was because of unsuitability of some categories in revision process of the science units. The SOS was prepared to search the students' views on the stimuli of science units in terms of curriculum, content and language familiarity. SOS included four items asking familiarity of students with the content in school and in daily life, and also language and lay-out (see Appendix G). The ratings were made on a Likert scale ranging from 1 (not familiar) to 5 (very familiar).

30 students filled the questionnaire for each of the science units, because of practical reasons to minimize the effort involved in rating the large number of pages and to keep motivation of students to fill the questionnaire and also with the low means of the stimuli found in the IRF compared with the remaining items, only stimuli (it means eight stimuli) in the item groups presented students. Students completed the questionnaire in a given time of one lesson (45 minutes). The data analysis was made on calculation of mean scores of the questionnaire items for each PISA science unit will be given.

At the beginning of the lesson, students were motivated by describing general aim of the questionnaire and explaining the importance of their input for the study. Students then instructed to assess the familiarity of PISA item groups' stimuli. The familiarity of the PISA item group stimuli were to be assessed simply by judging how familiar the texts in terms of their content learned from school and daily life, language usage and lay-out. The students seemed to be well motivated that they filled the boxes with a concentration.

6. DATA ANALYSIS AND RESULTS

As it is described by Burns and Grove (1993), data analysis is a type of mechanism to reduce and organize obtained data to generate findings that necessitate interpretation by the researcher (1998, p.744). Miles and Huberman (1994) refer three steps for the qualitative analysis that data reduction is the first step. It includes selection, condention and transformation of the data. The second step is data display that helps to the creation of organized data by thematic units. The last one is the verification step that it requires revisiting data to confirm the identified themes (categories). In the present study, the data for the first phase of the study was analyzed according to written comments of teachers for open ended questions of the IRF and numerical scores obtained from the three Likert-type questions of IRF. In general, the three steps defined by Miles and Huberman (1994) were followed. The data were analyzed using strategies similar to content analysis. The fundamental approach in content analysis is to reduce texts to categories describing phenomena on a general level that this is a method for producing inferences from messages that are present in the content itself. Content analysis is typically used in analyzing various kinds of textual data systematically (Ryan and Bernard 2000), especially free-flowing textual data such as free comments in response to open ended questions as in the case of this study. According to Yıldırım and Şimşek (2005), content analysis allows the analysis of the collected data in depth and enables the appearance of undefined or unclear themes or categories.

The unit of analysis in this study can be described to be thematic i.e. piece of the text that reflects a single theme /category. Boyatzis (1998) defined a theme as “a pattern in the information that at minimum describes and organizes the possible observations and at maximum interprets aspects of the phenomenon” (1998, p. 161). Thematic analysis is a search for themes (categories) that emerge as being important to the description of the phenomenon (Daly, Kellehear and Gliskman, 1997).

The process of the content analysis and names used in it is represented below figure to clarify the terminology used in the study.

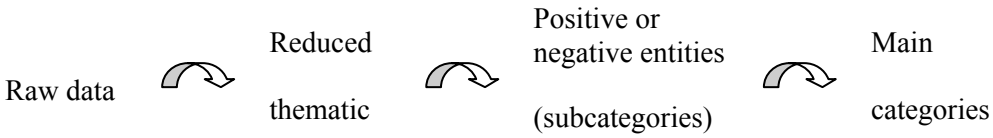


Figure 5.6. Process of content analysis

As it is seen from the Figure 5.6 the content analysis in the present study involved three stages. At the beginning there was raw data that thematic units were formed by reduction of the teachers' comments. Then, the thematic units were combined to form the sub categories that these were classified as positive and negative entities according to the teachers' writings. In the present study, it is preferred to use 'category' rather than 'theme' to make the reading easier. During the study 'sub category' will be used interchangeably with positive and negative entities. At the end of the content analysis, main categories were formed by the classification of the related sub categories under more general categories. All steps included in the content analysis will be explained in detail at following pages.

In practice, the thematic units are the codes that are driven form the free writings of teachers by reducing long sentence to meaningful pieces. The data in the first phase of the study in some cases consisted of only one sentence, sometimes of several sentences. Furthermore, one written comment might include one or more thematic units, as in the examples in following parenthesis, which shows three codes or thematic units (…*sentences are so long*…, …words *are uncommon*…, …*content is unfamiliar to students*…). In addition to this, same sentence may include more than one thematic unit. Each thematic unit was categorized according to entity it comments on, i.e. the long sentence, uncommon word and irrelevance with national curriculum. In general, an inductive approach (Straus and Gorbin, 1990; Pope, Ziebland and May, 2000) was used in the analysis that categories were induced from the data although researcher the researcher's prior knowledge and experience had an effect on the categories, the researcher paid attention to in analysis.

Data from the first phase study included written comments produced by 80 teachers, 10 teachers on each science units. The positive and negative entities formed during the content analysis of teachers comments at the sub category step. Teacher comments varied in length the shortest being an incomplete sentence, the longest one paragraphs. A typical item of short data is as follows: "*the content is irrelevant with curriculum"* (GREEN stimulus). An example of the longest data is like: "*the sentences of the paragraphs are so long, some words are not common for the students and also students are not met with the subject of the reading frequently in school that students spent much time to read and understand the reading passage so that it can be beneficial to shorten the sentences to make students' understanding easier to answer the questions*" (CLOTHES stimulus). As it is illustrated above examples, one comment made by teachers may deal with one or more thematic units (codes). These thematic units can be classified under same or different main categories. The first example includes one thematic unit (irrelevance with national curriculum) classified under one category (content) and the second includes three thematic units (long sentence; uncommon vocabulary; different objective form program) classified under two different categories (language and content). A more detailed example is presented in Appendix D.

Naming categories that resulted in the analysis is shaped by nonexistent labels that researcher labeled them without using existing ones in the literature. It may be said that the content analysis carried out in this study was an inductive one. Additionally, as Ryan and Bernard (2000) discuss because the coding (presenting thematic units) with the sub categories and main categories was made; categories can be quantified and thus also analyzed quantitatively.

Content analysis of the data was conducted by the researcher herself following careful line-by-line reading and one doctorate student with a master degree on secondary school science and mathematics education. As Rice and Ezzy (1999) mentioned the process of the analysis involves the establishment of themes (categories) through careful reading and re-reading of the data. Hence, the data was read through several times. During the first two readings, the data became familiar in content to the researcher. During the third round, the data was reduced to expressions representing subcategories induced from the data during which the categories present in the data began to take

shape. The subcategories for the positive views of the teachers formed the positive entities and named as 'positive entities' throughout the study and the subcategories emerged for the negative views of the teachers called to be 'negative entities'. The focus of the study was not on the degree of the comments made on the entities or categories but on themselves, i.e. if two comments both referred to the subcategory of worse item stem, they both categorized under the "worse item stem" subcategory while one may include *bad* comment and other *worst*. During the fourth reading, similarities and differences between the subcategories became clearer and the main categories given last shape. Then, one more expert was given the both comments, subcategories and the categories to be competent about the validity of the analysis. During the second analysis some refinements to the categories was made. Kuzel and Like (1991) offer four techniques to be used during data collection process. Member check is one of these techniques. In the present study, member checking was used to confirm the subcategories formed by the content analysis process. Five teachers reviewing five different science units were interviewed for 15-20 minutes and asked whether the subcategories based on their writings were consistent what they actually meant and what the researcher categorized. Teacher reviewers participated member check agreed with the thematic interpretations described in the study and stated that the quotes were representative.

For the instrument IRF, scores obtained from ratings of teachers about the validity, necessity and importance of the stimuli and items in the science units were analyzed by calculating mean scores of each stimulus and item separately and as the mean of science unit. It is outstanding to notify that there were 80 teachers filled the IRF and each of the eight science units were reviewed by 10 teachers.

In the course of analyzing results of the first phase, the need to study the extent for the negative entities could not been included in the revision process became apparent. It is found necessary to study the familiarity of students with the PISA science units in terms of content, language, curriculum relationship from the point of students.

Second phase of the study included application of original Turkish items and revised items. The test was completed by the students and their responses were

subsequently codified and marked, according to the Turkish version of PISA marking guides. One independent scorer other than the researcher codified and marked the students' responses. In case the marks of the two scorers varied, the final mark was determined with the discussion between two experts. Actually, there were very few differences between the marks of the two scorers. The comparison between the average values of scores coming from the independent groups was made using the independent sample t-test. In addition, the independent t-tests were carried out for each item to see whether a significant difference between the scores of comparison groups. The level of statistical significance selected for all the comparisons was the usual 0.05 (5 per cent). All the tests were conducted using the Statistical Package for Social Sciences (SPSS) software, Version 17.0

This part of the study is organized to present the results to the research questions of the first, second and third phases of the study respectively.

Because the qualitative design was dominant at the first phase of the study, the central question was written as the statement of the question being examined in its most general form. The central question was followed by three sub questions in order to narrow the focus of the study but leave open the questioning for the first two sub questions. The third sub question was a descriptive question related with the first two sub questions. Therefore, the questions become 'working guidelines' rather than the 'truths' to be proven (Thomas, 1993 p.35).

In order to answer the first two sub questions at the first phase of the study, writings of the teachers collected through IRF were used. Firstly, the descriptive statistic findings in relation to the content analysis will be given together with the examples from the teachers' writings.

In line with the first research question (What are the entities that have impact on the construct validity of Turkish version of released PISA science items?) three sub questions were investigated through data analysis.

As it was stated at part 4.1, three sub questions under the umbrella of first research question were composed.

i. What are the negative entities embedded within the items that affect measuring and evaluating of a person's science mindness?

ii. What are the positive entities embedded within the items that affect measuring and evaluating of a person's science mindness?

iii. How science items are valid, necessary and important in terms of measuring and evaluating of a person's science mindness?

Content analysis was used in order to answer first sub question related to the first research question. At this part, content analysis steps together with the results will be presented.

The content analysis included three steps at the present study as reduction of thematic units, formation of sub-categories and description of main categories. Creation of the thematic units was an inductive task, based on what respondents said. It began by reading responses from the teachers' writings and then reducing the statements from sentences to form thematic units. For the sub categories, which are called negative entities, thematic units compiled a code list containing a list of letter codes along with their definitions. For example, the letter "k" referred to the negative entity that item included multiple true alternatives. As each new idea or belief was encountered, it was added to the code list. Gorden (1992, p.181) has stated that a useful set of codes should be all-inclusive and mutually exclusive. The final code list contained 29 unique codes and their definitions, each corresponding to specific negative entity and includes thematic units stated at least once by one or more of the 10 teachers on the each science units and 80 teachers as total.

Once a working draft of the code list was developed, the next step was to ensure that different coders could independently replicate each other's work using the same instructions. To pretest the code list and estimate final intercoder agreement, second

expert coded the teachers' writings. Miles and Huberman (1994) and Gorden (1992) emphasize the importance of pretesting and revising code lists, because initial coding instructions often yield poor agreement. In the present study, two coders independently coded the data. The purpose of the first coding comparison was to pretest and remedy problems with the code list. At first, two researchers compared the sets of codes that each coder assigned to teachers' comments. A response was considered to be coded the same only if both coders used the identical set of codes. For example, if one coder assigned the *worse alternative*, *familiarity with item stem*, and *unfamiliar word* codes to a response, the other coder had to assign the same three codes in order for there to be agreement. Presence of one or more disagreements, such as not assigning one of these codes or assigning a fourth code, was counted as a coding discrepancy. Using this method, comparison of the code list pretest results showed that there were only four responses were coded the different by both coders. To eliminate this disagreement, two coders discussed the reasons for their disagreements, the two coders were able to identify and correct problems with the code list. The reasons for the discrepancies included problems such as redundant codes for the same writings and vague code definitions. After resolving the unclear parts of the code list, the two coders formed the final revised code list and reached the full agreement. After completing final list of subcategories, the 29 negative entities classified into major categories. There were three experts that assigned the subcategories into few categories. To carry out the classification, similar process with the formation of subcategories followed. Firstly, researcher and two other experts worked independently and then for the disagreements, coders discussed reasons and finally there were 91 % agreement for the main categories on the coding list. The agreement percentage of the coders were calculated according to the formula proposed by Miles and Huberman, (1994, p.64) Reliability= agreement /agreement + disagreement x 100. Miles and Huberman (1994, p.64) suggest that final intercoder agreement in qualitative data analysis should approach or exceed 90%.

As a summary, thematic units detected in the raw data were found to form 29 negative entities (see Appendix F). These sub categories were classified under the five main categories that they were shown at the Table 7.1 with the number of thematic units included in these main categories. The examples from the writings of teachers will be presented after the explanation of each main category. At the end of the examples, the

participation code of the teachers is given as T5 which stands for the fifth teacher and then abbreviation for the science unit and number of the item which example taken is written in parenthesis.

The main categories of negative entities are presented in Table 7.1. The table shows that there are five main categories (presentation, typicality, structure, content and language) that thematic units classified under.

Table 7.1. Frequency distribution for the categories of negative entities

Main Categories of Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
Presentation	102	17.6	17.6	17.6
Typicality	29	5.0	5.0	22.6
Structure	88	15.2	15.2	37.8
Content	218	37.7	37.7	75.5
Language	142	24.5	24.5	100.0
Total	579	100.0	100.0	

The main category including most thematic units is that of 'content', which refers to the what is said in the text in details as negative entities in terms of national curriculum, topic, culture, clarity of given information and concepts. Excerpts from teachers are displayed bellowed.

> Öğrenciler okulda yada günlük hayatta çok rastlamadıkları bir konu…-this is the subject that students do not meet in daily life or school. (irrelevance with topic, T33, CLOTHES stimulus)

> Okuma parçasına konu olan Büyük Kanyon gerek okul programlarında gerekse öğrencilerin günlük yaşamlarında sıkça karşılaşamayacakları bir olgu, kanyon coğrafi şekli programda çok vurgulanmayan ve Türkiye coğrafyasında Akdeniz bölgesinde rastlanan bir şekil….- the grand canyon forming the subject baseline of the reading passage is a concept that students can not meet in their daily lives, canyon is a geological term that it is not emphasized at the school program and this geographical shape can be seen only in the Mediterranean Region. (cultural unfamiliarity, irrelevance with national curriculum, T41, ACID stimulus)

> Ölçülmek istenen beceriler yeni programla beraber öğrencilerin üzerinde durduğu beceriler,şuan tam yerleşmediler…-…abilities aimed to be measured in this question are the objectives of new program that currently they are not stated exactly. (irrelevance with program objectives, T45, GRAND 2nd item)

The category with the second most thematic units was that of 'language'. 'Language' covers the thematic units commenting on the length of the sentences, unfamiliar words, difficulty in grammar and the quality of expressions. The examples from the teachers' writings are given below.

> Sorulara temel oluşturan okuma parçasının cümlelerinin tamamı çok uzun. Örneğin üçüncü paragrafın ilk cümlesi bir paragraflık yer kaplıyor. Parçada anlatılmak istenen daha kısa ve anlaşılır cümlelerle akatarılabilir. Yeni güçlü zararlı ot tamlaması parçanın bütününde anlamayı zorlaştırıyor…- …the all of the sentences in the reading passage constituting baseline for the questions are so long. For example, the first sentence of the third paragraph covers a place as a one paragraph. The event in the passage can be transfer with shorter and clearer sentences. 'yeni güçlü zararlı ot' clause makes the understanding difficult on the whole passage. (length of the sentence, clause difficulty, T1, GMC stimulus)

> Daha az koruyan demek uygundur, az koruma saglanmaz, ültraviole ışın yerine mor ötesi demek gerekir vede bu seçenekteki nasıl sorusu derecelendirme anlamına geldiğinden tam yanlıştır denemez..- it is more appropriate to say 'daha az koruyan' because 'az koruma sağlanmaz'; it is needed using 'mor ötesi' instead of 'ultraviole' and also in this alternative the question word of how mean a type of graduation and it can not be said this to be wrong exactly. (unfamiliar word, T71, GUNES2)

The category of 'presentation' was also frequently mentioned in terms of thematic units. This category covers comments on the quality of visual elements, lay-out, questioning style and unfamiliarity with the item presentation. Some of the teachers' examples as below:

> Grafikler belirsiz ve konunun sunumu ve sayfa içindeki duruşu değiştirilebilir…- the graphs are ambiguous and presentation of the subject and its lay-out can be changed. (quality of visual elements, lay-out, T23, GREEN stimulus)

> Verilen resmin net olmadığı ve renkli ve daha büyük bir resim kullanmanın daha iyi olacağı söylenebilir…- it can be said that given picture is not clear and it would be better to use a bigger picture. (quality of visual elements, T41, GRAND stimulus)

> Bu tür bir açık uçlu soruda Ali'nin ulaştığı sonucu soru cümlesinden hemen önce vermek gerekir, diğer türlü öğrencinin parçada Ali'nin sonucunu bulması için parçaya tekrar dönmesi gerekir ki bu da zaman harcamasına neden olur ve soruyu cevapsız bırakabilir…- in such an open constructed question, it is needed to give the result of Ali just before the item stem, in the other way students would return to the stimuli to find the result of the Ali and this can cause students to spend time and leave question unanswered. (questioning style, T25, GREEN 1st item)

The category of 'structure' refers to the items quality in terms of incompetent alternatives, multiple answers, incompetent item stem and worse alternatives. It was also quite often mentioned by the teachers. Some examples from writings of teachers are given below.

Eğer A seçeneği doğru kabul edilirse bir önceki soruda cevapta A seçeneği olabilir çünkü koruyucu maddelerin her birini diğerlerine kıyaslarsak ZnO ve minerale kıyaslamış oluruz, bu da soruyu eksik bırakıyor …- if A alternative was accepted as true answer, it is possible to sign the A alternative for the former question because if we compare the preventing matters with each other we compare them with also ZnO and mineral, this causes question to be inadequate. (incompetent item stem, T 74,GUNES2)

Hangisi en iyi nedendir sorusu daha uygun ,birden fazla cevabı var…-….the item stem of which one is the best is more suitable for the question, there are true alternatives more than one. (incompetent item stem and multiple answers, T2,GYD 2nd item)

The category of 'typicality' is concerned with the familiarity with the item stimuli, vague expectancy, extreme easiness and expectancy error. There were only 29 themes in this category. Examples from two teachers are presented below.

Fazlaca sorulan ve çok kolayca cevaplanabilen bir soru…- it is a question that frequently asked and can be replied easily. (extreme easiness, T56, BEDEN2)

Ögrenciler ne tür bir cevap beklendigini anlamaz…- students don't understand which kind of an answer is expected. (vague expectancy, T2, GYD 3rd item)

Bu tür sorular öğrencilerin alışık olduğu tarzda değil…-these types questions are not similar to the ones that students are familiar. (familiarity with item, T73, GUNES stimulus)

Based on the results of the content analysis, five categories representing the 29 negative entities were found. These categories were content, language, presentation, structure and typicality (see Table 7.1.). Content was the most common negative category that teachers commented on. Furthermore, it is noteworthy to mention that these categories are shaped by the researcher and their relation with the literature will be examined in discussion part. For the main categories and negative entities together (see Appendix H).

Research Question 1(ii) is about investigating the positive entities embedded within the items that affect measuring and evaluating of a person's science mindness.

In order to explore the answer for the second sub question (1.ii) data analyzed through the same procedure used for the first sub question (1.i). Positive entities found in the data consisting of the written comments provided by teachers are presented in the tables together with the number of the thematic units representing each category. At first,

total frequencies of the entities for all of the item texts and items will be presented as a whole. Then, one example of the eight science units, each item text separately and related items, as a unit will be analyzed by frequency distribution of the categories. Each category will be exemplified by selected teacher responses on selected item group. At the end of the examples, the participation code of the teachers is given as T21 which stands for the twenty first teacher and then abbreviation for the item which example taken is written in parenthesis.

The same way in the formation of thematic units was followed in the creation of the thematic units for positive entities that it was an inductive task, based on what the respondents said. We began by reading responses from the teachers' writings and then reduced the statements from sentences to form thematic units.

The agreement process between the two independent researchers was carried out for the formation of the sub-categories and main categories of the positive entities. The agreement percentage of the researchers was calculated according to the formula proposed by Miles and Huberman, (1994, p.64) Reliability= agreement /agreement + disagreement x 100. Miles and Huberman (1994, p.64) suggest that final intercoder agreement in qualitative data analysis should approach or exceed 90%. There was 97% agreement between the researchers that this percentage was higher than the agreement value on the main categories for the negative entities. This can be explained by the lower number of the positive entities mentioned by the teachers. Complete list of sub-categories for positive entities composed of 17 items.

As it is seen from the Table 7.2, subcategories of positive entities are classified under four main categories that these were *content, context, composition and science process.*

Table 7.2. Frequency distribution for the categories of negative entities

Main Categories of Positive Entities	Frequency	Percent	Valid Percent	Cumulative Percent
Content	73	23.4	23.4	23.4
Context	65	20.6	20.6	44.0
Composition	97	30.7	30.7	74.7
Science process	80	25.3	25.3	100.0
Total	315	100.0	100.0	

It is noteworthy that there were one item group stimuli and two items that teachers did not comment any positive entity about them. These were CLOTHES stimuli text, CLOTHES second item and GRAND first item.

In the following page, at the table 7.3 the main themes and sub categories of the positive entities are presented together.

Table 7.3. The main categories together with sub categories of positive entities

Main categories	Context	Content	Composition	Science Process
Sub-categories	a. Real issue	c. interesting topic	g. visual element usage	k. scientific investigation
	b. Relevance to possible situation	d.familiarity with subject	h. cognitive level	l. describing science event
		e.consistency with program	i. appropriate item stem	m. using evidence
		f.consistency with objectives	j.appropriate language	
		r.relation to history of science	n.item style	
			o. appropriate alternatives	
			p. only one answer	

The "composition" category includes most thematic units with 97 and most sub-categories that roughly the *composition* category refers to the appearance of text, level of desired ability to get the answer and item properties.

One example from teachers' writings on each thematic units of the composition category will be presented in the following quotations,

> Asit yağmurlarını anlatan bir parçada bunu destekleyici ve etkisinin daha iyi anlaşılmasını sağlayan bir fotoğraf kullanılması güzel….- in a paragraph about the acid rain, it is nice to use a photograph which helps to better understanding of the effect of acid rain. (visual element usage, T12, ACID stimulus)

> Sorunun cevabı için birkaç bilginin aynı ve yorumlanması gerekiyor üst düzey soru yazılması güzel….- in order to answer the question it is required to combine and interpret some pieces of knowledge, it is nice to write a high level question. (cognitive level, T59, PHYSICAL 3rd item)

> Soru kökü iyi yazılmış, sorunun cevaplanabilmesi için gerekli bütün bilgileri içeriyor….-the item stem is well written that it includes all necessary information to be able to answer the question. (appropriate item stem, T41, GRAND 3rd item)

Soruda kullanılan dil anlaşılır ve sade...-the language of the question is clear and correct. (appropriate language, T51, PHYSICAL 2nd item)

Seçmeli bir sorunun ardından açık uçlu bir soru ile bunun nedenini öğrenmeye çalışmak ….- having an open constructed item next to multiple choice one and trying to understand reasoning behind the multiple choice selection. (item style, T80, SUN 4th item)

Hazırlanan seçenekler çok yerinde ve uygun….- the all alternatives are plausible and appropriate. (appropriate alternatives, T45, ACID 2nd item)

"…. Sorunun sadece bir net cevabının olması….- having only one correct and exact answer for the question. (only one correct answer, T63, MARY 1st item)

The category "science process" also well presented in term of thematic units which are called as scientific investigation, describing science event and using evidence. The mental actions to solve the processes in the items were quite often mentioned by the teachers. The examples of the teachers' comments are presented in below quotations;

Asit yağmurunun kaynaklarını yazmak için kimyasal bir olayı tarif edebilmek…. To be able to describe a chemical process to write the sources of acid rain…." (describing science event, T17, ACID 1st item)

Soruya cevap verebilmek için verilen grafiklerdeki bilgilerin yorumlanmak durumunda olması….- the necessity of using information presented in graphical form to answer the question. (using evidence, T47, GREEN 1st item)

Deneyin amacına yönelik olarak değişken kavramının sorulması – asking the concept of variables according to aim of experiment. (scientific investigation, T16, GMC 1st item)

The category of 'content' refers to what is said in the text in terms of national curriculum, topic, culture, clarity of given information and concepts. Examples from two teachers are presented bellow.

Konusu farklı ve ilginç….- the subject is different and interesting…." (interesting topic, T85, SUN stimuli)

Spor yapmanın önemi ve sağlığa yararları müfredatla örtüşen bir konu….- the importance of physical exercise and its benefits to health is a subject that covered in the national program …." (consistency with national program, T54, PHYSICAL stimuli)

Aşı konusu ve aşının vücut içindeki işleyişi biyoloji derslerinde önemle üzerinde durulan ve sorulması önem arzeden bir soru…- the subject of vaccination and its function in the body is mostly noticed in the biology lessons and it is crucial to related questions. (consistency with objectives of national program, T62, MARY 2nd item)

> Fen konularında ve sorularında tarihsel süreç ve gelişimlere yer vermek olumlu….- it is positive to include historical events and developments in the science subjects and questions. (history of science, T68, MARY stimuli)

> Fiziksel aktivite ve bunun vücudumuz için yararı ve vucutta meydana gelen biyolojik değişiklikler öğrenciler için tanıdık bir konu…- physical activity and its benefits to our body and biological changes occurring in the body is a familiar subject for students. (familiarity with topic, T55, PHYSICAL stimuli)

The last category for the positive entities described for the science-units is 'context' that it contains two sub-categories of 'real issue' and 'relevance to possible situations'. Two teacher examples are below:

> Sadece kimya bilgisi olarak değil de gerçek hayattan alıntı yaparak sorulması güzel….- it is nice asking o a real life issue instead of asking as chemistry knowledge. (real issue, T13, ACID stimuli)

> Öğrencinin kendi yakın çevresinde veya şehrinde, ülkesinde gerçekten karşılaşabileceği bir durum olması….- because it is a real fact that students can meet in near environment or in their country. (relevance to possible situation, T16, ACID 1st item)

To sum up, there are four positive entity categories (content, context, composition and science process) with 315 thematic units in total. To show the content analysis process in science units, Genetically Modified Crops (GMC) item group is selected as example science unit that frequency tables with the examples of the teachers' comments will be introduced below. Statistics for the remaining seven science units will be presented in Appendix F.

Table 7.4. Frequency distribution for the positive entities: GMC Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. Real issue –*context*	6	54.5	54.5	54.5
k. Scientific investigation – *science process*	5	45.5	100.0	100.0
Total	11	100.0	100.0	

Table 7.4 shows the distribution of the 11 positive entities commented in the stimulus of GMC science unit. The *real issue* entity was found to be most frequent entity which is followed by *scientific investigation*. *Real issue* entity placed by six teachers, and the *scientific investigation* was written by five of the teachers. The examples of the teachers' writings will be given below.

One of the teachers commented on the *real issue* entity as;

> Teknoloji ve bilimin gelişimiyle beraber bazı sorunlarda ortaya çıkmıştır, konunun gerçek bir olayla ilgili olması güzel…- some problems have appeared with the development of science and technology, it is good the subject related with a real situation.... " (T4)

Another teacher wrote a similar clause;

> Yazıda anlatılanların gerçekte var olan bir problemle ilgili olması….- the relation of the subject in the written paragraph with a problem presence in real life. (T6)

Five of the teaches referred to the *scientific investigation* included in the stimuli, two examples from the teachers' comments as below;

> Sunulan probleme dair deneyin okuma parçası içinde yer alması….- the presence of a scientific experiment related with the problem in written paragraph. (T10)

> Araştırma metodolojisi üzerinde durulması ve bir ülkede zarar yaratabilecek bir sorunun bu şekilde araştırılması…- including methodology of the investigation and examining a problem that can cause harmful effects in a country. (T2)

Table 7.5 shows the only types of positive entity characterized in the first item of the GMC science unit. Six of the teachers described *scientific investigation* which belongs to the category of science process.

Table 7.5. Frequency distribution for the positive entities: GMC 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
k. Scientific investigation-*science process*	6	100.0	100.0	100.0
Total	6	100.0	100.0	

One example from the teachers' comments is like;

> Sorunun içeriği değişen iki koşulun deney metodlarına gore yorumlanması ile ilgili, bu güzel….- it is good that the item is asking interpretation of two changing situation in terms of experiment method. (T7)

Another example is like;

> Deney yaparken amaca yönelik değişken kotrolünü sorması….- asking the control of variables in relation to aim of experiment while carrying out an experiment. (T1)

As it is shown from the Table 7.6, second item of the GMC science-unit appears to have one positive entity of *describing science event*.

Table 7.6. Frequency distribution for the positive entities:GMC 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l.Describing science event-*science process*	8	100.0	100.0	100.0
Total	8	100.0	100.0	

One of eight teachers comments on the item as below;

> Yapılan deneyde anlatılan bir durumda bilimsel araştırmanın önemli özelliklerinden birini sorması- questioning to recognize one of the important features of scientific investigation for experiment described in the written paragraph. (T5)

Table 7.7. Frequency distribution for positive entitites: GMC 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
n. Item style-*composition*	7	100.0	100.0	100.0
Total	7	100.0	100.0	

The third item of the GMC science unit was found to include only one theme in the *composition* positive entity sub-category. Two of the teachers exemplified it as in the below quotes,

> Açık uçlu soru olmasını pozitif bir yön olarak söyleyebiliriz, öğrencilerin soru hakkında ne bildiklerinin daha detaylı anlama şansımız olabilir….-it can be said open constructed item to be a positive feature, in this way we can get a chance to learn what students know about the subject in detail. (T10)

> Soru kümesinde birde açık uçlu bir sorunun olması iyi….- it is good to be an open ended item in the item group. (T3)

In summary, there was found that there were 17 type positive entities described by the teachers for the PISA science units. These were classified under the four main categories of content, context, science process, and composition. It is noteworthy to state that categories defined for the positive entities had limited extent when they compared with the categories of the negative entities. Also, the number of the thematic units mentioned for the positive entities were lower than the number of the thematic units in negative entities.

In the following part, statistics derived from third part of the IRF about overall validity, necessity and importance of the items will be presented in order to find answer for the third sub question (1.iii).

Research Question 1(iii) was about investigating science items' validity, necessity and importance in terms of measuring and evaluating of a person's science mindness.

The third part of the IRF asks teachers to rate overall validity, necessity and importance of the each stimuli given for item groups and item for measuring and evaluating the range of the scientific mindness of the students. Teacher gave a rank between 1 and 9 for each of 8 texts and related 24 items. The results will be presented in terms of means for each science units as the combination of mean values of stimuli and items and then separately for the stimuli of science units and items included in the science units at the Table 7.8.

Table 7.8. Descriptive statistics of overall Validity, Necessity and Importance

Science units	Validity Mean	*Rank*	Necessity mean	*Rank*	Importance mean	*Rank*
GMC	3.30	7.	4.25	4.5	5.00	3.
ACID	4.75	4.	4.78	3.	4.92	4.
GREEN	5.01	1.	5.25	1.	5.27	2.
PHYSICAL	5.00	2.	5.15	2.	5.42	1.
MARY	3.57	6.	3.87	7.	4.83	6.
SUN	3.97	5.	4.25	4.5	4.57	7.
CLOTHES	1.63	8.	1.75	8.	2.00	8.
GRAND	4.92	3.	4.25	4.5	4.85	5.

Besides the mean values calculated, the rankings of each science units on the three evaluation criteria were given in the table that rankings serve as a tool for showing up what came at the top and the bottom of the lists as well as for describing the general tendencies on validity, necessity and importance of PISA item groups.

The overall indices of the item validity, necessity and importance show that the teachers participating in the study found that PISA items were moderately important, lowly necessary and had lower validity to improve scientific mindness of students. That is to mean, nearly all the item groups are deemed to have same value sequences from higher importance to lower validity with the exception of GRAND group which has highest value in validity. The validity ratings vary between 5.01 and 1.63 and item necessity ratings between 5.25 and 1.75. Additionally, item importance ratings are between 5.42 and 2.00. Thus, only two of the science units (GREEN and PHYSICAL) achieve a value just above the value 5. However, it is to be notice that overall indices for the science units are the means calculated using the values of stimuli and items included in science units for the three criteria of validity, necessity and importance. This obscures differences between the stimuli and items and also individual item differences in the science units. Because of that, it is decided that presenting separate tables showing values on the validity, necessity and importance for each item group in terms of its

stimuli and item mass (means of the remaining items' mean rather than the stimuli) will be meaningful.

Table 7.9. Descriptives on overall Validity, Necessity and Importance –Stimuli-

Stimuli of Science units	Validity mean	*Rank*	Necessity mean	*Rank*	Importance mean	*Rank*
GMC	2.80	3	4.40	2	5.80	1
ACID	3.60	2	3.40	4.5	5.00	3
GREEN	3.70	1	4.50	1	4.10	4
PHYSIC	2.50	5	1.70	8	2.00	8
MARY	2.70	4	3.40	4.5	5.20	2
SUN	2.40	6	3.60	3	3.10	5
CLOTHES	1.80	7.5	2.00	7	2.70	6
GRAND	1.80	7.5	2.10	6	2.20	7

Table 7.9 shows the ratings of the stimuli given at the beginning of the each science units, the validity of the stimuli range between 3.70 and 1.80, while necessity is between 4.40 and 1.70 and importance ratings are between 5.80 and 2.00. These values seem to vary a lot for three attributes. From the point of validity, stimuli of CLOTHES and GRAND science units have the lowest values with 1.80 while the stimulus of GREEN group has highest with 3.70. For the necessity, stimulus of GREEN science unit placed 1st (necessity value 4.50) while the PHYSICAL stimulus ranked last (necessity value 1.70). The GMC stimulus is 1st on the importance list (importance value 5.80) and PHYSICAL stimulus on the 8th with the value of 2.00. Also, it is noteworthy that there is no general tendency to form a particular sequence among the values of validity, necessity and importance. Some of the group stimuli have order from highest value of importance to the lowest value of validity (GMC, MARY, CLOTHES and GRAND). However, stimuli of GREEN and SUN science units have highest means for the necessity which is followed by importance and validity. PHYSICAL and ACID stimuli have different orders from mentioned has two groups and each other that ACID group stimulus has highest value for importance, then, for validity and necessity while

PHYSICAL science unit stimulus gets more for validity, importance and necessity respectively. At the Table 7.10 the means for the items in science units will be given.

Table 7.10. Descriptives on overall Validity, Necessity and Importance -Items-

Items of Science units	Validity	*Rank*	Necessity	*Rank*	Importance	*Rank*
GMC	3.47	6	4.20	4	4.70	5
ACID	5.13	3	5.23	3	4.90	3
GREEN	5.47	2	5.50	2	5.67	2
PHYSIC	5.83	1	6.30	1	6.57	1
MARY	3.97	5	4.13	5	4.87	4
SUN	3.38	7	3.35	7	3.80	7
CLOTHES	2.35	8	2.50	8	2.65	8
GRAND	4.48	4	3.73	6	4.30	6

Table 7.10 shows the ratings of the remaining items given after the stimuli of each science units, PISA science units include two, three or four items, in the table the means are the overall means for all items except the stimuli in the science units. The validity of the items varies between 2.35 and 5.83. The means of the necessity attribute is between 2.50 and 6.30. Importance ratings change between 2.65 and 6.57. It is noteworthy that all of the value ranges are higher than the value ranges of the stimuli. It is to be noticed that PHYSICAL group items have 1st rank for all categories while its stimulus has lowest degrees. From the point of rating tendency among the validity, necessity and importance values for each item group, five item groups -GMC, GREEN, PHYSICAL, MARY and CLOTHES- have importance, necessity and validity order from highest value to the lowest. SUN group follows the importance, validity, necessity; GRAND group has validity, importance, necessity and ACID group shows necessity, validity, and importance sequences.

It is seen from the Table 7.9 and Table 7.10 that stimuli have an effect on the overall means of the science units. The means of the stimuli decrease the overall validity means of the seven science units except the CLOTHES science unit. In a similar

way, they reduce the overall necessity mean of the six science units except CLOTHES and GMC science units. For the overall means of importance, the values of stimuli increase the overall means of the four science units (GMC, ACID, MARY and CLOTHES) whereas decrease overall means of the remaining four science units.

In order to check if teachers assign similar sequences to the attributes of validity, necessity and importance of the science units, the Friedman test was applied to the total ratings of science units and also to each science unit separately. So, it is seen that the mean values of the all three criteria (validity, necessity and importance) on the nine-point scale were lower than the average value of five. However, as it is seen form the Table 7.11, there was found significant difference on the means of the three criteria.

Table 7.11. Friedman test descriptives for science units on the Validity, Necessity and Importance criteria

	N	Mean	SD	Min	Max	Mean Rank
Validity	80	3.86	1.01	1.3	5.5	1.55
Necessity	80	4.10	1.07	2.0	8.0	1.85
Importance	80	4.44	0.95	2.3	6.0	2.60

The mean values for the validity, importance, and necessity criteria are 3.8, 4.1 and 4.4 respectively. The rank sequence found among the ratings of validity, necessity and importance show that the importance gets the significantly higher values than the validity and necessity. [The test statistics show that there is a statistically significant finding. The p-value (asymp. Sig. in the table above) is $p = 0.00$. (A p-value less than 0.05 is said to be statistically significant.)].

In addition to the results for the overall ratings, the results of the Friedman test for the each science unit will be presented on the following page.

Table 7.12. Friedman test results for science units on the Validity, Necessity and Importance

Science unit	Rank 1	Rank 2	Rank 3	P value (Asymp. Sig.)
GMC	Importance	Necessity	Validity	0,000*
ACID	Importance	Necessity	Validity	0,562
GREEN	Necessity	Importance	Validity	0,565
PHYSICAL	Importance	Necessity	Validity	0,018*
MARY	Importance	Necessity	Validity	0,000*
CLOTHES	Importance	Necessity	Validity	0,001*
SUN	Importance	Necessity	Validity	0,001*
GRAND	Importance	Validity	Necessity	0,002*

* p-value less than 0.05 is said to be statistically significant

As it is seen for the Table 7.12 for ACID and GREEN science units, there is no evidence that distributions of three types of scores are different ($p > .05$). That means, for these two science units teachers did not assign significant sequence to the three attributes. Other six science units have significantly different scores for the validity, necessity and importance variables ($p< 0.05$) that validity has lowest rank for five of them and necessity has lowest rank for one of them. This result implies that while the mean values of validity, necessity and importance criteria lower than average value of the nine-point scale, teachers' ratings among these three criteria differentiate for the importance of science units that they give meaningfully higher scores for the importance criterion rather than there is not a significant difference among the ranks of the scores about the validity and necessity criteria.

At the end of data analysis related with the first phase of the study, the revision of the science units was made to form the PISA-RT test which included eight science units including eight stimuli and 21 items. The PISA-RT test was used at the second phase of the study together with the other instrument, PISA-OT test, which consisted of the Turkish version of the eight science units used in PISA 2006 study and detected in the first phase of the present study. The three of the items excluded from the PISA-OT test to provide the equity between comparison groups.

With respect to the process of revision of the items to build the new items, it was created paying attention to the original Turkish revision that had taken place previous for data collection in PISA 2006. The process of revision of the items involved a complex procedure to keep the nature of the questions, the way in which they are planned, the order in which to specify the questions. As the Duverger (1996), all these elements conform and affect directly the results obtained. Attempts were made not to leave any important element aside that the whole process of formation of PISA-RT was heavy and difficult, it involved discussion with experts. Since, it could be reached other similar research about the revision of subject, the revision process carried out bearing in mind the main goals being brought up in the research. The revision of the items to form the new items involved the following steps:

- Construction of frequency tables to show the negative entities found by ten teachers for each science unit separately for each stimuli and item in it.
- Deciding negative entities will be included in the revision process of the science units through the judgments of researcher and one measurement expert.
- Revision of the items together with one Turkish language expert and two science teachers.

The results of the reviews about the negative entities found in the science units are organized by frequencies of the negative entities and presented in below tables for each stimuli and item in the science units. In the tables, there will be 'negative entity (sub-category) and main category' columns which show the codes of sub-category with its name that it is followed by name of the main category to which sub-category belongs. For example, a. Long Sentence –*Language* will mean 'long sentence' sub category with the code 'a' and this sub-category classified under the general category of 'Language'. Also, examples from the teachers' writings will be presented for each different negative entity. That is to mean, there will be no reputation of the examples for the same negative entity in different science units or items. At this part of the study, revision process of the CLOTHES science unit consisting of stimulus and two items will be explained in detail as example. For the revision of the remaining seven science units, see Appendix F.

CLOTHES science unit consists of one stimulus in the form of a reading passage and two items, one of them is in the yes-no item format and the other is a multiple choice item (see Appendix B).

Table 7.13 shows distribution of 40 thematic units classified under 10 negative entities in the stimulus of CLOTHES science unit. The most frequent negative entity was found to be *irrelevance with topic* which is followed by *inappropriate lay-out* and *unnecessary context*. *Difficulty in statement* and *irrelevance with the national curricula* were placed by three of the teachers. Only two teachers mentioned negative entities of *difficulty in clause*, *difficulty in grammar* and *irrelevant cue.* The least frequent ones were the *long sentence and unfamiliar item stem* that they were recognized by one teacher. For the revision of the stimulus, it was decided not to be able to include the negative entities of *irrelevance with the topic*, *irrelevance with the national curriculum and unfamiliar item stem*. Additionally, although five of the teachers found the stimulus to be unnecessary to answer the items, the researcher and the experts agreed on the revision of the stimuli. This was because of the practical reasons that it was necessary to provide students with the similar conditions in terms of reading material and time for the answering items. Therefore, revision of the stimuli based on the language, structure and typicality categories. Before detailed examination of the revision process applied, examples of the teachers' writings will be given.

Most of teaches referred to the reading passage as an irrelevant topic for students. Three examples from teachers' writings are presented below.

> Verilen konu hem hedeflenen öğrenci kitlesine çok uzak…- given subject is unfamiliar to the target student population. (T32)

> Ama öğrencilerin okulda ya da günlük hayatta çok rastlamadıkları bir konu….- but the topic is not so common in school or daily life for students. (T33)

> Bahsedilen konu itibariyle öğrencilere tanıdık değil….- in terms of the given topic it is not familiar to the students. (T 40)

Inappropriate lay-out was found to be another negative entity by the teachers. Some of the teachers commented as in the below two examples;

> Ilk olarak verilen yazının düzenine bakıldığında yazı stilinin okuma güçlüğü yarattığı söylenebilir, tüm cümleler her iki satır basına yaslanmasa okunması daha kolay olabilir…- At first, when it is looked to the lay-out of the passage, it can be said that the writing style creates difficulty, also it can be better if the paragraph alignment is chosen to be left rather than the justified. (T31)

> Paragraf biçimleri ilk bakışta birbirinden farklılık gösteren bir yazı…- at first glance, the styles of the each paragraph is different from each other. (T37*)*

Another negative entity was the *unnecessary context* that teachers found the reading passage not to be required for replying the questions. Two examples are presented below.

> Soruların cevaplanması için okuma parçası gerekli değil...-the reading passage is not necessary to answer the questions. (T34)

> Verilen parçada ile ikinci soru arasında zoraki bir ilgi var, birinci soru içinde parçanın son paragrafı yeterli hatta olmasa da olur…- there is a weak relationship between the second question and the reading , the last paragraph is enough to answer the first question even it can not be. (T30)

The following sentences are the examples of the negative entities of *difficulty in clause*, *difficulty in grammar* and *irrelevant cue* respectively.

> Parçada kullanılan bazı kelimeler yerine daha etkili ve anlaşılır olanları kullanılmalı; 'elektro tekstil' yerine……., konuşmalarının başkaları tarafından anlaşılır duruma gelmesini sağlamaktadır yerine söylemek istediklerinin başkaları tarafından anlaşılmasını sağlamaktadır gibi…-there must be used more effective and understandable words instead of some words used in the reading passage; for example …….in the place of electro textile and by omitting ‘duruma gelmesi’. (T35)

> Parçada kullanılan gramer anlaşılmayı oldukça zor hale getiriyor..-the grammar used in the reading text make the understanding hard. (T31)

> Son paragraf farklı yazılmış bu da birinci sorunun cevabını öğrencinin buradan kolayca bulmasına yol açıyor…- the last paragraph of the text is written in an different format that leads to students to search the answer of the first question in this part. (T33)

Table 7.13. Frequency distribution for negative entities: Clothes Stimulus

Negative entities	Frequency	Percent	Valid Percent	Cumulative Percent
a.Long Sentence -*Language*	1	3.3	3.3	3.3
b.Difficulty in clause-*Language*	2	4.7	4.7	10.0
c.Difficulty in statement-*Language*	5	14.7	14.7	20.0
d. Irrelevance with national curriculum- *Content*	3	6.0	6.0	30.0
e.Unnecessary context-*Content*	5	14.7	14.7	46.7
g.Difficulty in grammar-*Language*	7	26.7	26.7	53.3
q.Irrelevance with topic-*Content*	9	20.0	20.0	73.3
u.unfamiliar item stem- *Typicality*	1	3.3	3.3	76.7
w.irrelevant cue-*Structure*	2	4.7	4.7	83.3
z. inappropriate lay-out-*Presentation*	5	14.7	14.7	100.0
Total	40	100.0	100.0	

The stimulus of the CLOTHES science unit was found to include *long sentence* negative entity by one of the teachers and also to be an *unfamiliar item stem* by the teacher. One example is showed below.

> Bu gibi okuma materyallerinin öğrencilerin ilgilerine ne derecede hitap ettiği önemlidir, bu kadar uzun ve anlaşılması zor bir yazıyla ilgili soruları öğrencilerin cevaplaması için hayatında bu konuya ilgi duyması yada bu konuyla fazlaca karşılaşması gerekir ki sınavlarda çok karşılaşılan bir soru kökü değil bu…..-…. for such reading materials, how the passage is interesting for is an important point. For students, in order to answer questions related with the reading which is long and can not be easily understood the necessary to be interested in such a topic or to meet such texts frequently. However, it is not such a common item stem that students acquainted with in the exams. (T35)

Under the light of comments of the teachers, the revision process for the stimulus of CLOTHES science unit began with the change of the lay-out of the reading passage that all passages were arranged to have same writing character and alignment. For the negative entities classified in the language category, all sentences were revised to be written in the simple present tense, difficulties is clauses were diminished by replacing words with more appropriate ones and long sentences are divided into two sentences. For the revised version of the CLOTHES stimulus see Appendix F.

Table 7.14. Frequency distribution for negative entities: CLOTHES 1st Item

Negative entities	Frequency	Percent	Valid Percent	Cumulative Percent
g.Difficulty in grammar-*Language*	3	17.6	17.6	17.6
h.Uncommon vocabulary-*Language*	2	11.7	11.7	29.3
r. questioning style –*Presentation*	5	29.4	29.4	58.7
q.Irrelevance with topic-*Content*	1	5.8	5.8	64.5
3. measuring different objective from the aimed – *Structure*	4	23.5	23.5	88.0
u.unfamiliar item stem- *Typicality*	2	12.0	12.0	100.0
Total	17	100.0	100.0	100.0

As it is shown from the Table 7.14, first item of the CLOTHES science unit appears to have six types of negative entities. *Questioning style* was the most mentioned negative entity embedded in the item. *Different objective from the aimed* was the second frequent one that mentioned by four the teachers. Only two teachers identified the *uncommon vocabulary* and *unfamiliar item stem* in the item. Irrelevance with the topic was mentioned by one teacher. Teachers examples for the *uncommon vocabulary* and *irrelevance with topic* were as in the below quotations,

> 'sav' sözcüğü okullarda ve günlük hayatta çokça kullanılan bir kelime değil, ayrıca 'sav' zaten ileri sürelen şey demektir dolayısıyla ileri sürülen demeye gerek yoktur …-the 'sav' (claim) is not a common word used in the school or daily life and it means to say that something true so it is not necessary to repeat it in the sentence. (T40)

> Laboratuar da birşey test etme öğrencilerin soru bazında sıkça karşılaştıkları bir konu değil – testing something in the laboratory is not a frequent subject especially on the questions. (T36)

The revision of the first item based on the categories of language that two of the words were omitted and one of the words changed with the more common one. The presentation of the item was kept as original because it required changing the item style from yes-no question to the multiple choice. Due to the fact that Turkish students are more familiar with the multiple choice items, format was unchanged. For the revised version of the first item in CLOTHES science unit, see Appendix F.

Table 7.15. Frequency distribution for negative entities: CLOTHES 2nd Item

Negative entities	Frequency	Percent	Valid Percent	Cumulative Percent
r. questioning style –*Presentation*	3	23.0	23.0	23.0
3. measuring different objective from the aimed – *Structure*	8	61.6	61.6	84.6
2. error expectancy- *Typicality*	2	15.4	15.4	100.0
Total	13	100.0	100.0	100.0

Table 7.15 shows the three types of negative entities found in the second item of the CLOTHES science unit. The most frequent negative entity was *different objective from the aimed* which belongs to the category of structure. Other entity was noticed by less than five teachers. Three teachers mentioned *questioning style* of the item to be inappropriate. Two of the teachers referred to the item to be unclear in terms of students' expectance. There will be one example for the each of the sub-categories from the comments of the teachers below.

> Elektrik iletkenliğini deneyebilmek için gerekli araçlar arasında şıklardan herhangi biri yer alabilir öğrencilerin bir kısmı bu aletlerin çeşitli kombinasyonlarını düşünebilir; bu soru öğrencilerin yaratıcılıklarını dener...- there can be any of the alternatives among the answers of the question of which can be among the devices to try the electricity conduction, some of the students can think on the combination of the devices so that the question asks for the creativity of the students. (T36)

> Soru kökü farklı şekilde yazılsa daha iyi olur..- it can be better to rewrite the item stem. (T33)

> Çok basit bir soru fakat öğrenciler test bilinci ile farklı şekillerde düşünebilirler..- the question is simple but it can be leads students to think in another way with the awareness of the testing. (T40)

The second item of the CLOTHES science unit, the *questioning style* of the item was revised and the item was changed from the 'which can be among the necessary devices…' to 'which is the necessary device…'

As a summary of revision process, the frequency tables of the negative entities were prepared for each stimulus and items for all the eight science units and revisable negative entities were decided with the help of experts, at the revision process researcher worked cooperatively with one Turkish language expert and two science teachers. Then, the revisions for all science units were completed to form the PISA–RT test as one of the instruments for the second phase of the study.

First research question was about investigating the affect of negative entities on PISA science items' construct validity throughout the achievement scores of the students. The two related hypotheses were as below.

i. There will be significantly higher scores on the PISA-Revised Turkish test (PISA-RT) than scores on PISA-Original Turkish test (PISA-OT) for the whole tests.

ii. There will be significantly higher scores on the PISA-Revised Turkish test (PISA-RT) than scores on PISA-Original Turkish test (PISA-OT) for each item.

In order to find out whether negative entities examined by the teachers for the released PISA 2006 science questions have an effect on the achievement of the students, a selected group of students (n=60) was presented with the two groups of test in which students answered the original Turkish version of items (PISA-OT) and revised Turkish version (PISA-RT) of the items. As it is mentioned at the part 7.2, revision of the items performed with the negative entities that they were referred to be revisable entities. It is noteworthy that the all revised negative entities are accepted to be one variable in the third phase to investigate the effect on the students' achievement scores.

There were 30 students who were presented with the original Turkish version of the items and other 30 students with the revised Turkish versions. So, there were two comparison groups. In order to test the hypothesis of there is a significant difference between the achievement scores of the 15-year-old students who answer the original version of the items and revised version of the items on the whole test and at the item level, it is noteworthy to remember that the previous science knowledge of the students in the two groups compared by using GTT scores (see p.50) and there were no statistically significant difference between the general and science subject GTT scores. Therefore, it could be said that two equalized groups were formed.

The mean of scores in both groups is calculated, and it is found to be M= 0.790 in the first comparison groups and M=1.007 in second comparison group. Table 7.16 shows the descriptive statistics related to total scores of comparison groups.

Table 7.16. Descriptive statistics related to the PISA-OT and PISA-RT total tests

	Groups	N	M	SD	SE Mean
Total scores	Group PISA-OT	617	0.79	0.827	0.033
	Group PISA-RT	638	1.00	0.826	0.032

As it is seen from the Table 7.16 there is an increase in the mean score of the first comparison group which is called as Group PISA-RT which got the PISA- RT test when it is compared to the mean score of the first comparison group, which is called as Group PISA-OT which took PISA-OT test. Independent sample t-test was carried out between the scores of comparison groups in order to determine whether this increase is statistically significant or not (see Table 7.17).

Table 7.17. T-test results between the PISA-OT and PISA -RT scores

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference
Scores	Equal variances assumed	21.78	0.000	-4.64	1253	0.000	-0.21
	Equal variances not assumed			-4.63	1251.23	0.000	-0.21

Since the Levene' Test significance value for the tests (p= 0.00) is not greater than 0.05, equal variance cannot be assumed. This means that the t-test significance value is the one on the second rows of t-test statistics. The t-test significance values are less than 0.05 (p= 0.00) indicating that there is a statistically significant difference between the two groups in terms of test scores. In other words, it is found that there is a statistically significant difference between mean scores of PISA-OT test and PISA-RT test calculated for answers of 15-year-old students on each science items. It can be implied

that there is a significant effect of the negative entities on the total achievement scores of the comparison groups that the scores calculated for the whole tests. So, it was succeed to reject null hypothesis and the research hypothesis 2.i was supported.

Additionally, the analysis of individual items was completed. In order to test whether there is a significant difference between the scores of comparison group took the PISA-OT test and group took PISA-RT test on each item separately, independent sample t-tests carried on between scores of groups on each item. The group statistics for the change in each individual item are presented in Table 7.18A and Table 7.18B. The tables include the standard error means for each group's scores. Besides, the numbers of both comparison groups is presented because the numbers vary. For groups, this variation reflects the fact that at some instances students were directed to skip the questions. Since the analysis includes 22 individual items placed in the eight science units, descriptive statistics on the individual items will be presented in Table 7.19A and Table 7.19B by presenting results of ten items in one table and remaining 12 in the other. Then, the corresponding independent sample tests are presented in the following tables.

Table 7.18A. Descriptive statistics of groups on 10 individual items

Items in Science Units	Groups	N	Mean	SD
GMC2	Group PISA-OT	30	0.6667	0.47946
	Group PISA-RT	30	0.9000	0.30513
ACID1	Group PISA-OT	22	0.5909	0.85407
	Group PISA-RT	28	1.2143	0.78680
ACID2	Group PISA-OT	30	0.5333	0.50742
	Group PISA-RT	30	0.5667	0.50401
ACID3	Group PISA-OT	25	0.5600	0.65064
	Group PISA-RT	25	0.7600	0.72342
GREEN1	Group PISA-OT	22	0.5455	0.50965
	Group PISA-RT	28	0.7143	0.46004
GREEN2	Group PISA-OT	27	0.6296	0.74152
	Group PISA-RT	28	0.7857	0.68622
GREEN3	Group PISA-OT	23	0.3478	0.48698
	Group PISA-RT	26	0.4615	0.50839
PHYSICAL1	Group PISA-OT	30	2.1333	0.50742
	Group PISA-RT	30	2.2667	0.52083
PHYSICAL2	Group PISA-OT	30	1.8000	0.40684
	Group PISA-RT	30	1.9000	0.30513
PHYSICAL3	Group PISA-OT	30	0.3000	0.46609
	Group PISA-RT	30	0.6333	0.49013

As it is seen from the Table 7.18A, all students in both of the comparison groups answered the three items in the PHYSICAL science unit and second item of the GMC science units while there are unanswered items by the students for the other items. Besides, the number of the students do not answer the items were more in the Group PISA-OT than the Group PISA-RT. Moreover, the mean values for each of the 10 items show that there is an improvement in the scores of the second comparison group in terms of the science achievement.

Table 7.18B. Descriptive statistics of groups on 12 individual items

Items in Science Units	Groups	N	Mean	SD
GRAND2	Group PISA-OT	30	1.3000	0.65126
	Group PISA-RT	30	1.3667	0.63968
GRAND3	Group PISA-OT	30	0.6000	0.49827
	Group PISA-RT	30	0.7667	0.43018
GRAND4	Group PISA-OT	30	0.5667	0.50401
	Group PISA-RT	30	0.9000	0.30513
CLOTHES1	Group PISA-OT	30	2.5667	0.93526
	Group PISA-RT	30	3.2000	0.76112
CLOTHES2	Group PISA-OT	30	0.5667	0.50401
	Group PISA-RT	30	0.9000	0.30513
MARY1	Group PISA-OT	30	0.6000	0.49827
	Group PISA-RT	30	0.8333	0.37905
MARY2	Group PISA-OT	30	0.5333	0.50742
	Group PISA-RT	30	0.7667	0.43018
MARY3	Group PISA-OT	28	0.5929	0.66309
	Group PISA-RT	28	0.6271	0.49735
SUN1	Group PISA-OT	30	0.4667	0.50742
	Group PISA-RT	30	0.4667	0.50742
SUN2	Group PISA-OT	30	0.5333	0.50742
	Group PISA-RT	30	0.8333	0.37905
SUN3	Group PISA-OT	30	0.3333	0.47946
	Group PISA-RT	30	0.6667	0.47946
SUN4	Group PISA-OT	21	0.1429	0.47809
	Group PISA-RT	25	0.3200	0.55678

Table 7.18B presents the descriptive statistics for the remaining 12 individual items included in four science units. As it is seen from the table, all students in the comparison groups also answered all items in the GRAND and CLOTHES science units together with the first and second items of the MARY science unit and also first, second and third items of the SUN science unit. Only third item of the MARY science unit and forth item of the SUN science unit were no answered by all of the students. There was an improvement at the achievement scores of the second comparison group took the PISA-RT for 11 of the items that the mean scores of the groups were same for the first item of the SUN science unit.

It is noteworthy to mention that the all of the items students did not answer and skipped were the open constructed items; there were no missing response for the

multiple choice questions. In this situation, the findings differs from the PISA 2006 study for Turkish population, but the result for this difference can be explained by the small number of the items and small number of the students in the present study when it is compared with the original PISA study.

Table 7.19A shows the t-test results on the achievement scores of students in comparison groups for 22 items individually.

Table 7.19A. Independent Sample t-tests on individual 10 items

Items in Science Units		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	Df	Sig. (2-tailed)
GMC2	Equal variances not assumed	0.658	0.000	-2.249	49.180	0.029*
ACID1	Equal variances assumed	0.530	0.470	-2.678	48	0.010*
ACID2	Equal variances assumed	0.236	0.629	-0.255	58	0.799
ACID3	Equal variances assumed	0.079	0.780	-1.028	48	0.309
GREEN1	Equal variances not assumed	4.150	0.047	-1.213	42.839	0.232
GREEN2	Equal variances assumed	0.918	0.342	-0.811	53	0.421
GREEN3	Equal variances assumed	2.067	0.157	-0.797	47	0.430
PHYSICAL1	Equal variances assumed	1.180	0.263	-2.004	58	0.019*
PHYSICAL2	Equal variances not assumed	4.930	0.030	-1.077	53.783	0.286
PHYSICAL3	Equal variances assumed	1.143	0.289	-2.699	58	0.009*

As it is seen from the Table 7.19A, there were three items (GMC2, ACID1, PHYSICAL1 and PHYSICAL3) that scores of the comparisons were significantly different from each other and for the remaining eight items there found no significant difference between groups.

Table 7.19B. Independent Sample t-tests on Individual 12 Items

Items in Science Units		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	Df	Sig. (2-tailed)
GRAND2	Equal variances assumed	0.071	0.791	0.200	58	0.842
GRAND3	Equal variances not assumed	7.162	0.010	-1.387	56.791	0.171
GRAND4	Equal variances not assumed	45.298	0.000	-3.099	47.740	0.003*
CLOTHES1	Equal variances assumed	1.654	0.203	-2.877	58	0.006*
CLOTHES2	Equal variances not assumed	45.298	0.000	-3.099	47.740	0.003*
MARY1	Equal variances assumed	16.626	0.000	-2.041	54.144	0.046*
MARY2	Equal variances not assumed	10.933	0.002	-1.921	56.487	0.049*
MARY3	Equal variances assumed	0.622	0.434	0.871	54	0.448
SUN1	Equal variances assumed	0.000	1.000	0.000	58	1.000
SUN2	Equal variances not assumed	22.338	0.000	-2.594	53.680	0.012*
SUN3	Equal variances assumed	0.000	1.000	-2.693	58	0.009*
SUN4	Equal variances assumed	3.878	0.055	-1.145	44	0.258

At the Table 7.19B it is seen that there were seven items (GRAND4, CLOTHES1, CLOTHES2, MARY1, MARY2, SUN2 and SUN3) that scores on the PISA-RT were significantly higher than scores on the PISA-OT other and for the remaining five items there found no significant difference between groups. Hence, it can be concluded that the results support the hypothesis (2.ii) for the half of the items and there were no evidence for the remaining items.

Third research question was about investigating students' ideas on appropriateness of PISA stimuli with their learning experiences.

As it is mentioned in 7.2, during the course of analyzing results of the first phase of the study, the need to extend the study to which the views of the students on the some

negative entities which were not possible to revise for the new version of the items became apparent.

Based on the student ratings, each stimulus in science units received numerical indices in terms of familiarity on a scale from 1 to 5. Each text received an average value based on students' rating of the four items in the questionnaire (content in school and daily life, language, lay-out). Furthermore, a mean of all students' average ratings was calculated to form the overall values. The results are described below.

Table 7.20. Descriptive statistics related to SOS

Stimuli of Science Units	A school knowledge	*Rank*	B Relation to daily life	*Rank*	C Familiarity with language	*Rank*	D Familiarity with lay-out	*Rank*
GMC	2.83	6	3.30	2	2.93	3	2.07	8
ACID	3.46	2	2.37	7	2.86	4	2.86	3
GREEN	3.36	3	3.16	3	2.53	6	2.80	4
PHYSICAL	3.86	1	4.03	1	3.90	1	3.63	1
MARY	3.16	4	2.96	4	2.60	5	3.00	2
SUN	2.50	7	2.56	5.5	2.13	7.5	2.13	7
CLOTHES	2.90	5	2.56	5.5	2.13	7.5	2.36	6
GRAND	2.16	8	2.16	8	3.00	2	2.66	5
Overall	3.03		2.89		2.76		2.69	

Table 7.20 shows the means of the values for four items in the SOS and their rankings to see which science unit stimulus were most familiar in terms of their content, language and lay-out to students participating study. The rankings serve as a tool to see the stimuli at the top and bottom. The means on the table show that the students answered questionnaire found that the stimuli given at the beginning of the PISA science units to be moderately familiar to themselves. Familiarity ratings of science unit stimuli with the students' school content knowledge change between 2.16 (GRAND science unit stimuli) and 3.86 (PHYSICAL item group-stimuli). Familiarity means from the daily life

vary between 2.16 (GRAND science unit stimuli) and 4.03 (PHYSICAL item group-stimuli). For the third item of the questionnaire, students rate the familiarity to the language used in the stimuli of item groups. SUN and CLOTHES stimuli have the lowest mean with 2.13 and PHYSICAL science unit's stimulus has the highest value with 3.90. For the last item of the questionnaire, students found PHYSICAL stimulus to be more familiar in terms of lay-out with 3.63 value and GMC stimulus to get lowest mean value of 2.07.

The students who participated to the study found that the most familiar stimuli four all of the items in the questionnaire was PHYSICAL science unit stimulus which consists of one picture and one sentence emphasizing the importance of physical exercise for a healthy life. However, GRAND science unit stimulus was the least familiar one in terms of school content knowledge and daily life familiarity, which has a half page description of a canyon national park and a black and white picture from south side of the park. SUN and CLOTHES stimuli have the same rank and share the lowest value that shows the unfamiliarity with the language used in the passages. SUN stimulus includes one and half page description of an experiment designed to test the effectiveness of several sunscreen creams. CLOTHES stimulus is one page newspaper reading that is about an electro-textile product for blind children. In terms of lay-out, GMC stimuli got the lowest degree; it is part of an argumentation about using genetically modified crops that is one-third page long and additional explanation of the testing the counter ideas.

8. DISCUSSION AND CONCLUSION

This study is conducted in order to investigate three main goals and includes three phases with two predecided and one additional phase. The first goal is to examine construct validity of released PISA 2006 science units in terms of the positive and negative entities embedded within the stimuli and items of these science units. The construct validity was defined to be a qualitative property of test (e.g. Angoff, 1988; Cronbach & Quirk, 1976; Baykal, 2008) and examined through the entities in the present study. At the first phase of the study, positive and negative entities affecting the construct aimed to be defined in the science units, a form was developed and used to collect the teachers' reviews on stimuli and items of the science units. The second goal of the study is to investigate the effect of the negative entities through two comparison groups. The effect was measured in terms of the achievement scores of subjects. To be able to carry out the second phase of the study, reviews of teachers were collected and revisions were made according to the teachers' reviews at the end of the first phase of the study. Then effect of the negative entities was measured by examining achievement levels of the 15-year-old students who answered the original version of the items (PISA-OT test) and who answered revised version of the items (PISA-RT test). Achievement levels of the students were compared in terms of their scores on these two versions of tests. As a necessary extension of the study, a third phase was appended to the study in which a selected group of students were asked to rate the stimuli of the science units in terms of familiarity to their own learning experiences. The study presents both quantitative and qualitative data obtained from 80 teachers. Therefore eight science units were reviewed by 10 teachers per unit. Quantitative data was collected from the 60 students who are 15 year-old for the second phase of the study. Also, quantitative data collected form 30 students for third phase of the study.

The written comments produced by teachers on positive and negative entities of PISA science units were analyzed using the procedures of content analysis. Data obtained for the content analysis was collected through the two questions in the second part of a form which is called as Item Rating Form (IRF) and included both free comments of the teachers in response to the two questions. Also the descriptive statistics for three attributes of validity, necessity and importance for the each stimuli and item were collected and calculated with

the data coming from the third part of the IRF. After the revision of the items, PISA-RT test formed. Then, PISA-OT and PISA-RT were administered to gather data and to compare the effect of revision. To collect data about the students' views on the consistency of PISA stimuli with their unique learning experiences, Student Opinion Survey (SOS) was developed and administered.

The discussion of the first aim of the study will be based on the main categories evolved from the content analysis rather than including all of the sub categories. However, related literature and examples will be given when there is different sub category than it will be expected by the researcher. Lastly, the limitations and implementations of the study are presented.

Firstly, the study aims to determine positive and negative entities embedded in the PISA 2006 science units**.** From the teachers' writings positive entities embedded within the items affecting measuring and evaluating of a person science mindness, and negative entities embedded within the items affecting measuring and evaluating of person science mindness described and categorized. Additionally, teaches' ratings on the validity, necessity and importance of the items in terms evaluating of a person science mindness were calculated. Because the aim of the first phase is to detect and describe rather than testing some hypotheses, theoretical understanding was inductive within the qualitative approach.

The expert comments by the 80 teachers during the first phase of the study revealed 29 different negative entities relevant to the stimuli and items. These negative entities classified under the five different main categories. These were *language*, *typicality, presentation, structure* and *content*. Since the classification of the teachers comments were made without searching categories raised in the literature, there existed some consistencies and differences between the categories identified in the literature and in present study. It should be noted that in the literature the categories which they were similar to the negative entities or main categories in the present study described in two ways. Some studies were related with theoretical explanations of the construct irrelevance variance and construct under depression terms (e.g. Ferrera, 2007; Brebaum, 2007) and some others with the categories collected as the results of the empirical studies (e.g. Dudaite, 2006). Additionally, some of the studies referred in the literature based on the differences found in

translation and adaptation of the tests (e.g. McCreith, 2004; Olivery, 2007) while others from the investigation of properties of the items (e.g. Dempster & Reddy, 2007; Lemke, 1990; O'Halloran, 2000). Hence, the discussion will be based on all studies mentioned and will include examples from each.

In terms of the thematic units representing each negative entity, the most significant of these to the teachers were *content* and *language*. These negative entities were referred respectively 218 and 142 times in the teachers' comments on the PISA science units. The negative entity commented most often was *content* (see Table 7.1, p.67). In their comments about the *content* teachers referred to the things and ideas presented in science units. The *content* category included negative entities of irrelevance with national curriculum, irrelevance with topic, unnecessary context, wrong concept, cultural irrelevance, item-alternative inconsistency, different objective from program. Liberg et al. (2002) referred to *content* in a similar way. Besides, some of the negative entities were similar with the list of McCreith (2004). One of them was called as cultural irrelevance in the present study that McCreith described it as cultural relevance or cultural differences like a reading passage contains content to be more relevant to one of the groups taking the assessment. Culture also was emphasized in the study of Ferrera (2007) that it was referred to be one of important aspects of the target construct.

The *language* was also commented frequently. The *language* category covered expressions and words used in the science units. The negative entities were long sentence, difficulty in clause, difficulty in statement, difficulty in grammar, uncommon vocabulary (e.g. Turkish equivalence of the word ultraviolet). In their study of TIMSS 2003 focusing on South African students, Dempster and Reddy (2007) found similar sources for the underachievement of students. Additionally, some of the studies identified features that are specific to the science and mathematics items (Lemke, 1990; O'Halloran, 2000) and some of negative entities in the *language* category of the present study were similar to them. These studies showed that vocabulary, terminology, grammar and text structure were among these features creating obstacles for students. Besides, complexity of the text and connective markers categories mentioned by Liberg et al. (2002) were relevant to *language* category of present study.

The *presentation* and *structure* categories were commented relatively often, 102 and 88 times. One of the categories in the present study was *presentation* that referred to the elements used in the stimuli and items such as graphs, pictures, photos and lay-out of the science units. The negative entities of the category were lack of visual element, quality of visual element, inappropriate lay-out and arrangement unfamiliarity. Kress (2003) mentioned similar characteristics which cover forms of communication, pictures, tables and diagrams.

Comments on *structure* referred to features of the items like worse alternatives, multiple keyed answers, incompetent alternatives and irrelevant cues. Comments about the *typicality* refer to the quality of science units. The factors which disqualify the items are vague expectation, extreme easiness, expectation conflict and unfamiliarity with item stem. The unfamiliarity with item stem was also described to be negative entity under the typicality category that Ostelind (1998) mentions some disadvantages for real world problems and items to cause writing unambiguous items and maintaining a consistent grading standard. It was reported in some studies (Halayda, 1997; Dowling, 2006; Frey et al., 2007) that there are some basic rules to need to be obeyed to write good items in the tests such as using either the best answer or the correct answer format, avoiding cuing one item with another, avoiding window excessive verbiage in the stem, structuring an item so that the required response is concise. These rules are also mentioned by teachers in the present study, in the categories of *structure* and *language,* as negative entities included in the items.

From the studies present in the literature about the translation and adaptation of test, also Olivery (2007) mentioned some features of translated and adapted items that included in the language, structure and typicality categories of the present study. These properties making differences among groups can be summarized as changes in format including differences in punctuation, capitalization, item structure, typeface, and other formatting usages; omissions or additions including words, phrases, or expressions affecting the meaning of an item; differences in verb tense; differences in word difficulty, frequency or commonness of vocabulary; key words providing additional clues to guide examinees' thinking processes; differences in length or sentence complexity making the item more or less difficult and differences in words, expressions, or sentence structure inherent in one language or culture. However, there are differences in the way of using of these properties

with present study that Olivery (2007) described the differences between original and translated version of the items. In the present study these negative entities referred to be included in Turkish version of the items without comparing with original the English version. In a similar study, Ercikan (2002) mentioned about the insufficient translation to be able to raise difficulty levels of items with factors such as linguistic differences, curricular differences and cultural differences. Also, Ercikan et al. (2004) described four main points that they included to some extent in the present study such as familiarity of context and vocabulary is related with uncommon vocabulary seen in the present study as a sub category of language category; meaning, this can be seen in relation to combination of some sub categories such as grammar difficulty and expression; how the key information provided affect examinee's thinking processes, the sub categories of questioning style and vague expectation in the present study may be seen as threats to this.

In the present study, the most common negative entity category was the content and it can be concluded that this category was one of the most important ones affecting the construct aiming to be measured by PISA test.

In the first phase of the study, positive entities were also investigated (1.ii) by reviews of teachers. The same content analysis procedure was conducted to generate four main categories. These were classified as *context, content*, *composition* and *science processes*. In terms of the number of thematic units representing each category of the positive entities, the most common to the teachers was *composition*, *science process*, *content* and *context* respectively. These main categories referred at least 65 and at most 97 times by teachers. The most commented one was the *composition* category which included positive entities of using visual presentation; appropriate item stem, alternatives and language, item style, high cognitive level item and only one correct answer (see Table 7.3, p.72). As Haladyna (1997) mentions, clarity of the language, appropriate item stem, alternatives are among the necessary features to write valid items and teachers referred some items to be written appropriate according to these criteria. Interestingly, teachers perceive open ended items as a positive entity in the item style sub category. The reason behind this position depends on the abuse of multiple choice items in national central selection examinations and classroom assessments. As Güven (2001) reports, primary school Turkish teachers prefer to use mainly multiple choice items or traditional way of essay questions in their classes. Besides, in a study carried out by Dindar (2000) it is concluded that the most common item styles

used by the Turkish biology teachers were filled response, true-false and multiple choice in turn.

The category of *science process* as also cited frequently (80 thematic units) and the category contained comments about scientific investigation, describing a science event and using evidence to reach the answer. This category was unexpected because it was similar to the competencies described for each item in PISA study. As it is mentioned in the literature review (see part 2), PISA described three competencies and all of the items are classified as *explaining phenomena scientifically*, *using scientific evidence* and *identifying scientific issues* (see part 2.2.1). There were 25 items and eight stimuli in the present study and PISA defined competencies for each items, however teachers in the present study referred to the science processes for both items and stimuli. There were 19 items and two stimuli commented by the teachers to include science processes. 87 per cent of the teachers provided responses including worth of science processes defined in stimuli and items. Some of teachers' comments could be considered closer to competencies described by PISA. However some teachers referred to another science processes which are different from PISA or not included in the descriptions of PISA. This situation can be reasoned in relation to the claims of the Lau (2009) that there was found a difference between description of knowledge about science as in the PISA framework and its appearance in the sample items. In general, it may be concluded that the teachers participated in present study had perceptions of scientific processes considered relevant for scientific literacy in the PISA framework. The recognition of science processes by teachers may be seen as appealing since there are some studies referring the capability of Turkish teachers in terms of assessment techniques and skills on the measurement and evaluation. For example, Yeşil (2006) presents that Turkish social science teachers have limited competency about the measurement and evaluation processes in their classes. In another study, Tabak and Karakoç (2004) concluded similar results. Nevertheless, it is important to mention that in the present study teachers background knowledge may be accepted to be sufficient because 95 per cent of the teachers were graduated from the universities (e.g. Bogaziçi, ODTÜ) where the quality of teacher training education is compatible with the Western standards.

The categories of *content* and *context* were also commented upon. Comments on *content* category (73 thematic units) cumulated on the positive entities of interesting topic, familiarity with subject, consistency with objectives and relation to history of science.

Comments about *context* (65 thematic units) focused on two sub categories which they were real issue and relevance to possible situations. In relation to *context* category, it can be said that usage of texts related with the real situations and subjects/topics driven from the daily life experiences of students were positive values of the PISA science units. In a similar way, as a result of their study about the preservice-teachers' real world problems and their reactions to the students' solutions in mathematics, Verschaffel et al (1997) reported that teachers encounter similar problems in solving real- world problems but teachers believe in the value of real-world knowledge and realistic considerations when solving these type problems. Similarly, relation to the real world problems, Osterlind (1998) claims the alternative methods to affect test taker positively to build a response to a particular stimuli rather than recalling facts by using complex, real-world problems.

It can be inferred from the number of the thematic units referred by teachers for negative and positive entities that the number of negative entities (579 thematic units) was more than the number of positive entities (315 thematic units). From the point of entities, there were 29 negative entities and 17 positive entities formed from thematic units. These results can be interpreted in such a way that views of the teachers on the stimuli and items of the PISA 2006 science units tend to include more negative units than the positive ones in terms of numbers and diversity.

The third question (1.iii) searched in the first phase of the study aimed to investigate how Turkish version PISA 2006 science-units were valid, necessary and important in terms of the science mindness. Quantitative data for the first phase of the study came from the third part of the IRF given to the teachers who were asked to rate on the validity, necessity and importance of items on the three Likert-type nine point scales. That is to mean, teachers rate three properties (validity, necessity, importance) and they rate these over 9 points. The results presented in terms of the mean scores for each science unit and then for stimuli and items in the groups separately. It is noteworthy that seven of the eight science units received higher importance values than necessity and validity values, respectively. This result can be seen in relation to the category of *context* with positive entities of real issue and relevance to possible situation that there were 65 thematic units for these positive entities. The two science units (PHYSICAL and GREEN) had the average values converging to mid point. For the necessity criterion, the situation was similar to importance criterion; there were two science units (PHYSICAL and GREEN) above 5.00. Six of the science units received

average validity value below the 5.00 and one of the remaining two science units had 5.00 mean values while the highest value was 5.01. It can be important to mention that only two science units (PHYSICAL and GREEN) had the values over 5.00 for all of three criteria on a nine-point scale. The reasons for those two science units to get highest values roughly can be seen as the results of teachers' reviews. PHYSICAL science unit included positive entities of consistency with national curriculum and objectives in the program. For the GREEN science unit, the results may depend on the positive entities of real issue and science process which were frequently mentioned by teachers. In general view, based on the results for the science units, it can be concluded that in terms of necessity, importance and validity of the science units to measure and evaluate the presence and degree of science mindness (science attitude, ability, achievement etc.) of a person, teachers found that science units tend to have low validity, necessity and importance.

After all, since the main focus of the study on the items rather than being on the science units, the mean values for stimuli and items were calculated separately. So, one can see whether there will be any difference in the mean values and ranks. It is important to mention that the highest value for the validity of stimuli was 3.70, while it is 4.50 for necessity and 5.80 for importance. There were only three values higher than 5.00 and all was in ratings of the stimuli about importance. It is seen that the validity values for stimuli of SUN and CLOTHES groups were so low (both 1.80). This can be explained by results about the negative entities described by teachers. For the necessity and importance, the lowest value belongs to the PHYSICAL group. This can be connected the nature of stimuli mentioned by teachers, because the stimuli includes only the picture of two exercising people and one sentence explaining the need of exercising for a healthy life, most of the teachers had mentioned the stimuli to be unnecessary to answer the questions in the science units. Also, it can be inferred from this result that PHYSICAL science unit could have had highest scores for all of three criteria if its stimuli had designed similar to other stimuli (e.g. including information related with the items) and it would be not surprising because there were few negative entities described for this science unit.

The mean values of items in science units differentiated from values of their stimuli. The three of the science units, ACID, GREEN and PHYSICAL, were valuated to have scores over 5.00. It can be seen that while the stimulus of PHYSICAL science unit have lowest validity, its items gets the highest points. The items of these three groups also have

the highest points of the necessity and importance. As it is mentioned above the basis of the highest values for the items of PHYSICAL items can be examined in terms of the relevance of the subject to the national curriculum and connection with objectives of the science programs. On the contrary, items in CLOTHES group have the lowest validity, importance and necessity values. This can be understood in relation to the teachers' comments on the unit that they referred to negative entities like inappropriate questioning style, measuring different objective from the aimed. Additionally, the reason for items of CLOTHES having very low validity, necessity and importance values may be depend on effect of teachers' own evaluation of the stimulus.

In order to see whether there is a meaningful sequence in the choice of teachers about validity, necessity and importance of the science units, Friedman tests were carried out. Friedman tests results showed that teachers gave the significantly higher rates on the importance. It is to mean, teachers rate the science units to be important while they rate the validity of items to be low. The result is notable and consistent with analysis of the first and second research questions above. Almost all of the science units had referred to have lowest validity and necessity and a little higher values on importance.

The second aim of the study is to determine the effect of negative entities on the achievement scores of the students on the whole test and each item. The two tests, one including original Turkish version of items (PISA-OT), and the other consisting revised Turkish version of items (PISA-RT) were administered to two groups of students. It is noteworthy to remember that all of the negative entities defined for each item were not used in the revision process. That is to mean, only selected negative entities (by researcher and experts) included in the revision process. Based on the scores taken from whole test, a significant difference is found between the scores of students who were given the test including revised version of items and the test containing original Turkish version of items ($p=0.000$). It can be concluded that negative entities have a significant effect on the achievement scores of the students on the whole test. However, at the item level, significant difference is not found for all of the items. In the other words, there were not statistically significant differences between the scores of eleven revised and original items over 22 items. Furthermore, although there is no significant difference for the means of the all items, it is founded that the mean scores of the students in the comparison group which took the PISA-RT test were higher than the mean scores of comparison group which answered

the PISA-OT test at the item level. The reason for the results at the item level may depend on several factors. First of all, because all of the negative entities could not be covered in the revision process that these excluded ones could change the results. Secondly, negative entities such as irrelevance with national curriculum, unfamiliar subject, cultural unfamiliarity included mainly in the content category in the teachers' comments may lead to these results. For example, significant difference could not be found for any of the items in the GREEN science unit which had higher validity, necessity and importance scores in comparison to other science units at the first phase of the study as discussed above. However, the science unit was referred to have negative entities like irrelevance with national curriculum and irrelevance with topic by most of the teachers. Those may constitute base for the similar scores of the students in comparison groups. In a contrary situation, for one of the items in PHYSICAL science unit, it was not found significant difference between the scores of students in comparison groups. This science unit has again highest validity, necessity and importance values given by teachers. Also, it had few negative entities such as questioning style referred by teachers together with positive entities like consistency with national program, familiarity with subject and consistency with objectives that the reason. So, the significant difference may depend on the item properties rather than the curricular difference in some cases. It should also be noted that there was one item, second item of GREEN science unit, that there was no change made on the item by experts. There were changes on the science unit stimuli of the item but there was no significant difference between the comparison groups on this item. Hence, it can be concluded that some of differences may be because of the negative entities like curriculum, objective and topic relevance while some differences can be explained in terms of structure and typicality of the items.

The findings of the second phase of the study contradicted with the report of Olivery (2007) that items found to contain high level differential item functioning (contaminants in items) did not lead to performance differences between examinee groups for PISA 2003 problem solving area. However, at the item level the findings were parallel with the study of Dempster and Reddy (2007) that they found sentence complexity to be negatively correlated with the percent correct in the nine items (not all of items) where more than 40 per cent of the students chose the wrong concept. Furthermore, in a similar way, findings of Gipps and Murphey (1994) showed students' failure on tasks in science not because they made errors or they did not know, but because they tries to answer a completely different

question in their mind. As a specific example from the present study, MARY science unit can be shown, because the marker (sentence above the stimuli) asked to answer the questions according to the reading passage, it cannot be sure that students answer the right questions. Additionally, the results of the present study were consisted with Dudaite (2006) study in which item stem format and item answer format caused differences on the results of the TIMSS items that the TIMSS items had been changed in such a way.

Messick (1989) suggests that to gather data for the validity of the assessments processes underlying item response and task performance could be illuminated by asking students how they cope with items or tasks. In a similar way, the third aim of the study is to investigate the views of the 15 year-old students about the stimuli of the science units. Specifically, it is examined familiarity of the students with the PISA stimuli in terms of language, lay-out, school knowledge and daily life experiences. Students rated the eight stimuli for these four properties on a five point Likert-type form. The results showed that students' school experiences were most consistent with the PHYSICAL stimulus (M = 3.86) and less familiar with the GRAND stimulus (M=2.16). It may be inferred from these results that comments of students and teachers were consistent with each other. This can be exemplified with PHYSICAL stimulus that it had higher ratings form teachers while the grand group stimuli referred to be irrelevant with the national curriculum by nine of the ten teachers. From the point of daily life experiences, the ratings formed the similar ranks for PHYSICAL and GRAND stimuli. At that point, students' ratings with the GMC stimuli was interesting which had the second rank, this can depend on the accumulation of popular knowledge that the subjects of healthy life and organic nourishment was so actual at the period that study conducted.

From the results of first two questions at third phase of the study, it can be concluded that the views of the students show similarity with the comments of the teachers on the science units. It can be said that Turkish teachers and students do not tend to rate some of the stimuli (e.g. GRAND) as it is implied in the PISA context development. The language of the CLOTHES and SUN groups were found to be less familiar to the student that they had the same mean values of 2.13. These results are consistent with the comments of the teachers that two stimuli had lowest validity scores given by teachers. Some of the studies referred similar criteria like language and its components (Hambleton et al., 1999; Lemke, 1990; O'Hallonan, 2000).

The last criterion was the familiarity with lay-out of the stimuli. It is found that the students were most familiar with the PHYSICAL stimulus. However it can be because of the shortness of the stimuli when it is compared to the other stimuli. The second most familiar stimuli was that of MARY group which composed of only written stimuli of three paragraphs that other stimuli including graphs, pictures, photos had lower ratings. The result was also consistent with the findings of Dindar (2000) and Güven (2001) mentioned the properties of items use by the Turkish teaches not to include graphs, visual elements etc.

Based on the results of present study it can be cautiously concluded that Turkish version of the some of the released PISA 2006 science items tend to measure different than what they aim. It somewhat lacks validity in the national culture of Turkish 15 year-old students from the point of teachers and students participated in present study. This is reflected in the number of negative comments and low ratings for science units as evaluated by a group of teachers and also low ratings of the students in terms of familiarity with language, lay-out, school knowledge and daily life experiences. In particular, the low ratings of some science units reveal that they do not match with the Turkish students' unique learning experiences in their schools and in their lives. For example, GRAND group was almost unfamiliar to the students. In an international study such as PISA that is not based on the curriculum analysis, the question of the content coverage of the test is related with the national culture as a whole that includes national curriculum but not limited with it. As Hamilton and Barton (1999) focus, it is important to create the stimuli and items at the intersection points of all participating countries that country specific item can make international comparisons difficult.

This study illustrates the power of a blind item review process in detecting negative and positive entities embedded in the items of Turkish version of the PISA 2006 science units. The study also gives clues to item developers of such international assessments and translators/adaptors of the home country about the properties of items and the results of the processes.

It is the fact that large scale assessments such as PISA are used in order to inform curriculum, program development and evaluation and decisions concerning educational policies, and to make comparisons of student achievement across countries. Given this

picture, researchers and educators have much to learn about the weaknesses of home county students, e.g. in the present study Turkish students. It is noteworthy to mention about the nature of the items as 'translated' and translation makes the international studies more complex than it becomes. As Grisay (2003) claims, shortcomings in translation process can lead to the poor items and so some of the items can be more difficult for some countries. Also, Ruddock (2006) refers to the comparability of the constructs in the large scale assessments at both national and international levels that as stated in the present study high and unfamiliar reading demand of questions in PISA may lead to the lower demand in science required and lead to differences among the students answers in new contexts as opposed to very familiar ones. Since the present study deals with only translated version of the items, it may be said that the findings can belong to the both original and translated versions.

The present study encourages the recognition of the translation process of the international assessments with great attention and continuous research on the national findings as suggested by most of the studies (e.g. Hambleton, 1999; Ercikan, 2002; Simola, 2005; Golstein, 2004; Sireci, 1997).

The results of the study also encourage ongoing curriculum development studies in Turkey and instructional contexts. The new national framework curriculum has been prepared since 2008 aiming to constitute a reform action in (MEB, 2008) in Turkey that it provides opportunity to widen the conception of assessment styles and tools (e.g. pictures, photograph, written documents from different areas) used in the mother tongue.

8.1. Discussion of the Processes in Qualitative Part of the Study

In this section of the study, the validity of the overall study will be examined with a special focus on the qualitative part. As the study employs a mixed methodology in which qualitative and quantitative data analysis used in the same study (Teddlie and Tashakkori, 2003), the criteria for the evaluating the processes of the study used in both quantitative and qualitative approaches have to be taken into account. Since the aim of the study to discover and describe for the first phase of the study, the qualitative analysis was the dominant. In

the second phase of the study the quantitative analysis was present and instruments were discussed in terms of reliability and validity at parts 5.3.1 and 5.3.2.

The study can be evaluated in terms of its validity and appropriateness of the interpretations based on the results for the qualitative part. The Glaser and Staruss (1967) describes the trustwortness of the qualitative concepts to be extent to which one can believe the in the research findings. According to LeCompte and Goetz (1982) translatability is one of the important elements in the validity of the qualitative research that it means confidently making comparisons by clearly explaining the research methods, analytic categories, characteristics of group studied.

For the qualitative part of the study, criteria were defined as in the study of Miles and Huberman (1994, p.278). The criteria are;

- reliability /dependability/audiability
- objectivity /conformity
- credibility / authenticity/internal validity
- transferability/fittingness/external validity
- utilization /application/action orientation

In order to evaluate present study, these five criteria will be used. Miles and Huberman (1994) define the objectivity of the study to be connected with the relative neutrality and independence from researcher bias. These require transparency of the processes in the present study through the clear indication of the researcher's role in the study. It has been attempted in the present study to meet the demands of mentioned criteria by specific and detailed description and reasoning for each phase of the study. Firstly, the context of the study was introduced in the beginning of the study and research questions set out in the early phase of the the present study. Then, in the methodology section methods used in the study and reasons behind them made explained. Data gathering explained in detail, including the developments of the instruments, information about the sample. Furthermore, data analysis was described in detail by giving examples from the written comments of the teachers. From the point of role of researcher in the study, the researcher was aware of her prior knowledge and experience necessarily affecting the categories found in the content analysis and it is tried to reduce this interaction one more expert studied on

formation of the categories. At that point some critics may appear about the expertise of the teachers and students to rate the items in instruments used in the present study. It is important to mention that student and/or teacher reviews are frequently used to assess various aspects of the teaching learning process. From a theoretical perspective, both student and teacher report measures have face validity. However, some of the studies have reported the value of the reviews of students especially on the students' learning gains or motivational development (Kunter and Baumert, 2006). Form the point of teacher's ratings some of the researchers clarified that teachers' reviews have a predictive validity on the students' gains of the instructions. As Mayer (1999) and Porter (2002) emphasizes teachers have professional training and knowledge that these features lead them to be experts on various instructional approaches, methods and lesson features.

The criteria for the reliability have a role to clarify the consistency in the study (Miles & Huberman, 1994). As Creswell (2003) mentions, the qualitative and quantitative methods used together increase the validity and reliability of the study. In the present study, clarification of the research questions and choosing the appropriate methods in the light of the study aims contributed to the coherence between theory, research questions and methods used. As an example, teachers in the first phase of the study mentioned the texts and items not to be consistent with the students experience in either school or daily life and in the third phase of the study students were asked to evaluate the familiarity of the text from their own perspectives. There were practical reasons like time that teachers were given a tool to collect their comments rather than interviews. This process includes some limitations in it and these will be discussed in the next section. The requirement for the coding checks can be seen to include in the criterion of reliability that content analysis carried out in the study was repeated by another expert and the interviews were made with a small number of teachers to clarify the appropriateness of the codes with the teachers' writings.

Additionally, the credibility of the results of the present study has been strengthening by methodological triangulation. It has been argued that using mixed methods increases the validity of the study since by applying multiple methods to the same phenomenon a more complete understanding can be achieved. The indices used for validity, necessity and importance ratings of the teachers that comments of the teachers supported with numerical findings of their ratings on these three criteria. Additionally, teachers' comments on the

familiarity of the students with the content, language, lay-out of the science-units were asked students and the numerical findings were presented. These increased the authenticity of the study.

Transferability refers to the generalizability of the study findings. In the present study, because of the small number of the teaches commented on the science units and because of the selection type of the student sample as convenience sample without obeying the principle of probability sampling results cannot be generalized beyond the group of students attended to the private exam preparation center.

The utilization criteria for assessing the quality of the conclusions are related with the results of the study and conclusions forming basis to the further studies. The findings of the study can be applicable in the context of the forming more qualified items in designing national and international assessments. The requirement for more equity of testing is the validity issue that suggests items to measure the intended constructs, in an international assessment it is more critical to provide the validity for all of the participants that contributions on the adaptation of items for the unique learning experiences of the Turkish students will increase the utility of findings.

8.2. Limitations

This study was conducted under certain circumstances so that it includes some limitations. First of all, in the first phase of the study there were 80 teachers in total but only ten teachers reviewed the each science unit. Evaluating the conduct of the study afterwards shows that it would have been wise to collect more data and thus increase the number of respondents, i.e. reviewers per text, to the extent that more sophisticated quantitative methods of analysis could have been used in the study.

Another limitation for this study is related to the generalizability of the results of the study. For the first phase of the study, because of the limited numbers of released PISA items, the types and number of the entities cannot be generalized to the all PISA items. Additionally, for the second phase of the study, the conclusions of the study cannot be

generalized to all fifteen-year-old students. The sample size is at the limit value that this study is conducted with 60 students with 30 per in comparison groups in a private exam preparation center (dersane). Furthermore, the samples were not selected randomly. Findings are valid only for this sample. On general PISA study was a large scale study in terms of sample, purpose of study, cost and complexity of data analysis while the present study was a small scale one. Hence, present study includes all limitations coming from these differences.

Another limitation is due to the fact that investigating effect of negative entities focusing only on the answers from the writing comments of the teachers. The effect of the revision could be explored and clarified by conducting think-aloud procedures between students answered original Turkish version and revised versions of the items. Think-aloud procedures might have provided further information regarding whether the two student groups understood the items as they were intended to by researcher.

Another limitation can be found in the third phase of the study that it was an extent section to clarify the views of the teachers related with the familiarity of the students with stimuli and items in the science units. However, because of the practical reasons like number and length of the items, to keep students motivation high while they were filling scale, only the stimuli text were given to the students. This possibly caused to collect the limited data on the familiarity of the students on the science-units while teachers mentioned also the familiarity with the questioning style or item format found in the items rather than the stimuli.

Another limitation on the presenting teachers' comments and coding that usually the comments were so vague and short, member check process carried out with five teachers on the selected five items that is a small number to guarantee the coding process.

Moreover, the third part of the IRF enclose its limitation in it that teachers were asked to rate the validity, necessity and validity of the stimuli and items in terms of their contribution to develop students' science mindness. There was no additional explanation or description of these three constructs, they were left to the teachers' understandings and interpretation of the teachers.

One more limitation related with the third phase of the study that all revised negative entities are accepted to be one variable. This situation overshadows the individual effect of the negative entities on the items that some negative entities which have main effect can be overlooked or some which have little effect can be overused.

8.3. Recommendations for Further Research and Implications

International large-scale assessment data lead to important decisions in most countries concerning educational policies, comparisons of student achievement across demographics, regions, school types etc. Additionally, program development and evaluation have been affected from the results of such assessments. Particularly, reports from the PISA aim to provide indicators related to how well students are prepared for productive participation in future's world. Results of these analyses suggest that some of the countries are prepared well than the others to adapt the requirements of the future.

The most critical issues related with the any exam are accepted to be the validity and reliability. For international tests, the equity in testing also appears as a necessary condition that it also contribute valid comparisons of the results in an international context. On the test level, it is clear that only valid tests lead to meaningful and valid data. So that low performing countries may direct appropriate resource allocation and policy development to raise their standards or modify their systems. Hence, they can well prepare to meet the challenges of the future. For countries like Turkey who are performing near bottom, only valid and reliable data may provide information related to what are the holes in the education system and provide information to monitor the students to set up policies and programs that would allow them to maintain and continue to developing standards.

The present study includes a theoretical review on the subject of construct validity and provides information on the positive and negative entities related with the constructs found in the PISA 2006 science items from the point of Turkish science teachers. Based on these, the study provides new perspectives on the selection, formation and translation of the items for the science literacy tests.

Further research is recommended on the gender and school type differences to see if there is any review difference between teachers to find out the positive and negative entities or between teachers who work in public or private schools. Additionally, researches based on the gender and school type differences can be carried out with the students who answered the school, daily life, language and lay-out familiarity of PISA items.

Further research is also recommended on curricular and cultural differences found in the stimuli texts or item stems of the science units. Negative or positive entities described in terms of national curriculum and unique Turkish culture may include some cues to interpret or understand the results of the international assessments, particularly PISA study.

Although this study focuses on validity of science domain of PISA 2006, other test domains (such as mathematics and reading) and other large-scale assessments (such as TIMSS and PIRLS) may be investigated. Also, the same investigation process can be used to see whether there is any trend in terms of negative or positive entities addressed in the present study.

To summarize, validity and meaningfulness of scores should be based upon studies of construct, content, and cognitive parts. These studies may focus on one or more sources of validity evidence and include factors beyond test items such as culture, language, and the assessment context. Due to important decisions based upon the use of large-scale assessments, and significant amount of resources invested into developing, interpreting and using these measures; it is important that validity of these tests need to be met at highest standards. This study contributed towards addressing and examining these requirements at the item level from the point of teachers in the home country. Overall, the suggested researches would continue to contribute to more valid international tests in a broad perspective. More steps and efforts need to be invested in this area of research in the future.

APPENDIX A: LEVELS OF COMPETENCIES IN PISA

Level	**Lower score limit**	***What students can typically do***
6	707,9	At Level 6, students can consistently identify, explain and apply scientific knowledge and knowledge about science in a variety of complex life situations. They can link different information sours and explanations and use evidence from those sources to justify decisions. They clearly and consistently demonstrate advanced scientific thinking and reasoning, and they demonstrate willingness to use their scientific understanding in support of solutions to unfamiliar scientific and technological situations. Students at this level can use scientific knowledge and develop arguments in support of recommendations and decisions that centre on personal, social or global situations
5	633,3	At Level 5, students can identify the scientific components of many complex life situations, apply both scientific concepts and knowledge about science to these situations, and can compare, select and evaluate appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, link knowledge appropriately and bring critical insights to situations. They can construct explanations based on evidence and arguments based on their critical analysis.
4	558,7	At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions and they can communicate decisions using scientific knowledge and evidence
3	484,1	At Level 3, students can identify clearly described scientific issues in a range of contexts. They can select facts and knowledge to explain phenomena and apply simple models or inquiry strategies. Students at this level can interpret and use scientific concepts from different disciplines and can apply them directly. They can develop short statements using facts and make decisions based on scientific knowledge.
2	409,5	At Level 2, students have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They are capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem solving
1	334,9	At Level 1, students have such a limited scientific knowledge that it can only be applied to a few, familiar situations. They can present scientific explanations that are obvious and that follow explicitly from given evidence

APPENDIX B: TURKISH VERSION OF RELEASED QUESTIONS FROM PISA 2006

1. SORU KÜMESİ

GENETİK YAPILARI DEĞİŞTİRİLEN TARIM ÜRÜNLERİ

GENETİK YAPISI DEĞİŞTİRİLEN (GYD) MISIR YASAKLANMALIDIR

Doğayı koruma grupları, yeni ortaya çıkan genetik yapısı değiştirilmiş (GYD) mısırın yasaklanmasını istemektedirler.

GYD mısır, geleneksel mısır bitkilerini öldüren yeni ve güçlü bir zararlı ot ilacındanetkilenmeyecek şekilde geliştirilmiştir. Bu yeni zararlı ot ilacı, mısır tarlalarında kullanıldığında büyüyen zararlı otların pek çoğunu öldürecektir.

Doğayı koruma yanlısı olanlar, yeni ilacın öldüreceği zararlı otlar küçük

Yukarıdaki yazıda sözü edilen bilimsel incelemenin bazı ayrıntıları şunlardır:

- Mısır, ülkenin değişik yerlerindeki 200 tarlaya ekilmiştir.
- Her tarla önce iki eşit parçaya ayrılmıştır. Tarlanın bir parçasında yeni güçlü zararlı ot ilacı ile ilaçlanmış olan genetik yapısı değiştirilmiş (GYD) mısır yetiştirilmiştir. Tarlanın diğer parçasında da geleneksel zararlı ot ilacı ile ilaçlanmış geleneksel mısır yetiştirilmiştir.
- Yeni zararlı ot ilacı ile ilaçlanan GYD mısır içinde bulunan böceklerin sayısı, geleneksel zararlı ot ilacı ile ilaçlanmış olan geleneksel mısır içinde bulunan böceklerin sayısı ile hemen hemen aynıdır.

Soru 1:

Yukarıdaki yazıda sözü edilen bilimsel incelemede, hangi faktörler, bilinçli olarak değişikliğe uğratılmıştır? Her faktör için "Evet" ya da "Hayır" seçeneklerinden sadece birini yuvarlak içine alınız.

Bu faktör, incelemede bilinçli olarak değiştirilmiş midir?	**Evet yada Hayır?**
Çevredeki böcek sayısı	Evet / Hayır
Kullanılan zararlı ot ilacı türleri	Evet / Hayır

Soru 2:

Mısır ülkenin değişik yerlerindeki 200 tarlaya ekilmişti. Bilim adamları niçin birden fazla yerde ekim yapmışlardır?

A Yeni GYD mısırı, birçok çiftçinin deneme fırsatı bulması için

B Ne kadar GYD mısır yetiştirebileceklerini görmeleri için

C GYD mısır ekimini olabildiğince geniş bir alana yaymak için

D Mısırın değişik yetiştirme koşullarda nasıl büyüyeceğini görmek için

Soru:3

Tarlanın bir yarısına yeni ve güçlü bir zararlı ot ilacıyla ilaçlanan GYD mısır, tarlanın diğer yarısına da geleneksel zararlı ot ilacıyla ilaçlanan geleneksel mısır ekilmiştir.

Her bir ekim alanının iki yarıya ayrılarak bu şekilde kullanılması, çalışma sonuçlarının tarafsız olmasına nasıl bir katkıda bulunmuştur?

2. SORU KÜMESİ

ASİT YAĞMURU

Aşağıda, Caryatids adı verilen ve Atina Akropolünde 2500 yıl önce inşa edilmiş olan heykellerin fotoğrafı görülmektedir. Heykeller, mermer adı verilen bir cins kayadan yapılmıştır. Mermer kireçtaşından (kalsiyum karbonattan) oluşmaktadır.

Orijinal heykeller 1980 yılında kopyalarıyla değiştirilerek Akropol müzesinin içine alındı. Bu heykeller asit yağmurundan zarar görmüşlerdi.

Soru 1:

Normal yağmur, havadan bir miktar karbon dioksit emdiği için zayıf asit özelliği gösterir. Asit yağmuru, kükürt oksitler ve azot oksitler gibi gazları da emdiği için normal yağmura göre daha güçlü bir asit özelliği gösterir.

Havadaki kükürt oksitler ve azot oksitler nereden gelmektedir?
.....

Soru 2:

Asit yağmurunun mermer üzerindeki etkisi, bir gece boyunca mermer parçalarını sirke içine koyarak gösterilebilir. Sirke ve asit yağmuru yaklaşık aynı derecede asit özelliğine sahiptir. Mermer parçaları sirke içine bırakıldığında gaz kabarcıkları oluşur. Kuru mermer parçasının deneyden önce ve sonraki kütlesi bulunabilir

Bir mermer parçasının gece boyunca sirke içine konmadan önceki kütlesi 2,0 gramdır. Sonraki gün bu parça sirkeden çıkarılarak kurutulmuştur. Kurutulmuş olan bu mermer parçasının kütlesi ne kadar olabilir?

A 2,0 gramdan daha az
B Tam olarak 2,0 gram
C 2,0 ile 2,4 gram arasında
D 2,4 gramdan fazla

Soru 3:

Bu deneyi yapan öğrenciler mermer parçalarını bir gece boyunca saf (damıtılmış) su içerine bıraktılar.

Öğrencilerin, deneylerine bu işlemi de katmalarının nedeni nedir?
.....

3. SORU KÜMESİ

SERA

Okuma parçalarını okuyunuz ve ilgili soruları yanıtlayınız.

SERA ETKİSİ: GERÇEK Mİ YOKSA DÜŞSEL Mİ?

Canlılar yaşamak için enerjiye gereksinim duyarlar. Dünya üzerinde yaşamın devamını sağlayan enerji, çok sıcak olduğu için enerjisini uzaya yayan Güneş'ten gelir. Bu enerjinin çok küçük bir oranı Dünya'ya ulaşır.

Dünya'nın atmosferi, gezegenimizin üzerinde koruyucu bir örtü etkisi yaratır, havasız bir ortamda olabilecek sıcaklık değişimlerini engeller.

Güneş'ten gelen, ışınlar halinde yayılan enerjinin çoğu Dünya'nın atmosferinden geçer. Dünya bu enerjinin bir bölümünü emer, bir bölümü de Dünya yüzeyinden tekrar yansıtılır. Bu yansıtılan enerjinin bir bölümü atmosfer tarafından emilir.

Bunun sonucunda Dünya yüzeyi üstündeki ortalama sıcaklık, atmosferin yokluğu durumunda olabilecek sıcaklıktan daha yüksektir. Dünya'nın atmosferi bir sera ile aynı etkiye sahiptir, bundan dolayı *sera etkisi* terimi kullanılmaktadır.

Yirminci yüzyılda sera etkisinden daha çok bahsedildiği söylenmektedir.
Dünya atmosferinin ortalama sıcaklığının arttığı bir gerçektir. Karbon dioksit yayılımındaki artışın, yirminci yüzyıldaki sıcaklık artışının temel kaynağı olduğu gazete ve dergilerde sıklıkla söylenmektedir.

Soru 1:

Grafiklerde Ali'nin ulaştığı sonucu destekleyen nedir?
.....
.....

Soru 2:

Ceren adında başka bir öğrenci, Ali'nin varmış olduğu sonuca katılmamaktadır. O, iki grafiği karşılaştırır ve grafiğin bazı bölümlerinin Ali'nin sonucunu desteklemediğini söyler.
Grafiklerin, Ali'nin sonucunu desteklemeyen bölümlerine bir örnek veriniz. Yanıtınızı açıklayınız.
.....

Soru 3:

Ali, Dünya atmosferinin ortalama sıcaklığındaki artışın, karbon dioksit yayılımındaki artıştan kaynaklandığı konusunda vardığı sonuçlarda ısrar etmektedir. Ama Ceren, onun sonuca varması için henüz erken olduğunu düşünmektedir. Ceren, şöyle söylemektedir: "Bu sonucu kabul etmeden önce, sera etkisine neden olabilecek diğer etkenlerin sabit olduğundan emin olmalısın."

Ceren'in söylemek istediği etkenlerden birini belirtiniz.

4. SORU KÜMESİ

Soru 1:

Düzenli beden eğitimi hareketlerinin yararları nelerdir? Her ifade için "Evet" ya da "Hayır" seçeneklerinden sadece birini yuvarlak içine alınız.

Aşağıda verilenler düzenli beden eğitimi hareketlerinin sağlayacağı bir yarar mıdır?	**Evet ya da Hayır?**
Beden eğitimi hareketleri, kalp ve dolaşım hastalıklarından korunmaya yardımcı olur.	Evet / Hayır
Beden eğitimi hareketleri, sağlıklı bir beslenmeye götürür.	Evet / Hayır
Beden eğitimi hareketleri, fazla kilolardan korunmada yardımcı olur.	Evet / Hayır

Soru 2:

Kaslar çalıştırıldığı zaman ne olur? Her ifade için "Evet" ya da "Hayır" seçeneklerinden sadece birini yuvarlak içine alınız.

Kaslar çalıştırıldığında aşağıdaki olaylar gerçekleşir mi?	**Evet ya da Hayır?**
Kaslara gelen kan akışının artması	Evet / Hayır
Kaslarda yağların oluşması	Evet / Hayır

Soru 3.

Dinlenmedeki durumunuzla karşılaştırıldığında, beden eğitimi hareketleri yaparken daha sık nefes alıp verme zorunda olmanızın nedeni nedir?

.....

5. SORU KÜMESİ

MARY MONTAGU

Aşağıdaki gazete yazısını okuyunuz. Soruları bu yazıya göre yanıtlayınız.

AŞININ TARİHÇESİ

Mary Montagu güzel bir kadındı. 1715 yılında çiçek hastalığına yakalandı. Hastalığı geçirdi; fakat izleri kaldı. 1717 yılında Türkiye'de yaşarken, bu ülkede yaygınca kullanılmakta olan ve adına aşılama denen bir tedaviyi gördü. Bu tedavide sağlıklı gencin derisi çizilerek ona zayıflatılmış çiçek virüsü veriliyordu. Kişi kısa bir süre için hasta oluyor, ancak hastalığı genellikle çok hafif bir şekilde geçiyordu.

Mary, bu aşılama yönteminin güvenli olduğuna inandı ve kendi oğlu ile kızının da bu şekilde aşılanmasına izin verdi.

1796 yılında Edward Jenner çiçek hastalığına karşı antikor geliştirmek için insandaki çiçek hastalığı virüsünü değil, ineklerde görülen çiçek hastalığı virüsünü kullanarak aşılama yöntemini geliştirdi. Jenner'in bulduğu bu aşılama yönteminin, çiçek hastalığı virüsü verilmesine kıyasla, yan etkileri daha azdır ve tedavi gören kişi virüsü başka insanlara bulaştıramaz. Bu tedâvi biçimi aşılama adıyla tanındı.

Soru 1:

İnsanlar hangi çeşit hastalıklara karşı aşılanabilir?

A Hemofili gibi kalıtsal hastalıklar
B Çocuk felci gibi virüslerin neden olduğu hastalıklar
C Şeker hastalığı gibi vücudun işlevsel bozukluklarından kaynaklanan hastalıklar
D Tedavisi olmayan her çeşit hastalık

Soru 2:

Hayvanlar ya da insanlar bakterilerin neden olduğu bulaşıcı bir hastalığa yakalanır ve iyileşirse, hastalığa neden olan bakteriler genellikle onlarda tekrar hastalık oluşturamaz.

Bunun nedeni aşağıdakilerden hangisidir?

A Vücudun, aynı çeşitten bir hastalığa neden olabilecek bütün bakterileri öldürmüş olması
B Vücudun, bu tür bakterileri çoğalmadan önce öldürecek antikorlar yapmış olması
C Alyuvarların, aynı çeşit hastalığa neden olabilecek bütün bakterileri öldürmesi
D Alyuvarların, vücuttaki bu tip bakterileri yakalayarak vücuttan atması.

Soru 3:

Özellikle küçük çocuklar ve yaşlı insanların gribe karşı aşılanmaları önerilmektedir. Aşağıya bu öneri ile ilgili bir neden yazınız.

.....

6. SORU KÜMESİ

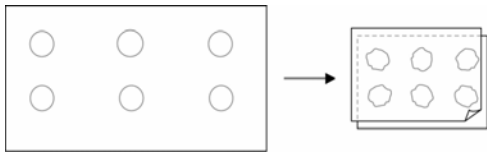
GÜNEŞTEN KORUYUCULAR

Jale ve Osman, güneşten koruma ürünlerinden hangisinin ciltleri için en iyi korumayı sağladığını merak ettiler. Güneşten koruma ürünleri için, her ürünün güneş ışığındaki ültraviyole ışınlarını ne derecede emdiğini gösteren bir *Güneşten Koruma Faktörü (GKF)* tanımlanmıştır. GKF'si yüksek olan bir güneşten koruyucu, GKF'si düşük olan bir güneşten koruyucuya göre cildi daha uzun süre korur.

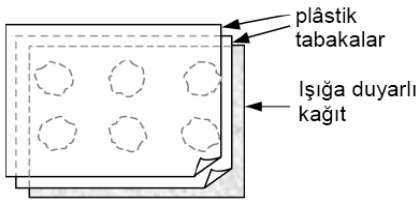
Jale, bazı güneşten koruma ürünlerini birbiriyle karşılaştırmak için bir yol düşündü. Osman ile birlikte aşağıdaki malzemeleri topladılar:

- güneş ışığını emmeyen (geçiren) iki temiz plastik tabaka;
- bir adet ışığa duyarlı kağıt;
- mineral yağ (M) ve çinko oksit (ZnO) içeren bir krem
- S1, S2, S3 ve S4 adını verdikleri dört farklı güneşten koruma ürünü.

Jale ve Osman, mineral yağı güneş ışınlarının çok büyük bir kısmını geçirdiği için, çinko oksidi de güneş ışınlarının tamamına yakınını geçirmediği için seçtiler. Osman, bir plastik tabaka üzerinde yuvarlak içine alınmış yerlerin her birine her maddeden birer damla koydu sonra bunların üzerini ikinci bir plastik tabaka ile kapattı. Bu plastik tabakaların üzerine büyük bir kitap yerleştirerek üstten iyice bastırdı.



Daha sonra,Jale hazırladıkları plastik tabakaları ışığa duyarlı kâğıdın üzerine koydu. Işığa duyarlı kâğıt, güneş ışığında tutulduğu süreye göre koyu griden beyaza (ya da çok açık griye) doğru renk değiştiren bir kâğıttır. En sonunda da, Osman hazırladıkları bu tabakaları güneşli bir yere koydu.



Soru 1:

Aşağıdaki ifadelerden hangisi, güneşten koruyucuların etkililiğini karşılaştırma amacıyla yapılan bir çalışmada mineral yağ ve çinko oksidin rolünün bilimsel tanımıdır?

A Mineral yağ ve çinko oksidin ikisi de etkisi araştırılan birer etkendir.
B Mineral yağ test edilen bir etken, çinko oksit ise karşılaştırma için kullanılan bir maddedir.
C Mineral yağ karşılaştırma için kullanılan bir madde, çinko oksit ise test edilen bir etkendir.
D Mineral yağ ve çinko oksidin ikisi de karşılaştırma için kullanılan birer maddedir.

Soru 2:

Jale ve Osman'ın yanıtlamaya çalıştığı soru aşağıdakilerden hangisidir?

A Güneşten koruyucu maddelerden her birinin koruma gücü diğerlerine kıyasla nasıldır?
B Güneşten koruyucular cildi ültraviyole ışınlarından nasıl korur?
C Mineral yağdan daha az koruma sağlayan bir güneşten koruyucu var mıdır?
D Çinko oksitten daha çok koruma sağlayan bir güneşten koruyucu var mıdır?

Soru 3:

İkinci plastik tabakanın üzerine neden iyice bastırılmıştır?

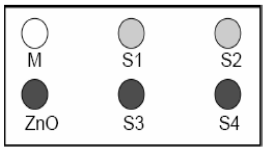
A Damlaların kurumasını önlemek için
B Damlaları mümkün olduğunca yaymak için
C Damlaları yuvarlaklar içinde tutmak için
D Damlalara eşit kalınlık vermek için

Soru 4:

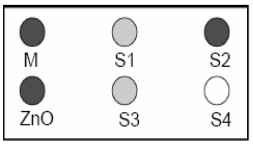
Işığa duyarlı kâğıt koyu gri renktedir; biraz güneş ışığında tutulduğu zaman açık gri renge dönüşür, güneş ışığında uzun süre tutulduğunda beyaz renk alır.

Aşağıdaki şekillerden hangisi elde edilebilecek sonucu göstermektedir? Neden bunu seçtiğinizi açıklayınız.

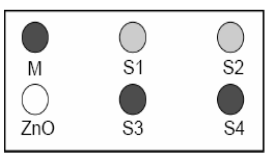
A



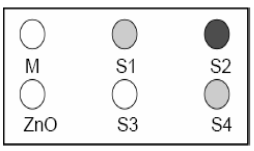
C



B



D



7. SORU KÜMESİ

GİYSİLER

Parçayı okuyunuz ve ilgili soruları yanıtlayınız.

GİYSİLERLE İLGİLİ BİR YAZI

Bir grup İngiliz bilim adamı, konuşma engelli çocuklara 'konuşma' gücü verecek 'akıllı' giysiler üretiyor. Benzeri olmayan bir elektro tekstil ürününden yapılan ve ses üreten bir aygıta bağlanmış yelek giyen çocuklar, dokunmaya duyarlı kumaşa hafifçe vurarak konuşmalarının başkaları tarafından anlaşılabilir duruma gelmesini sağlamaktadırlar.

Bu kumaş, normal kumaş ve içine kusursuz bir şekilde yerleştirilmiş karbon iplikçikler sayesinde elektriği iletebilen bir fileden yapılmıştır. Kumaş üzerine basınç uygulandığında, iletken iplikçiklerden geçen sinyaller değiştirilir ve bir bilgisayar devresi kumaşa nerede dokunulduğunu belirler. Daha sonra, bu devre kendisine bağlı olan ve iki kibrit kutusundan daha büyük olmayan bir elektronik aracın tetiklemektedir.

Bilim adamlarından birisi şöyle söylemektedir: "İşin en çarpıcı kısmı, kumaşı nasıl dokuduğumuz ve sinyalleri onun içinden nasıl gönderdiğimizdir - onu normal bir kumaşta var olan dokunuş şekli içerisine, kimsenin göremeyeceği şekilde yerleştirebiliriz."

Bu kumaş, zarar görmeksizin yıkanabilir, nesnelerin etrafına sarılabilir ya da sıkılıp top durumuna getirilebilir. Bilim adamları, onun toptan üretiminin ucuz olacağını da ileri sürmektedirler.

Kaynak: Steve Farrer, 'İç etkileşimli kumaş, kıyafetlerde malzeme hediyesi umudu uyandırıyor'(*'Interactive fabric promises a material gift of the garb'*), *Avustralya,* 10 Agustos 1998.

Soru 1:

Makalede ileri sürülen aşağıdaki savlar, laboratuardaki bilimsel araştırmalarla test edilebilir mi?

Her biri için "Evet" ya da "Hayır'ı" daire içine alınız.

Kumaş	**Sav, laboratuardaki bilimsel araştırmalarla test edilebilir mi?**
zarar görmeden yıkanabilir.	Evet / Hayır
zarar görmeden nesnelerin etrafına sarılabilir.	Evet / Hayır
zarar görmeden sıkılıp top biçimine getirilebilir.	Evet / Hayır
toptan üretimi ucuzdur.	Evet / Hayır

Soru 2:

Aşağıdaki laboratuar araçlarından hangisi kumaşın elektriği ilettiğini deneyebilmemiz için gerekecek araçlar arasında yer alabilir?

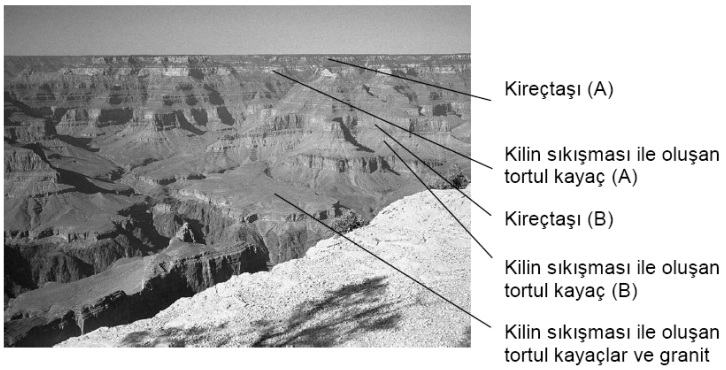
A Voltmetre
B Işık kutusu
C Mikrometre
D Ses ölçer

8. SORU KÜMESİ

GRAND KANYON (BÜYÜK KANYON)

Grand Canyon (Büyük Kanyon) Amerika Birleşik Devletleri'ndeki bir çöldedir. Burası, birçok kaya katmanını içeren çok geniş ve derin bir kanyondur. Geçmiş bir zaman diliminde yerkabuğunda meydana gelen hareketler bu katmanları yukarıya doğru itmiştir. Günümüzde bu kanyonun bazı bölümleri 1.6 km derinliğindedir. Kanyonun dibinde Colorado Nehri akmaktadır.

Aşağıda Büyük Kanyon' un güney kenarından çekilmiş bir resmi görülmektedir. Kanyon 'un bu resminde birkaç değişik kaya tabakası görülebilmektedir.



Soru 1:

Büyük Kanyon'u oluşturan nedir?
.....

Soru 2:

Büyük Kanyon millî parkını her yıl yaklaşık beş milyon dolayında insan ziyaret etmektedir. Bu kadar çok ziyaretçinin parka zarar vereceğinden kaygı duyulmaktadır.

Aşağıdaki sorular bilimsel araştırmayla yanıtlanabilir mi? Her soru için "Evet" ya da "Hayır" kutularından birini yuvarlak içine alınız.

Bu soru, bilimsel araştırma ile cevaplanabilir mi?	Evet ya da Hayır?
Yürüyüş yolları ne kadar toprak erozyona neden olmaktadır?	Evet / Hayır
Park alanı 100 yıl önce olduğu kadar güzel mi?	Evet / Hayır

Soru 3:

Büyük Kanyon' da hava sıcaklığı 0 ^{o}C 'ın altındaki sıcaklıklardan 40 ^{o}C'ın üstündeki sıcaklıklara kadar değişebilmektedir. Burası bir çöl alanı olmasına karşın, kayalardaki çatlaklarda bazen su bulunabilmektedir. Bu sıcaklık değişimleri ve çatlaklardaki su kayaların parçalanmasını nasıl hızlandırabilmektedir?

A Donan su, sıcak kayaları eritir.
B Su, kayaları birbirine yapıştırır.
C Buz kayaların yüzeyini düzleştirir.
D Kaya çatlaklarında donan su genleşir

Soru 4:

Büyük Kanyon'un "Kireçtaşı (A)" olarak belirtilen tabakasında deniztarağı, balık ve mercan gibi birçok deniz hayvanının fosilleri bulunmaktadır. Bu fosillerin orada bulunabilmeleri için milyonlarca yıl önce ne olmuştur?

A Eski zamanlarda insanlar okyanustan oraya su ürünleri getirmişlerdir.
B Bir zamanlar okyanuslarda büyük dalgalar oluştu ve bunlar deniz yaşamını karalara sürükledi.
C O zamanlarda okyanus buraları kaplamıştı, sonra sular eski yerine çekildi.
D Bazı deniz hayvanları, denize göç etmeden önce bir süre karada yaşadılar

APPENDIX C: ITEM RATING FORM

MADDE DERECELEME FORMU

Ad soyad: **Yaş:**

Okul (devlet/özel): **Tecrübe (yıl):**

Cinsiyet:

Değerli öğretmenimiz,

Ekteki sınav sorularını lütfen dikkatle okuyunuz. Kendinizi öğrencilerin yerine koyarak cevaplayınız. Bir insanın "***fen kafasını"*** (fen ilgisi, yeteneği, başarısı vb.) geliştirme hedefi bakımından sınav sorularının içeriği, ölçmek istediği yeterlilik, soruluş biçimi, sözel söylemi, görsel unsurları vb. ölçütler bakımından gözünüze çarpan özelliklerini belirtiniz. Gördüğünüz olumlu ve olumsuz nitelikleri anlaşılacak kadar belirtmeniz yeterlidir.
Kendinizi tam cümleler kurmak zorunda hissetmeyiniz.
Olumlu:

1) ……...

2) ……..

3) ……...

Olumsuz:

1) ……...

2) ……..

3) .....

Özetle, bu soru bir bütün olarak, öğrencilerde ***"fen kafasının"*** varlığını ve düzeyini ölçmek ve değerlendirmek bakımından ne kadar geçerli, gerekli ve önemlidir? Derecelendirmelerinizi aşağıdaki ölçeklerde 1 ile 9 arasında bir sayıyla belirtiniz.

Geçerli

Tümüyle Geçersiz								Tümüyle Geçerli
1	2	3	4	5	6	7	8	9

Gerekli

Tümüyle Gereksiz								Tümüyle Gerekli
1	2	3	4	5	6	7	8	9

Önemli

Tümüyle Önemsiz								Tümüyle Önemli
1	2	3	4	5	6	7	8	9

APPENDIX D: EXAMPLE OF THE CONTENT ANALYSIS OF THE DATA CONSISTING OF WRITTEN COMMENTS BY TEACHERS

Teacher number	**Science Unit (Stimuli/Item)**	**Written comment**	**Reduced thematic units**	**Sub-categories (negative/positive entity)**	**Final categories**
P41	Okuma parçasına konu olan Büyük Kanyon gerek okul programlarında gerekse öğrencilerin günlük yaşamlarında sıkça karşılaşamayacakları bir olgu, kanyon coğrafi şekli programda çok vurgulanmayan ve Türkiye coğrafyasında akdeniz bölgesinde rastlanan bir şekil. ek olarak, verilen resmin net olmadığı ve renkli ve daha büyük bir resim kullanmanın daha iyi olcağını söylenebilir	Okuma parçasına konu olan Büyük Kanyon gerek okul programlarında gerekse öğrencilerin günlük yaşamlarında sıkça karşılaşamayacakları bir olgu, kanyon coğrafi şekli programda çok vurgulanmayan ve Türkiye coğrafyasında akdeniz bölgesinde rastlanan bir şekil. ek olarak, verilen resmin net olmadığı ve renkli ve daha büyük bir resim kullanmanın daha iyi olcağını söylenebilir	…not in school program… ….not seen on daily life… …worse visual element….	*Negative* *Irrelevance with national curriculum* *Irrelevance with topic* *Inappropriate visual quality*	Content Content Presentation

APPENDIX E: POSITIVE ENTITIES FOR SCIENCE UNITS

The Appendix E includes the frequency distributions for remaining seven science units that one of them (GMC) presented at the part 7.

1. ACID RAIN (ACID) SCIENCE UNIT

A. STIMULUS

Acid Rain (ACID) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. Real issue-*context*	7	58.3	58.3	58.3
b. Relevance to possible situations-*context*	1	8.3	8.3	66.6
g. Visual element usage-*composition*	4	23.4	23.4	100.0
Total	12	100.0	100.0	

B. 1st ITEM

Acid Rain (ACID) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. Describing science event-*science process*	6	54.5	54.5	54.5
n. Item style-*composition*	5	45.5	45.5	100.0
Total	11	100.0	100.0	

C. 2nd ITEM

Acid Rain (ACID) 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
j. Clear language- *composition*	9	36.0	36.0	36.0
f. Consistency with program objectives- *Content*	5	20.0	20.0	56.0
i. Appropriate item stem- *composition*	4	16.0	16.0	72.0
m. Using evidence- *science process*	5	20.0	20.0	92.0
o. Appropriate alternatives- *composition*	2	8.0	8.0	100.0
Total	25	100.0	100.0	

D. 3rd ITEM

Acid Rain (ACID) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
n. item style-*composition*	4	40.0	40.0	40.0
l. Describing science event- *science process*	6	60.0	60.0	100.0
Total	10	100.0	100.0	

2. GREENHOUSE (GREEN) SCIENCE UNIT

A. STIMULUS

Green House (GREEN) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. Real issue- *context*	6	60.0	60.0	60.0
g. Visual element –*composition*	4	40.0	40.0	100.0
Total	10	100.0	100.0	

B. 1st ITEM

Green House (GREEN) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
m. using evidence - *science process*	7	38.8	38.8	38.8
n. item style-*composition*	7	38.8	38.8	77.6
f. consistency with program objectives –*content*	4	23.4	23.4	100.0
Total	18	100.0	100.0	

C. 2nd ITEM

Green House (GREEN) 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
h. Cognitive level- *composition*	5	62.5	62.5	62.5
m. Using evidence-*science process*	3	38.5	38.5	100.0
Total	8	100.0	100.0	

D. 3rd ITEM

Green House (GREEN) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. Real issue- *context*	4	100,0	100,0	100,0
Total	4	100,0	100,0	

3. GRAND CANYON (GRAND) SCIENCE UNIT

A. STIMULUS

Grand Canyon (GRAND) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
g. Visual element usage-composition	4	100.0	100.0	100.0
Total	4	100.0	100.0	

B. 1st ITEM

Grand Canyon (GRAND) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
-	0	0,0	0,0	100,0
Total	0	0,0	0,0	

C. 2nd ITEM

Grand Canyon (GRAND) 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. Describing science event-*science process*	4	100.0	100.0	100.0
Total	4	100.0	100.0	

D. 3rd ITEM

Grand Canyon (GRAND) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
i. appropriate item stem – *context*	5	25.0	25.0	25.0
a. real issue-*context*	6	30.0	30.0	55.0
j. appropriate language-*composition*	4	20.0	20.0	75.0
l. Describing science event-*science process*	5	25.0	25.0	100.0
Total	20	100.0	100.0	

E. 4rd ITEM

Grand Canyon (GRAND) 4rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. Describing science event-science process	5	62.5	62.5	62.5
p.only one correct answer- item structure	3	38.5	38.5	100.0
Total	8	100.0	100.0	

4. PHYSICAL EXERCISE (PHYSICAL) SCIENCE UNIT

A. STIMULUS

Physical Exercise (PHYSICAL) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
g. Visual element usage-*composition*	7	25.9	25.9	25.9
e. consistency with national program- *content*	8	29.6	29.6	55.5
a. real issue-context	7	25.9	25.9	81.4
d. familiarity with subject	5	18.6	18.6	100.0
Total	27	100.0	100.0	

B. 1st ITEM

Physical Exercise (PHYSICAL) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. Real issue-*context*	6	21.4	21.4	21.4
e. consistency with program-*content*	6	21.4	21.4	42.8
f. consistency with objectives-*content*	4	14.2	14.2	57.0
l. Describing science event-*science process*	5	17.8	17.8	74.8
d. familiarity with subject- *content*	7	25.2	25.2	100.0
Total	28	100.0	100.0	

C. 2nd ITEM

Physical Exercise (PHYSICAL) 2st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. describing science event- *science process*	5	21.7	21.7	21.7
i. appropriate item stem- *composition*	2	8.7	8.7	30.4
e. consistency with program- *content*	6	26.2	26.2	56.6
f. consistency with objectives- *content*	5	21.7	21.7	78.3
j. appropriate language- *composition*	5	21.7	21.7	100.0
Total	23	100.0	100.0	

D. 3rd ITEM

Physical Exercise (PHYSICAL) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
h. cognitive level- composition	5	34.8	34.8	34.8
m. Using evidence- science process	4	32.6	32.6	67.4
a. real issue- context	4	32.6	32.6	100.0
Total	13	100.0	100.0	

5. MARY MONTAGU (MARY) SCIENCE UNIT

A. STIMULUS

Mary Montagu(MARY) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
r. history of science- content	6	40.0	40.0	40.0
a. real issue	5	33.3	33.3	73.3
c. interesting topic- content	4	24.7	24.7	100.0
Total	15	100.0	100.0	

B. 1st ITEM

Mary Montagu (MARY) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. real issue-*context*	3	18.7	18.7	18.7
l. describing science event- *science process*	5	31.2	31.2	49.9
e. consistency with program-*content*	6	37.5	37.5	87.4
p. only one correct answer-*composition*	2	13.6	13.6	100.0
Total	16	100.0	100.0	

C. 2nd ITEM

Mary Montagu (MARY) 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
f. consistency with program objectives-*content*	6	46.1	46.1	46.1
j. appropriate language – *composition*	5	38.5	38.5	84.6
l. describing science event- *science process*	2	15.4	15.4	100.0
Total	13	100.0	100.0	

D. 3rd ITEM

Mary Montagu (MARY) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
n. item style – *composition*	4	33.3	33.3	33.3
a. Real issue-*context*	6	50.0	50.0	83.3
c. interesting subject- *content*	2	16.7	16.7	100.0
Total	12	100.0	100.0	

6. SUNSCREEN (SUN) SCIENCE UNIT

A. STIMULUS

Sun Screens(SUN) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
a. real issue-*context*	4	30.8	30.8	30.8
m. scientific investigation- *science process*	4	30.8	30.8	61.6
g. visual element usage-*composition*	3	23.1	23.1	84.7
c. interesting topic-*content*	2	15.3	15.3	100.0
Total	13	100.0	100.0	

B. 1st ITEM

Sun Screens(SUN) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. describing science event- *science process*	4	57.1	51.7	51.7
h. cognitive level- *composition*	3	42.9	42.9	100.0
Total	7	100.0	100.0	

C. 2nd ITEM

Sun Screens(SUN) 2nd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. describing science event- *science process*	6	100,0	100,0	100,0
Total	6	100,0	100,0	

D. 3rd ITEM

Sun Screens(SUN) 3rd Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. describing science event- *science process*	4	80.0	80.0	80.0
j. appropriate language-composition	1	20.0	20.0	100.0
Total	5	100.0	100.0	

E. 4rd ITEM

Sun Screens (SUN) 4th Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
h. cognitive level-*composition*	3	23.1	23.1	23.1
m. using evidence-*science process*	3	23.1	23.1	46.2
g. visual element usage-*composition*	4	30.8	30.8	77.0
n. item style- *composition*	3	23.0	23.0	100.0
Total	13	100.0	100.0	

7. CLOTHES (CLOTHES) SCIENCE UNIT

A. STIMULUS

Clothes (CLOTHES) Stimulus

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
-	0	0.00	0.00	100.0
Total	0	100.0	100.0	

B. 1st ITEM

Clothes (CLOTHES) 1st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
l. describing science event-science process	3	60.0	60.0	60.0
c. interesting subject- content	2	40.0	40.0	100.0
Total	5	100.0	100.0	100.0

C. 2nd ITEM

Clothes (CLOTHES) 2st Item

Positive entities	Frequency	Percent	Valid Percent	Cumulative Percent
-	0	0.00	0.00	100.0
Total	0	100.0	100.0	

APPENDIX F: REVISED VERSION OF THE ITEM GROUPS

1. CLOTHES SCIENCE UNIT

A. Stimulus: CLOTHES (Okuma Parçası: GİYSİLER)

GİYSİLER

Parçayı okuyunuz ve ilgili soruları yanıtlayınız.

GİYSİLERLE İLGİLİ BİR YAZI

Bir grup İngiliz bilim adamı, konuşma engelli çocuklara 'konuşma' *imkanı* verecek 'akıllı' giysiler üretir.

Çocuklar benzer*siz* bir **elektro tekstil** ürününden *dokunan* ve ses üreten bir aygıta bağlan*an bir* yelek giyer. Dokunmaya duyarlı kumaşa hafifçe vurunca, *söylemek istedikleri* başkaları tarafından anlaşılabilir duruma gelir.

Bu kumaş iki *malzemeden oluşur.Bunlardan biri normal* kumaş ,diğeri de içine kusursuz bir şekilde yerleştirilmiş karbon iplikçikler sayesinde elektriği iletebilen filedir. Kumaş üzerine basınç uygulandığında, iletken iplikçiklerden geçen sinyaller değişir. Aynı anda, bir bilgisayar devresi kumaşa nerede dokunulduğunu belirler. Daha sonra, bu devre kendisine bağlı *olan iki kibrit kutusu büyüklüğünde* bir elektronik aracı tetikle*r.*

Bir bilim adamı şöyle söylemektedir: "İşin en çarpıcı kısmı,

B. 1st ITEM

Yazıda ileri sürülen aşağıdaki iddialar, laboratuardaki bilimsel araştırmalarla test edilebilir mi?

Her biri için "Evet" yada "Hayır'ı" daire içine alınız.

Kumaş,	İddia, laboratuardaki bilimsel araştırmalarla test edilebilir mi?
zarar görmeden yıkanabilir.	Evet / Hayır
zarar görmeden nesnelerin etrafına sarılabilir	Evet / Hayır
zarar görmeden sıkılıp top biçimine getirilebilir	Evet / Hayır
toptan üretimi ucuzdur.	Evet / Hayır

C. 2nd ITEM

Aşağıdaki laboratuar araçlarından hangisi kumaşın elektriği ilettiğini *ölçebilmemiz* için gerekli araçlar arasında yer alır?

A. Voltmetre
B. Işık kutusu
C. Mikrometre
D. Ses ölçer

2. GENETICALLY MODIFIED CORPS (GMC) SCIENCE UNIT

A. Stimulus: GENETICALLY MODIFIED CORPS SHOULD BE BANNED (Okuma Parçası: GENETİK YAPISI DEĞİŞTİRİLEN (GYD) MISIR YASAKLANMALIDIR)

Genetically Modified Corps (GMC) Stimulus

Negative entity sub codes and codes	Frequency	Percent	Valid Percent	Cumulative Percent
a.Long Sentence -*Language*	8	29.6	29.6	29.6
b.Difficulty in clause-*Language*	6	22.2	22.2	51.9
c.Difficulty in statement-*Language*	5	18.5	18.5	70.4
d.Irrelevance with national curriculum- *Content*	5	18.5	18.5	88.9
q.Irrelevance with topic-*Content*	3	11.1	11.1	100.0
Total	27	100.0	100.0	

Above table shows the distribution of the 27 negative entities identified in the stimulus of GMC science unit. The most frequent negative entity was found to be *sentence length* which is followed by *difficulty in clause*. *Difficulty in statement* and *irrelevance with the national curricula* placed by five teachers, and the least frequent one was the *irrelevance with the* genetically modified corps topic. For the revision of the stimulus, it was decided not to be able to include the negative entities of *irrelevance with the national curriculum* and *irrelevance with the topic*. Therefore, revision of the stimuli based on the language category. Before detailed examination of the revision process applied, examples of the teachers' writings will be given.

Most of the teaches referred to the *long sentence length* used in the stimuli;

> Kurulan cümleler parçanın tamamında çok uzun. Özellikle, 'Doğayı koruma yanlısı olanlar, yeni ilacın öldüreceği zararlı otlar küçük hayvanların ve özellikle böceklerin beslenmesine yaradığından, bu yeni zararlı ot ilacının GYD mısır ile birlikte kullanılmasının çevre için kötü olacağını söylemektedirler'. cümlesinin anlaşılması birkaç defa okunması gerekiyor…the sentences are so long on the whole paragraph. Especially, the first sentence of the third paragraph needs to be read several times to have a clear understanding. (T 4)

> Özellikle üçüncü paragrafın giriş cümlesi birkaç parçaya bölünmeli…- In particular, the first sentence of the third paragraph should be divided into several sentences. (T6)

Difficulty in clause was found to be another negative entity by the teachers. Some of the teachers commented as in the below example;

> 'yeni ve güclü bir zararlı ot ilacı' , 'yeni ilacın öldüreceği zararlı otlar kücük hayvanların ve özellikle…' gibi ifadeler gereksiz yere zor…-(teacher mentions about the Turkish version of the clauses) clauses like powerful new herbicide, these weeds are feed for small animals, especially insects, the use of the new herbicide…are difficult unnecessarily. (T2)

> Mısır birkilerini öldüren yeni ve güçlü zararlı ot ilacı gibi cümleciklerin algılanması çok zor temel bilgi gözden kaçıyor..- it is so difficult to realize the sentences like powerful new herbicide that kills conventional corn plants (Turkish version of the sentence) that the main point is missing. (T9)

> Yeni güçlü zararlı ot tamlaması parçanın bütününde anlamayı zorlaştırıyor…- powerful new herbicide (Turkish version) makes the understanding difficult over the whole passage. (T1)

Teachers' comments on the *difficulty in statement* and *irrelevance with the national curriculum* reveal that half of the teachers classified these as factors make students understanding difficult.

> Öğrencilerin yazının tamamındaki kurguyu paragrafın akışı nedeniyle anlaması zor…-it is complicated for students to recognize the organization of the reading text because of the flow of the text. (T7)

> Parçadaki söylem yetersiz, akıştan olayın gidişatını takip etmek çok zor, daha sade ve akıcı hale getirilse güzel olur...- the expression in the reading isn't enough and it is difficult to follow what is going on.. It would be better to rewrite it to make more clear and fluent. (T10)

> Ayrıca bu gibi güncel konular yeni programda biraz ağırlık kazanmış durumda,suanda öğrencilerin görmediği bir konu bu..- such temporary subjects are emphasized more in the new national curriculum but at this time students have not been met with such a subject. (T5)

> Yeni programda da çok vurgulanmayan becerilere dayandığı için öğrenciler zorlanır henüz tam olarak da uygulanmıyor ayrıca, belirli olarak bu konuya rastlamadım müfredatta..- Students can have problem to understand because it is rarely related with the objectives of the curriculum. Additionally, I didn't experience this topic in curriculum. (T4)

The least frequent negative entity was the *irrelevance with the topic* that only three teachers mentioned about the students' familiarity with the genetically modified corps topic that provides no sensible reason for the reading the stimuli.

Genetik yapıları değiştirilmiş tarım ürünleri öğrencilerin çokça karşılaştıkları bir konu değil…- the genetically modified corps is not a customary topic for students. (T2)

Konu öğrencilere yabancı..- the topic is not typical for students. (T5)

Below table shows the eight types of negative entities found in the first item of the GMC science unit. The most frequent negative entities were *uncommon vocabulary* and *long sentence* which are belong to the category of language. Other entities were noticed by less than five teachers. Four teachers mentioned *alternatives to be worse*, three found the item to be *irrelevant with the national curriculum*. *Unnecessary context*, *difficulty in grammar* and *wrong information* were commented by two of the teachers. However, *wrong concept* entity criticized only by one teacher. There will be some examples from the comments of the teachers below.

Genetically Modified Corps (GMC) 1st Item

Negative Entity	Frequency	Percent	Valid Percent	Cumulative Percent
a.Long Sentence -*Language*	5	20.0	20.0	20.0
d.Irrelevance with national curriculum- *Content*	3	12.0	12.0	32.0
e.Unnecessary context-*Content*	2	8.0	8.0	40.0
f.Worse Alternative-*Structure*	4	16.0	16.0	56.0
g.Difficulty in grammar-*Language*	2	8.0	8.0	64.0
h.Uncommon vocabulary-*Language*	6	24.0	24.0	88.0
i.Wrong information-*Content*	2	8.0	8.0	96.0
j.Wrong concept-*Content*	1	4.0	4.0	100.0
Total	25	100.0	100.0	

It seems that more than half of the teachers referred to the *uncommon words* used in the item.

Faktör kelimesi yerine etken yada benzeri bir kelime kullanılabilir…It can be used ’etken’ instead of ’factor’. (T2)

Kullanılan zararlı ot ilacı türleri sadece ilac türleri denip paragrafın icinde acıklama yapılsa daha iyi olurdu..- I would be better to write ‘ilaç türleri’ instead of ‘kullanılan zararlı ot ilacı türleri’. (T4)

Faktör yerine daha öğrencilerin daha aşina olduğu bir terim kullanılmalı- another word that students are most familiar must be usedinstead of ‘faktör’. (T6)

Some teachers found that item had *worse alternative*, *wrong information,* and *difficulty in grammar* which effects what the question trying to ask.

> Çevredeki böcek sayısına dair bir bilgi verilmiyor parçada sadece mısırda böcek sayısı deniliyor- there is no information related with the number of insects on environment but in reading the number of the insects on crops had been given . (T10)

> Inceleme yerine araştırma faktör yerine bilesen ögrencilerin daha kolay anlayabilecegi ve daha cok karsılastıgı kelimeler ..-'araştırma' instead of 'inceleme' , 'bileşen' instead of 'faktör' are the words that students face more. (T7)

> Inceleme bir şeyi yada işi ele alma, gözden geçirme işidir oysa araştırma ise yöntemli çalışma yapmak demektir buradaki de araştırmadır..- 'inceleme' means the work of relooking while 'araştırma' means making methodological research and here the true word is 'araştırma'. (T4)

As it is shown from the below table, second item of the GMC science unit appears to have two negative entities of multiple true alternatives and incompetent item stem.

Genetically Modified Corps (GMC) 2nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
k.multiple true alternatives-*Structure*	8	61.5	61.5	61.5
l.incompetent item stem-*Structure*	5	38.5	38.5	100.0
Total	13	100.0	100.0	

Teachers recognized, as illustrated in the following quotes, the only two negative entities disturb the students' understanding of item and answering in a true way.

> Hangisi en iyi nedendir sorusu daha uygun, birden fazla cevabı var…- asking the question in the form of 'which is the best answer' is more appropriate; there are more than one answers for this question. (T2)

> Sorunun gercek ve tam cevabı seceneklerde verilmemis…- there is no exact and complete answer in the alternatives. (T6)

> Niçin soru kökü soru için pek uygun degil…- the question word of 'why' is not suitable for this question. (T7)

> C seçeneğin de nedenlerden biri..- C alternative is also the answer of the question. (T9)

Genetically Modified Corps (GMC) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d.Irrelevance with national curriculum- *Content*	3	15.8	15.8	15.8
g.Difficulty in grammar-*Language*	6	31.6	31.6	47.4
h.Uncommon vocabulary-*Language*	3	15.8	15.8	63.2
m.lack of visual element	2	10.5	10.5	73.7
n. vague expectancy	5	26.3	26.3	100.0
Total	19	100.0	100.0	

The third item of the GMC science unit found to include 19 themes in the five negative entity sub-categories. Difficulty in grammar was the most mentioned negative entity embedded in the item. Teachers exemplified as in the below quotes,

> Nasıl bir katkıda bulunmustur sorusu tam anlasılmıyor…- the question of how it contributes is not clear enough. (T4)

> Tarlanın bir yarısına yeni ve güçlü bir zararlı ot ilacıyla ilaçlanan GYD mısır, tarlanın diğer yarısına da geleneksel zararlı ot ilacıyla ilaçlanan geleneksel mısır ekilmiştir ifadesinin dilbilgisi kötü..- the grammar of the statement 'Tarlanın bir yarısına yeni ve güçlü bir zararlı ot ilacıyla ilaçlanan GYD mısır, tarlanın diğer yarısına da geleneksel zararlı ot ilacıyla ilaçlanan geleneksel mısır ekilmiştir' is of poor quality. (T7)

Teachers found ambiguity of expectation as another negative entity. Further, few of the teachers mentioned about the uncommon vocabulary, irrelevance with the national curriculum and lack of visual elements.

> Sadece yazılı ifadeden ziyade sorunun görsel sekilde, ikiye bölünmüş bir tarla cizimi gibi sorunun anlasılmasını kolaylastırır ..- It can be preferable to use a drawing of two parts devided area instead of using only writing explanation to make question more understandable. (T8)

> Ögrenciler ne tür bir cevap beklendigini anlamaz..- Students will not realize the which answer writer expected. (T3)

REVISED VERSION OF THE GMC SCIENCE UNIT

GENETİK YAPILARI DEĞİŞTİRİLEN TARIM ÜRÜNLERİ

GENETİK YAPISI DEĞİŞTİRİLEN (GYD) MISIR YASAKLANMALIDIR

Doğayı koruma grupları, yeni ortaya çıkan genetik yapısı değiştirilmiş (GYD) mısırın yasaklanmasını istemektedirler.

GYD mısır, geleneksel mısır bitkilerini öldüren yeni ve güçlü bir zararlı ot ilacından etkilenmeyecek şekilde geliştirilmiştir. Bu yeni zararlı ot ilacı, mısır tarlalarında kullanıldığında büyüyen zararlı otların pek çoğunu öldürecektir.

Doğayı koruma yanlısı olanlar, yeni ilacın öldüreceği zararlı otlar küçük hayvanların ve özellikle böceklerin beslenmesine yaradığından, bu yeni zararlı ot ilacının GYD mısır ile birlikte kullanılmasının çevre için kötü olacağını söylemektedirler. GYD mısırın kullanılmasını destekleyenler buna cevap olarak bilimsel bir *araştırmanın,* sonucun bu şekilde olmayacağını gösterdiğini söylemektedirler

Yukarıdaki yazıda sözü edilen bilimsel *araştırmanın* bazı ayrıntıları şunlardır:

- Mısır, ülkenin *farklı* yerlerindeki 200 tarlaya ekilmiştir.
- Her tarla önce iki eşit parçaya ayrılmıştır. Tarlanın bir parçasında yeni güçlü zararlı ot ilacı ile ilaçlanmış olan genetik yapısı değiştirilmiş (GYD) mısır yetiştirilmiştir. Tarlanın diğer parçasında da geleneksel zararlı ot ilacı ile ilaçlanmış geleneksel mısır yetiştirilmiştir.
- Yeni zararlı ot ilacı ile ilaçlanan GYD mısır içinde bulunan böceklerin sayısı, geleneksel zararlı ot ilacı ile ilaçlanmış olan geleneksel mısır içinde bulunan böceklerin sayısı ile hemen hemen aynıdır.

A. 1st ITEM

Yazıda sözü edilen bilimsel *araştırmada*, aşağıdaki etkenler, *araştırmacılar tarafından* değişikliğe uğratılmış mıdır?

Her faktör için "Evet" yada "Hayır" seçeneklerinden sadece birini yuvarlak içine alınız.

Çevredeki böcek sayısı	Evet / Hayır
Kullanılan "zararlı ot ilacı" türleri	Evet / Hayır

B. 2nd ITEM

Mısır ülkenin değişik yerlerindeki 200 *tarlaya ekilmesinin en önemli nedeni aşağıdakilerden hangisidir?*

A Birçok çiftçiye GDY mısırı deneme fırsatı vermek
B Ne kadar GYD mısır *yetiştirilebileceğini görmek*
C GYD mısır ekimini olabildiğince geniş bir alana yaymak
D Mısırın değişik yetiş*me* koşullarda nasıl büyüyeceğini görmek

C. 3rd ITEM

Tarlanın bir yarısına *yeni "zararlı ot ilacı"* ile ilaçlanan GYD mısır, tarlanın diğer yarısına da geleneksel "*zararlı ot ilacı"* ile ilaçlanan geleneksel mısır ekilmiştir.

Her bir *tarlanın eşit iki parçaya* ayrılarak kullanılması, çalışma sonuçlarının tarafsız olmasına nasıl bir katkıda bulunmuştur?

3. ACID RAIN (ACID) SCIENCE UNIT

A. Stimuli: ACID RAIN
(Okuma Parçası: ASİT YAĞMURU)

Acid Rain (ACID) Stimulus

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d. Irrelevance with national curriculum- *Content*	7	20.6	20.6	20.6
e.Unnecessary context-*Content*	4	11.8	11.8	32.4
g.Difficulty in grammar-*Language*	2	5.9	5.9	38.2
h.Uncommon vocabulary-*Language*	5	14.7	14.7	52.9
j.Wrong concept-*Content*	6	17.6	17.6	70.6
o.quality of visual stimuli-*Presentation*	4	11.8	11.8	82.4
p. Cultural irrelevance- *Content*	2	5.9	5.9	88.2
q.Irrelevance with topic-*Content*	2	5.9	5.9	94.1
z. inappropriate lay-out-*Presentation*	2	5.9	5.9	100.0
Total	28	100.0	100.0	

Based on the teachers' comments above, it seems that there were 28 negative entities which are the members of nine different sub-categories and three different categories. *Irrelevance with the national curriculum* was mentioned by the seven teachers. The *wrong concept* was the second one written in the comments of the teachers. *Uncommon vocabulary* was referred by five teachers and it was followed by u*nnecessary context* and *quality of visual stimulus.* The negative entities of *difficulty in grammar*, *cultural irrelevance, irrelevance with topic and inappropriate lay-out* were remarked by two teachers.

The following are the example sentences that teachers' stated to be found in the stimulus of ACID science unit. Due to the examples belonging to most of the sub-groups were given in CLOTHES and GMC science units, only examples for the other sub-groups will be given below.

Four of the teachers commented on the quality of visual elements that stimulus of ACID science unit includes a picture of statuses damaged by acid rain.

Resimde verilen heykellerin görüntüsü çok uzaktan..-the picture of the status given in the stimulus is taken from so far. (T12)

Fotoğraf net değil ve renkli olması daha etkili olabilir…-the photograph is not clear enough. (T15)

Another negative entity was *cultural irrelevance* that example teacher comment had not been given before.

Heykel kavram olarak öğrencilere çok yakın değil özellikle Türk kültüründe mermer heykel ve asit yağmuru çok ihtiva edilmiyor….- status as a concept is not a familiar term for the students , especially in Turkish culture marble status and acid rain is not included in so much. (T14)

Acid Rain (ACID) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
g.Difficulty in grammar-*Language*	1	10.0	10.0	10.0
q.Irrelevance with topic-*Content*	1	10.0	10.0	20.0
r. questioning style –*Presentation*	8	80.0	80.0	100.0
Total	10	100.0	100.0	

There were three sub-groups of negative entities found in the first item of the ACID science unit. Eight of the teachers commented on the negative entity of questioning style and irrelevance with topic and difficulty in grammar mention by one teacher. One example from the writings of teachers related with the questioning style was given below.

Soru asit yagmurlarına neden olan gazların kaynagı nelerdir şeklinde sorulabilir…- the question can be asked as what are the sources of gases that cause the acid rain. (T20)

The revision of the item based on the changing the item stem and questioning style to make more a clear question.

Acid Rain (ACID) 2nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
c. Difficulty in statement-*Language*	10	100.0	100.0	100.0
Total	10	100.0	100.0	

There was only one negative entity found in the second item of the acid rain science unit that teachers commented on the *difficulty in statement* to be present. As one of the teachers reviewed;

> Sadece de sorunun üzerinde verilen sirke-mermer deneyinde daha etkili bir dil kullanılabilir..- only I can say about the question, the statements in the short paragraph can be clearer. (T11)

The item revised to make the expression more clearly by changing the wording of explanation and item stem.

Acid Rain (ACID) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
g.Difficulty in grammar-*Language*	1	10.0	10.0	10.0
n. vague expectancy	9	90.0	90.0	100.0
Total	10	100.0	100.0	

There were two types of negative entities in the third item of the ACID science unit. The *vague expectancy* was mentioned by nine of the teachers. Only one teacher wrote *difficulty in grammar* to be found in the item. One teacher exemplified the *vague expectancy* as in the below;

> Soruda verilen bilgilerin diğerleri ile ilişkisi açık değil, fazlası ile kapalı bir soru öğrencinin soruyu algılama şekline göre amaçtan uzak cevaplar gelebilir…-the relation between the information given in question and others is not clear, the question is highly flu that answers of the students can change according to their own understandings. (T16)

REVISED VERSION OF THE ACID SCIENCE UNIT

A. Stimulus: ACID RAIN

(Okuma Parçası: Asit Yağmuru)

ASİT YAĞMURU

Aşağıda, Caryatids adı verilen ve Atina *Kalesi'nde* 2500 yıl önce *yapılan* heykellerin fotoğrafı görülmektedir. Heykeller, mermer adı verilen ve kireçtaşından (kalsiyum karbonattan) oluşan bir *taş türünden* yapılmıştır.

Heykeller asit yağmurundan zarar görmüş ve orijinal heykeller 1980 yılında kopyalarıyla değiştirilerek Akropol müzesinin içine

B. 1st ITEM

Normal yağmur, havadan bir miktar karbon dioksit *ile birleşerek* zayıf asit özelliği gösterir.

Asit yağmuru, kükürt oksitler ve azot oksitler gibi gazlar*la da birleştiği* için normal yağmura göre daha güçlü asit özelliği gösterir

C. 2nd ITEM

Sirke ve asit yağmuru **yaklaşık aynı derecede** asit özelliğine sahiptir.
Asit yağmurunun mermer üzerindeki etkisi, mermer parça*sını bir gece boyunca* sirke *içinde bekletilerek* gösterilebilir. Mermer parç*ası* sirke içine bırakıldığında gaz kabarcıkları oluşur. Kuru mermer parçasının deneyden önce ve sonraki kütlesi bulunabilir

Bir mermer parçasının gece boyunca sirke içine konmadan önceki kütlesi 2,0 gramdır. Sonraki gün bu parça sirkeden çıkarılarak kurutulmuştur. Kurutulmuş olan bu mermer parçasının kütlesi ne kadar olabilir?

A 2,0 gramdan daha az
B Tam olarak 2,0 gram
C 2,0 ile 2,4 gram arasında
D 2,4 gramdan fazla

D. 3rd ITEM

Bu deneyi yapan öğrenciler **farklı** mermer parçalarını **da** bir gece boyunca saf (damıtılmış) su içerine bıraktılar

Öğrencilerin, deneylerine bu işlemi de katmalarının nedeni nedir?

.....

.....

4. GREENHOUSE (GREEN) SCIENCE UNIT

A. Stimuli: THE GREENHOUSE EFFECT: FACT OR FICTION?

(Okuma Parçası: SERA ETKİSİ: GERÇEK Mİ YOKSA DÜŞSEL Mİ?)

Green House (GREEN) Stimuli

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
c.Difficulty in statement-*Language*	6	16.7	16.7	16.7
d. Irrelevance with national curriculum- *Content*	4	11.1	11.1	27.8
g.Difficulty in grammar-*Language*	4	11.1	11.1	38.9
i.Wrong information-*Content*	2	5.6	5.6	44.4
o.quality of visual stimuli-*Presentation*	4	11.1	11.1	55.6
q.Irrelevance with topic-*Content*	7	19.4	19.4	75.0
t. arrangement unfamiliarity-*Presentation*	4	11.1	11.1	86.1
z. inappropriate lay-out-*Presentation*	5	13.9	13.9	100.0
Total	36	100.0	100.0	

In the stimulus of GREEN item group, there were found totally 36 negative entities that classified in language, content and presentation categories. The most frequent referred negative entities were *irrelevance with topic, difficulty in statement* and *inappropriate lay-out.* The remaining four negative entities found by four teachers and *wrong information* was the least referred entity. Examples of three teachers in the *irrelevance with topic, quality of visual stimuli* and *inappropriate lay-out* will be given below in sequence.

> Ek olarak sera etkisi öğrencilerin küresel ısınma konusu altında az süre ile öğrendikleri bir konu…- additionally, greenhouse effect is a subject that students learn little about it under the topic of global warming. (T21)

> Grafikler belirsiz ve bu öğrencilerin grafikleri okumasını zorlaştırır…-the graphs are ambiguous and this makes more difficult students to be able to read the graphs. (T25)

> Konunun sunumu ve sayfa içindeki duruşu değiştirilmeli..- the presentation and position of the written description of the subject must be changed. (T27)

The revision of the stimuli of GREEN item group included changes of the statements, clases, omitting some clauses and word to make reading of passage more

fluent. The information of the 'sun transmitting its energy because of its temperature' was completely changed.

Green House (GREEN) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
n.vague expectancy- *Typicality*	2	18,2	18,2	18,2
r. questioning style –*Presentation*	9	81,8	81,8	100,0
Total	11	100,0	100,0	

The first item of the GREEN science unit included two types of the negative entities which were *vague expectancy* and *questioning style*. One example from the teachers' writings for *vague expectancy* is as below;

> Soruda istenen aslında öğrencinin grafiklerden bir örnek vermesidir fakat bu durum yeterince açık değil…- mainly, students are needed to give examples by using the graphs but this is not clear. (T24)

Another example for the *questioning style* is presented below;

> Bu tür bir açık uçlu soruda Ali'nin ulaştığı sonucu soru cümlesinden hemen önce vermek gerekir, diğer türlü öğrencinin parçada ali'nin sonucunu bulması için parçaya tekrar dönmesi gerekir ki bu da zaman harcamasına neden olur ve soruyu cevapsız bırakabilir…- ın a such open ended question, it is necessary to give the results that Ali found just before the item stem, otherwise students will have to turn the reading passage to remember the results Ali found. As a conclusion this situation can cause time consuming and students can pass answer without answering. (T29)

Green House (GREEN) 2nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
e.Unnecessary context-*Content*	3	30.0	30.0	30.0
g.Difficulty in grammar-*Language*	7	70.0	70.0	100.0
Total	10	100.0	100.0	

There were two types of negative entities found in the second item of the GREEN item group which were *unnecessary context* and *difficulty in grammar.* One example given for the was *difficulty in grammar* as below;

> Daha düzgün Türkçe dilbilgisi kullanılabilir…- there can be better Turkish grammar usage. (T22)

Additionally, another example for the *unnecessary context* was like;

> Bu sorunun cevaplanması için sadece grafiklerin verilmesi yeterli..-Only graphs are enough to answer these questions. (T27)

Green House (GREEN) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
c.Difficulty in statement-*Language*	2	12.5	12.5	12.5
d. Irrelevance with national curriculum- *Content*	9	56.3	56.3	68.8
e.Unnecessary context-*Content*	3	18.8	18.8	87.5
s. item-alternative inconsistency – *Content*	2	12.5	12.5	100.0
Total	16	100.0	100.0	

Teachers referred to four negative entities to be found in the third item of the GREEN science unit. There will be an example of teachers' comment about the *irrelevance with national curriculum;*

> Sera etkisinin nedenleri öğrenciler için zor bir soru çünkü değişen programla beraber üzerinde durulmaya ve tarışılmaya başlanıldı- the question of reasons for the greenhouse effect is a difficult question for students because the subject has been begun to discuss with the changing curriculum. (T26)

REVISED VERSION OF THE GREEN SCIENCE UNIT

A. STIMULUS

SERA

Okuma parçalarını okuyunuz ve ilgili soruları yanıtlayınız

SERA ETKİSİ: GERÇEK Mİ YOKSA DÜŞSEL Mİ?

Canlıların yaşamak *için enerjiye ihtiyaçları vardır*.Dünya üzerinde yaşamın devamını sağlayan enerji, Güneş'ten gelir. Bu enerjinin çok küçük bir oranı Dünya'ya ulaşır.

Dünya'nın atmosferi, gezegenimizin üzerinde koruyucu bir örtü etkisi yaratır, havasız bir ortamda olabilecek sıcaklık değişimlerini engeller.

Güneş'ten gelen, ışınlar halinde yayılan enerjinin çoğu Dünya'nın atmosferinden geçer. Dünya bu enerjinin bir bölümünü emer, bir bölümü de Dünya yüzeyinden

yansıtılır. Yansıtılan **bu** enerjinin bir bölümü atmosfer tarafından emilir.

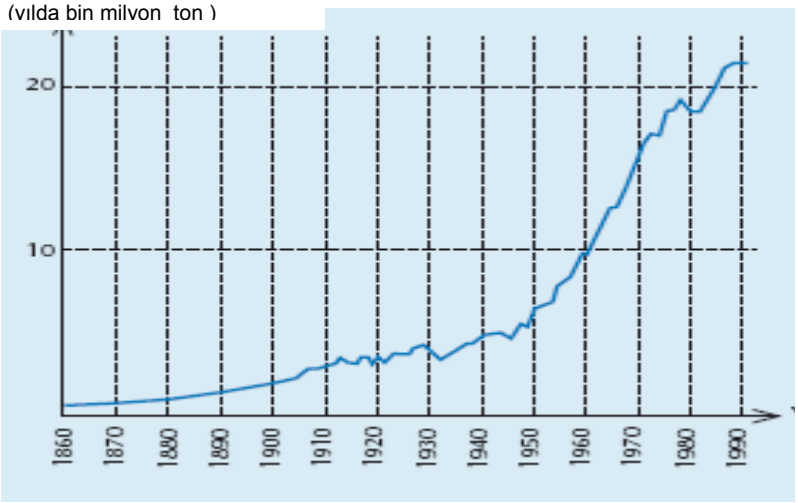
Bunun sonucunda Dünya yüzeyi üstündeki ortalama sıcaklık, atmosferin yokluğu durumunda olabilecek sıcaklıktan daha yüksektir. Dünya atmosferi bir sera ile aynı etkiye sahiptir, bu yüzden *sera etkisi* terimi kullanılmaktadır.

Yirminci yüzyılda sera etkisinden daha çok bahsedilmektedir Dünya atmosferinin ortalama sıcaklığının arttığı bir gerçektir. Atmosfere bırakılan Karbon dioksit miktarındaki artışın, yirminci yüzyıldaki sıcaklık artışının temel kaynağı olduğu gazete ve dergilerde sıkça söylenmektedir.

Ali adında bir öğrenci, Dünya atmosferinin ortalama sıcaklığı ile Dünya üzerinden atmosfere bırakılan karbon dioksit miktarındaki artış arasında bir ilişki kurar.
O, bir kitaplıkta aşağıdaki iki grafiğe rastlar.

Karbon dioksit yayılımı

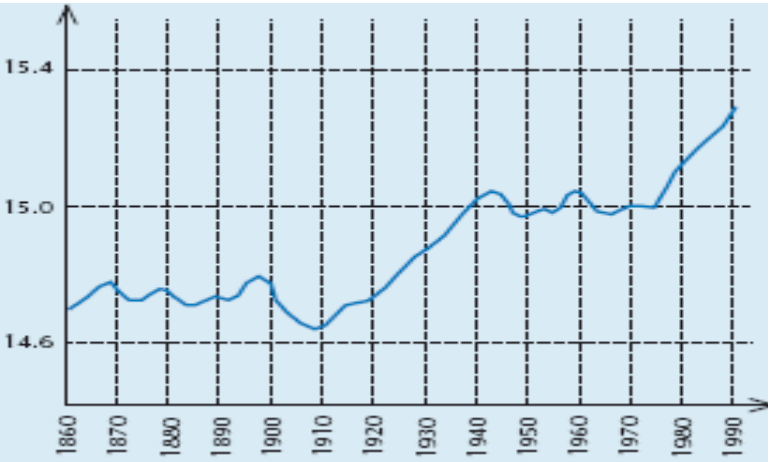
(yılda bin milyon ton)



Yıllar

Dünya atmosferinin

ortalama sıcaklığı



B. 1st ITEM

İki grafiği karşılaştırarak Ali'nin ulaştığı sonucu destekleyen kısımlara örnek veriniz.

.....
.....

C. 2nd ITEM

Ceren adında bir öğrenci, Ali'nin varmış olduğu sonuca katılmamaktadır.

Ceren, iki grafiği karşılaştırır ve grafiğin bazı bölümlerinin Ali'nin sonucunu desteklemediğini söyler.

Ceren, Ali'nin sonuca varması için henüz erken olduğunu düşünmektedir.

Ceren, şöyle söylemektedir: "Bu sonucu kabul etmeden önce, sera etkisine neden olabilecek diğer etkenlerin sabit olduğundan emin olmalısın."

6. GRAND CANYON (GRAND) SCIENCE UNIT

A. Stimuli: GRAND CANYON

(Okuma Parçası: GRAND CANYON (BÜYÜK KANYON)

Grand Canyon (GRAND) Stimulus

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d. Irrelevance with national curriculum- *Content*	5	19.2	19.2	19.2
e.Unnecessary context-*Content*	5	19.2	19.2	38.5
o.quality of visual stimuli-*Presentation*	7	26.9	26.9	65.4
q.Irrelevance with topic-*Content*	9	34.6	34.6	100.0
Total	26	100.0	100.0	

Based on the teachers' comments above, it seems that there were 26 negative entities which are the members of four different sub-categories. *Irrelevance with topic* was the most frequent one that nine teachers commented on it. *Irrelevance with the national curriculum* and *unnecessary context* was mentioned by the five teachers. There will be presented one example about the *quality of visual stimulus.*

> Verilen resimdeki katmanlar yeterince açık değil, sorular resmin netliği ile ilgili olmasa bile öğrencilerin verilen paragrafı okuması sırasında zaman kaybetmesine yol açar..-the layers in the given picture is not clear enough, even if the questions are not related with the clearness of the picture it cause student to consume more time when reading. (T42)

Grand Canyon (GRAND) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
l.incompetent item stem-*Structure*	6	46.2	46.2	46.2
n.vague expectancy- *Typicality*	1	7.7	7.7	53.8
q.Irrelevance with topic-*Content*	2	15.4	15.4	69.2
3. different measure form aimed	4	30.8	30.8	100.0
Total	13	100.0	100.0	

There were four sub-groups of negative entities found in the first item of the GRAND science unit. Six of the teachers commented on the negative entity of *incompetent item stem.* Different measure from aimed in program was mentioned by four teachers. Two teachers commented on the *Irrelevance with topic* and only one teacher referred to the *vague expectancy*. One example from the writings of teachers related with the different measure from aimed was given below.

Bu soruda ölçülmek istenen beceri çok temel düzeyde ve ölçülmek istenen beceri fen ile ilgili değil…- the objective aimed to be measured is at the basic level and it is not one of the science program objectives. (T46)

Grand Canyon (GRAND) 2st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
e.Unnecessary context-*Content*	4	16.0	16.0	16.0
g.Difficulty in grammar-*Language*	6	24.0	24.0	40.0
l. incompetent item stem-*Structure*	4	16.0	16.0	56.0
p.cultural irrelevance- *Content*	2	8.0	8.0	64.0
u.unfamiliar item stem- *Typicality*	4	16.0	16.0	80.0
x. different objective from program- *Content*	5	20.0	20.0	100.0
Total	25	100.0	100.0	

The second item of the GRAND item group seemed to include six types of negative entities. Teachers remarked *difficulty in grammar* more frequently. Five teachers referred to the different *objective from program*. *Unnecessary context, incompetent item stem* and *unfamiliar item stem* were commented by four teachers. Two of teachers mentioned to the *cultural unfamiliarity.*

As in the below example, the item stem seemed to be unfamiliar to the students:

Bu soru formatı öğrenciye yabancı..- the format of the question is not well-known by the students. (T50)

The third item of the GRAND science unit was found to embody two types of the negative entities which were *irrelevant cue* and *different objective from program.* An example from the teachers' comments on *irrelevant cue* was presented below:

Grand Canyon (GRAND) 3nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
w.irrelevant cue-*Structure*	9	75.0	75.0	75.0
x. different objectives from program- *Content*	3	25.0	25.0	100.0
Total	12	100.0	100.0	

> Sorunun seçenekleri cevap olabilecek nitelikte yazılmamış ve bu öğrenciyi doğru seçeneği işaretlemeye yönlendiriyor…-the alternatives of the question were not written as good alternatives and this leads students to mark the right alternative. (T48)

In the fourth item of the GRAND science unit, teachers found 24 themes of negative entities. *Uncommon vocabulary* was the most frequent one of them that followed by questioning style. Another negative entity was the *cultural unfamiliarity*. *Irrelevance with national curriculum* and *difficulty in grammar* were mentioned by four two teachers. Only one teacher referred to the irrelevance with topic.

Grand Canyon (GRAND) 4rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d. Irrelevance with national curriculum- *Content*	2	8.3	8.3	8.3
g.Difficulty in grammar-*Language*	2	8.3	8.3	16.7
h.Uncommon vocabulary-*Language*	9	37.5	37.5	54.2
p.cultural irrelevance- *Content*	3	12.5	12.5	66.7
q.Irrelevance with topic-*Content*	1	4.2	4.2	70.8
r. questioning style –*Presentation*	7	29.2	29.2	100.0
Total	24	100.0	100.0	

One of the teachers' comments revealed that the item found to be difficult for students due to the questioning style:

> Ayrıca B seçeneğide doğru cevaplardan biri olmaya aday bunun asıl sebebi sorunun verilen durum en iyi hangi seçenek ile açıklanabilir diye sorulmaması…-…Additionally, B is one of the possible true alternatives. Asking question in the form of 'best reason' is the main reason for this. (T42)

7. PHYSICAL EXERCISE (PHYSICAL) SCIENCE UNIT

A. Stimuli: PHYSICAL EXERCISE

(Okuma Parçası: BEDEN EĞİTİMİ HAREKETLERİ)

Physical Exercise (PHYSICAL) Stimuli

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
o.quality of visual stimuli- *Presentation*	10	71.4	71.4	71.4
p. cultural irrelevance- *Content*	4	28.6	28.6	100.0
Total	14	100.0	100.0	

For PHYSICAL science unit stimuli, teachers mentioned only two types of negative entities that they were quality of visual stimuli and cultural unfamiliarity. Quality of stimuli was written ten times and cultural unfamiliarity was mentioned four times by the teachers. One of the examples for quality of stimuli as below,

> Kullanılan resmin ve cümlenin niteliği soruya anlam kazandıracak yada temel teşkil edecek şekilde değil….-the quality of picture and sentence is not enough to constitute meaning or a base for the question. (T52)

The following is an example for the cultural unfamiliarity that one of the teachers stated,

> Resim ve başlık bizim öğrendiğimiz yada yaşadığımız ve yaptığımız şeklinle benzer değil…- the picture and sentence are not familiar to the one that we learn, live and do like. (T57)

Based on the comments of the teacher it is decided to omit the picture from stimuli and change the sentence as more frequent one used in Turkish.

For the first item of the PHYSICAL science unit, teachers mostly stated that the item included *incompetent alternatives*. The *unnecessary context* was the second negative entity mentioned by teachers and *irrelevance with national curriculum* took third place. The *difficulty in grammar* was stated only one of the teachers.

Physical Exercise (PHYSICAL) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d. Irrelevance with national curriculum- *Content*	3	18.8	18.8	18.8
e.Unnecessary context-*Content*	4	25.0	25.0	43.8
g.Difficulty in grammar-*Language*	1	6.3	6.3	50.0
v. Incompetent alternatives-*Structure*	8	50.0	50.0	100.0
Total	16	100.0	100.0	

One of the teachers made comment on the incompetent alternatives as below,

> Ikinci ve üçüncü kutulardakiler birbirini tamamlıyor fazla kilolardan kurtulmak için beslenme düzenine de geçmek gerek bu ifadeler tam değil değişmeli….- the second and third items complete each other, that is to mean, to be prevented form extra weights it is also necessary to have a regular diet, so these statements is required to be rearranged. (T59)

The revision on the item was made by changing the verb of the second statement to provide a clear understanding for sentence.

Physical Exercise (PHYSICAL) 2st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
e.Unnecessary context-*Content*	5	35.7	35.7	35.7
y. extreme simple item-*Typicality*	9	64.3	64.3	100.0
Total	14	100.0	100.0	

The third item of the science-unit was commented to include two types of negative entity which are unnecessary context and extremely simple item. Nine of the teachers stated that the item was so simple for Turkish students. One of the teachers explained this as,

> Müfredatta vurgulanan bir konu olduğundan öğrenciler için hayli basit bir soru….- it is a very simple question for my students that it is mostly emphasized in the curriculum. (T53)

Another negative entity was *unnecessary context* that five teachers stated the repeating the question stem to be unnecessary in the item.

Birinci sorudakine benzer olarak soruyu ikinci kez tekrar yazmak gereksiz…. -again like in the first question, it is not necessary to rewrite the question stem for the second time. (T51*)*

For the revision of the item it is decided to delete the question sentence in the box and to keep the question sentence outside of the question statements.

Physical Exercise (PHYSICAL) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
r. questioning style –*Presentation*	10	100.0	100.0	100.0
Total	10	100.0	100.0	

The third item of the science unit only mentioned to cover one type of negative entity which is *questioning style. One of the teachers comments like,*

Bu soruyu öğrencinin açıklaması ile birlikte sorsak daha iyi olur…- it would be better to ask the question by adding a statement which asks students to explain the reason. (T60)

The revision of the item based on the refinement of the item stem statement.

REVISED VERSION OF THE PHYSICAL SCIENCE UNIT

A. STIMULI

DÜZENLİ SPOR YAPMAK

B. 1st ITEM

Düzenli spor yapmanın yararları nelerdir? Her ifade için "evet" yada hayır" seçeneklerinden sadece birini yuvarlak içine alınız.

Düzenli spor yapmak, kalp ve dolaşım hastalıklarından korunmaya yardımcı olur.	Evet / Hayır
Düzenli spor yapmak, sağlıklı bir beslenmeye *yönlendirir/yol açar.*	Evet / Hayır

C. 2nd ITEM

Kaslar çalıştırıldığı zaman aşağıdaki olaylar gerçekleşir mi?

Her ifade için "evet" yada hayır" seçeneklerinden sadece birini yuvarlak içine alınız

Kaslara gelen kan akışının artması	Evet / Hayır

D. 3rd ITEM

Dinlenmedeki durumumuzla karşılaştırıldığında, *spor yaparken daha sık nefes alıp vermemizin nedeni nedir? Kısaca açıklayınız.*

8. MARY MONTAGU (MARY) SCIENCE UNIT

A. Stimuli: THE HISTORY OF VACINATION
(Okuma Parçası: AŞININ TARİHÇESİ)

Mary Montagu(MARY) Stimulus

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
c.Difficulty in statement-*Language*	6	28.6	28.6	28.6
d. Irrelevance with national curriculum- *Content*	2	9.5	9.5	38.1
e.Unnecessary context-*Content*	8	38.1	38.1	76.2
i.Wrong information-*Content*	5	23.8	23.8	100.0
Total	21	100.0	100.0	

Difficulty in statement, *irrelevance with national curriculum*, *unnecessary context* and *wrong information* were four negative entity subcategories that mentioned for the stimulus of MARY item group. Most of the teachers commented that the context of the stem was not necessary. As one of the teachers wrote,

> Ama sorulara bakıldığında aslında parçaya ihtiyaç olmadığı da söylenebilir….- but when questions are examined, in fact it can be said that there is no need for stimulus article to answer the questions. (T61)

However, for the revision of the item group stimulus experts decided that it was necessary to keep the article in the item group. The reason for this explained to be the providing equal assessment environment, especially in terms of time usage, for both comparison groups.

Another example of teacher comment on the *wrong information* that is like,

> Aşının tarihçesinde ondan bir tedavi şekli olarak bahsetmek yanlış…-it is a fault to mention the vaccination as a curing method in the history of it. (T69)

Six teachers stated that stimulus included difficult statements. For example,

> Son paragrafta anlam düşüklüğü var, ifadelerde aynı zaman kalıplarının kullanılması daha güzel olurdu....-there is a missing meaning in the last paragraph, it would be better to use same time fractions on the whole writing. (T64)

The changes on the item stimulus were based on the difficulty in statements and wrong information negative entities.

Mary Montagu (MARY) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
k. multiple true answers- *Structure*	3	20.0	20.0	20.0
s. item-alternative inconsistency – *Content*	6	40.0	40.0	60.0
w.irrelevant cue-*Structure*	6	40.0	40.0	100.0
Total	15	100.0	100.0	

In the first item of the MARY item group, teachers realized three types of negative entity sub-categories which were *multiple true answers, item-alternative inconsistency* and *irrelevant cue.*

One of the teachers commented on the *multiple true answers* as below,

> Öğrenciler soruyu kendi bildiklerine gore cevaplayacaksa bir şıkkı, parçadan çıkarım yapacaklarsa diğer bir şıkkı da işaretleyeblirler….- if students answer the question according to their knowledge they can sign alternative, otherwise if they answer b using their own knowledge they will chose another. (T65)

For the *irrelevant cue* one teacher stated as in the quote below,

> Virüs sadece B seçeneğinde verildiğinden ve okuma parçasında da sadece virus denildiğinden öğrenci için gereksiz ipucu olmuş…-virus is given only in alternative B that only the term virus is mentioned in the paragraph so it given an unnecessary cur for students. (T69)

In the revision part, the word ‘virus’ was omitted from the alternative B in order to eliminate negative entities of *irrelevant cue* and *item-alternative inconsistency.*

Mary Montagu (MARY) 2nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
f.Worse Alternative-*Structure*	6	40.0	40.0	40.0
i.Wrong information-*Content*	3	20.0	20.0	60.0
s. item-alternative inconsistency – *Content*	6	40.0	40.0	100.0
Total	15	100.0	100.0	

Worse alternative, wrong information and *item-alternative inconsistency* were three negative entity sub-categories that were stated by teachers. One example from comments of worse alternative sub-category will be given.

> Doğru cevap olan C seçeneğinde kullanılan cümle çok karışık….- the sentence in the alternative C which is the true alternative is so complicated. (T61*)*

The word bacteria changed with the virus also the statement in the alternative B changed *item-alternative* inconsistency.

Mary Montagu (MARY) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
r. questioning style –*Presentation*	8	66.7	66.7	66.7
s. item-alternative inconsistency – *Content*	4	33.3	33.3	100.0
Total	12	100.0	100.0	

Questioning style and *item-alternative inconsistency* were two negative entity sub-categories that they mentioned by teachers. The example of the questioning style is like below,

> Bu önerinin sebebini açıklayınız şeklinde sorulsa daha açık olurmuş….-it would be clearer if the question was asked as explain the reason of this suggestion. (T64)

The revision of the third item based on the questioning style of the item as writing the reason of this suggestion.

REVISED VERSION OF THE PHYSICAL SCIENCE UNIT

A. STIMULUS

Aşağıdaki gazete yazısını okuyunuz. Soruları yanıtlayınız

AŞININ TARİHÇESİ

Mary Montagu güzel bir kadındı. 1715 yılında çiçek hastalığına yakalandı. Hastalığı geçirdi; fakat izleri kaldı. 1717 yılında Türkiy*e'deyken*, bu ülkede yaygın olarak kullanılan ve adına aşılama denen bir *yöntem* gördü. Bu *yöntemle* sağlıklı gencin derisi çizilerek ona zayıflatılmış çiçek virüsü veriliyordu. Kişi kısa bir süre için hasta oluyor, ancak hastalığı genellikle çok hafif bir şekilde geçi**ri**yordu.

Mary, bu aşılama yönteminin güvenli olduğuna inandı ve kendi oğlu ile kızının da bu şekilde aşılanmasına izin verdi.

B. 1st ITEM

İnsanlar hangi çeşit hastalıklara karşı aşılanabilir?

A. Hemofili gibi kalıtsal hastalıklar

B. Çocuk felci gibi *virüslerin neden olduğu* hastalıklar

C. Şeker hastalığı gibi vücudun işlevsel bozukluklarından kaynaklanan hastalıklar

C. 2nd ITEM

Hayvanlar yada insanlar *virüslerin* neden olduğu bulaşıcı bir hastalığa yakalanır ve iyileşirse, hastalığa neden olan *virüsler* genellikle onlarda tekrar hastalık oluşturamaz.

Bunun nedeni aşağıdakilerden hangisidir?

A Vücudun, aynı çeşitten bir hastalığa neden olabilecek bütün bakterileri öldürmüş olması

B Vücudun, bu **tür** virüsleri *vücutta çoğalmadan öldürebilen antikor yapabilmesi*

D. 3rd ITEM

Gribe karşı özellikle küçük çocuklar ve yaşlı insanların aşılanmaları önerilmektedir.

Bunun nedenini yazınız.

8. SUNSCREEN (SUN) SCIENCE UNIT

A. Stimuli: SUNSCREEN
(Okuma Parçası: GÜNEŞTEN KORUYUCULAR)

Sun Screens(SUN) Stimulus

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
b.Difficulty in clause-*Language*	2	7.1	7.1	7.1
c.Difficulty in statement-*Language*	4	14.3	14.3	21.4
g.Difficulty in grammar-*Language*	9	32.1	32.1	53.6
h.Uncommon vocabulary-*Language*	5	17.9	17.9	71.4
q.Irrelevance with topic-*Content*	3	10.7	10.7	82.1
u.unfamiliar item stem- *Typicality*	5	17.9	17.9	100.0
Total	28	100.0	100.0	

Teachers identified six types of negative entities in the stimulus of Sun item group. The *difficulty in grammar* was the most frequently mentioned sub-categories that it was followed by *unfamiliar item stem* and *uncommon vocabulary*.

The following is an example for the *difficulty in grammar*,

> Türkçesi tekrar gözden geçirilmeli- the language (Turkish) in the writing needs to be revised. (T76)

Another example for *uncommon vocabulary,*

> Ultraviole demek yerine morötesi ışık desek daha güzel olur, ayrıca güneşten koruma ürünü yerine güneş kremi daha yaygın bir kullanım olur ve okumayı kolaylaştırır…-it would be better to say 'morötesi' instead of 'ultraviole',and also using 'güneş kremi' instead of 'güneşten koruma ürünü' is a more common usage and would make reading easier. (T78)

The revision on the stimulus of SUN item group depended on the making changes on the lay-out of the writing and language category.

Sun Screens(SUN) 1st Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
d. Irrelevance with national curriculum- *Content*	4	16.7	16.7	16.7
g.Difficulty in grammar-*Language*	5	20.8	20.8	37.5
l.incompetent item stem-*Structure*	6	25.0	25.0	62.5
s. item-alternative inconsistency – *Content*	5	20.8	20.8	83.3
v. Yetersiz Alternatives- *Structure*	1	4.2	4.2	87.5
x. different objective from program- *Content*	3	12.5	12.5	100.0
Total	24	100.0	100.0	

In the first item of the item group, teachers provided comments on the negative entities of *irrelevance with national curriculum*, *difficulty in grammar*, *incompetent item stem*, *item-alternative inconsistency*, *incompetent alternatives* and *different objective from program.*

One teacher commented on the *incompetent item stem* that,

> Soru ile cevaplar tam birbirlerini karşılamıyor, okuma parçasında hiç etken den bahsedilmemişken ve okuma parçasının anlaşılması bu kadar zorken sorunun daha iyi yazılması gerekir…- the item stem and alternatives do not complete each other that the question is needed to be written better because in the reading paragraph (stimulus) the word 'etken' is not mentioned and the reading paragraph(stimulus) is so difficult to understand. (T80)

The revision of the first item made by eliminating the word 'etken' from the alternatives and using 'güneş kremi' instead of 'güneşten koruyucular'.

The second item of the group was commented to include six types of negative entity which were *connection with other item, worse alternative,* u*ncommon vocabulary, wrong concept, incompetent item stem and irrelevant cue.*

Sun Screens(SUN) 2nd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
1. connection with other item-*Structure*	4	18.2	18.2	18.2
f. Worse alternative- *Structure*	4	18.2	18.2	36.4
h.Uncommon vocabulary-*Language*	4	18.2	18.2	54.5
j.Wrong Concept- *Content*	4	18.2	18.2	72.7
l.incompetent item stem-*Structure*	3	13.6	13.6	86.4
w.irrelevant cue-*Structure*	3	13.6	13.6	100.0
Total	22	100.0	100.0	

An example form the teachers' statements on the *incompetent item stem* as below,

> Verilen çalışma ile birkaç soruya cevap bulunabilir o yüzden en iyi cevap A seçeneğidir, o şekilde sorulmalı…- it is possible to answer some of the question by depending on study in the reading passage, because of that the alternative A is the best question to be answered that it should be asked like this. (T77)

The revision of the item placed on the language and structure sub-categories of negative entities that uncommon word of 'ultraviole' changed, layout of the passage are redesigned and grammar of the alternatives was changed.

Sun Screens(SUN) 3rd Item

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
r. questioning style –*Presentation*	6	60.0	60.0	60.0
s. item-alternative inconsistency – *Content*	3	30.0	30.0	90.0
x. different objective from program- *Content*	1	10.0	10.0	100.0
Total	19	100.0	100.0	

Teachers stated five different negative entity sub-categories on the third item of the group. One example from teacher comments related with *questioning style* as below,

> Tabakanın üzerine büyük bir kitap ile bastırılmasının temel nedeni nedir gibi sorulsa daha net olacak gibi…- it seems to be better to ask the question as what is the main reason to press down on the plastic layers by a big book. (T76)

The revision on the item was made by reorganizing the item stem and alternatives in that direction and also revising the B alternative.

Sun Screens (SUN) Stimulus

Negative Entities	Frequency	Percent	Valid Percent	Cumulative Percent
f.Worse Alternative-*Structure*	6	28.6	28.6	28.6
i.Wrong information-*Content*	3	14.3	14.3	42.9
o.quality of visual stimuli-*Presentation*	7	33.3	33.3	76.2
r. questioning style –*Presentation*	5	23.8	23.8	100.0
Total	21	100.0	100.0	

The last item of the sun item group was commented to involve *worse alternative*, *wrong information*, *quality of visual stimuli* and *questioning style* sub-categories. One teacher stated that *questioning style* was not appropriate for students,

> Bu gibi durumlarda soruyu açık ve net yazmak önemlidir, burada soru kökünün yapılan deney sonucunda güneş ışığı altında yeterince kalmış olması durumda sorulduğu belirtilmelidir….-in situations like that, it is really important to write a clear question that in the item stem it needs to be notified that the answer will be based on the results of the experiment while the paper stayed enough under the sunlight. (T72)

The revision of the item carried out via enlarging the visual stimulus and revising the item stem.

REVISED VERSION OF THE SUN SCIENCE UNIT

A. STIMULI

Jale ve Osman,cildi güneşten koruyan güneş kremlerinden hangisinin en iyi korumayı sağladığını merak ettiler. **Bu araştırma için şu bilgileri edindiler;**

Güneş kremleri için, her ürünün güneş ışığındaki morötesi ışınlarını ne derecede emdiğini gösteren bir *Güneşten Koruma Faktörü (GKF)* tanımlanmıştır. GKF'si yüksek olan bir güneşten koruyucu, GKF'si düşük olan bir güneşten koruyucuya göre cildi daha uzun süre korur.

Jale, bazı güneş kremlerini birbiriyle karşılaştırmak için bir yol düşündü.

Osman ile birlikte aşağıdaki malzemeleri topladılar:

- güneş ışığını emmeyen (geçiren) iki temiz plastik tabaka;
- bir adet ışığa duyarlı kağıt;
- S1, S2, S3 ve S4 adını verdikleri dört farklı güneş kremi.
- mineral yağ (M) ve çinko oksit (ZnO) içeren **birer** krem

Jale ve Osman, mineral yağı güneş ışınlarının çok büyük bir kısmını geçirdiği için, çinko oksidi de güneş ışınlarının tamamına yakınını geçirmediği için seçtiler.

Osman, **birinci plastik tabaka üzerinde yuvarlak içine alınmış yerlere maddelerin herbiri sadece bir yuvarlakta olacak şekilde birer damla koydu.Bunların üzerini ikinci plastik tabaka ile kapattı.** Sonra tabakaların üzerine büyük bir kitap yerleştirerek üstten iyice bastırdı.

B. 1st ITEM

Aşağıdaki ifadelerden hangisi, güneş kremlerinin etkililiğini karşılaştırma amacıyla yapılan bir çalışmada mineral yağ ve çinko oksidin rolünün bilimsel tanımıdır?

A Mineral yağ ve çinko oksidin ikisi de etkisi araştırılan birer maddedir.

B Mineral yağ test edilen, çinko oksit ise karşılaştırma için kullanılan bir maddedir.

C. 2nd ITEM

Jale ve Osman'ın yanıtlamaya çalıştığı soru aşağıdakilerden hangisinde en iyi ifade edilmiştir?

A **Güneş kremlerinden** her birinin koruma gücü diğerlerine kıyasla nasıldır?

B Güneş kremleri cildi **morötesi** ışınlarından nasıl korur?

C Mineral vağdan daha az **koruvan bir günes kremi** var mıdır?

E. 3rd ITEM

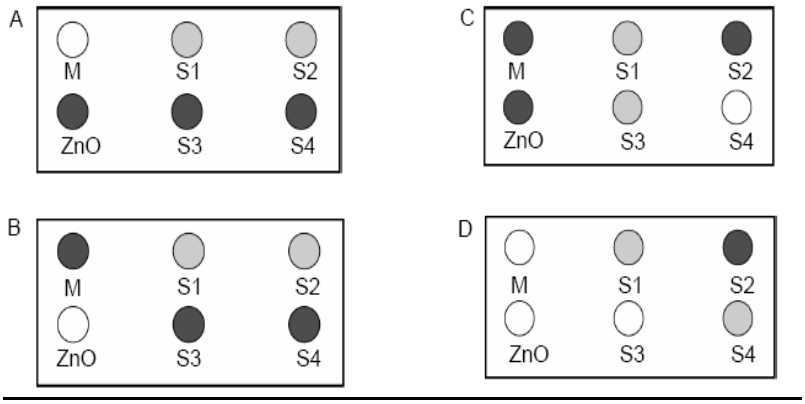
İkinci plastik tabakanın üzerine büyük bi kitap ile bastırılmasının temel nedeni nedir?

A Damlaların kurumasını önlemek

B Damlaları **tabakaya** yaymak

F. 4rd ITEM

Aşağıdaki şekillerden hangisi **yapılan araştırma sonucunda ışığa duyarlı kağıt üzerinde** elde edilebilecek sonucu göstermektedir? Neden bunu seçtiğinizi açıklayınız.



APPENDIX G: STUDENT OPINION SURVEY

ÖĞRENCİ GÖRÜŞ ÖLÇEĞİ

Ad / Soyad: **Yaş:**

Okul (devlet/özel): **Sınıf:**

Cinsiyet:

AÇIKLAMA;

Sevgili öğrenciler,

Aşağıda size verilen her bir okuma parçasını dikkatle okuyunuz. Her bir okuma parçasını, sayfa düzenlemesi, Türkçe kullanımı ve günlük hayatta veya okulda bu tip bir konuyla karşılaşma sıklığınız bakımından değerlendiriniz. Derecelendirmenizi 1 ile 5 arasındaki bir sayıyı işaretleyerek gösteriniz.

	Asla	Nadiren	Bazen	Sık sık	Her zaman
	1	2	3	4	5
Okulda bu konuyu işliyoruz.					
Günlük hayatta(ev,tv, internet) buna benzer konularla karşılaşıyorum.					
Okuma parçasında kullanılan dile alışkınım.					
Okuma parçasının sayfa düzenine (şekil,yazı karakterleri vb.) alışkınım.					

APPENDIX H: NEGATIVE ENTITIES AND MAIN CATEGORIES

LANGUAGE long sentence difficulty in clause difficulty in statement difficulty in grammar uncommon vocabulary	The main category comments on the negative entities related with sentence length, grammatical difficulty in statements, clauses and grammar together with vocabulary that it includes uncommon words for Turkish students.
CONTENT irrelevance with national curriculum irrelevance with topic unnecessary context wrong concept cultural irrelevance item-alternative inconsistency different objective form program wrong information	The main category includes negative entities related with the topic and national curricula present in Turkish education system together with properties of what is said in the stimuli or item.
STRUCTURE worse alternative multiple true alternatives incompetent item stem incompetent alternatives irrelevant cue connection with other item different objective from aimed questioning style	The main category refers to the negative entities related with the item stem and alternative properrties that it includes subcategories presented left side.
TYPICALITY vague expectancy extreme simple item unfamiliar item stem expectancy error	The main category includes negative entities with the references clearity of purpose of the stimuli or items together with consistency of writing aims.
PRESENTATION lack of visual element quality of visual element Inappropriate lay-out arrangement unfamiliarity	The main category composed of negative entities related with the overall structure of the stimuli or items that the way things are presented.

APPENDIX I: MARKING GUIDE FOR TURKISH VERSION OF RELEASED PISA 2006 SCIENCE UNITS

The second phase of the study included implementation of two tests (PISA-OT and PISA-RT) that answers of the students for these tests required to be scored. For the marking process the Turkish version of answer key used that it is check with the English version of the answer key prepared by PISA.

SERA PUANLAMA 3.1

Tam Puan

Kod 11: Hem (ortalama) sıcaklık hem de karbon dioksit yayılımındaki artışlara değinir.
Gaz yayılımları arttıkça sıcaklık arttı.
Her iki grafik de artıyor.
Çünkü 1910 yılında her iki grafik de artmaya başladı.
CO_2 yayılımı oldukça sıcaklık artıyor.
Grafiklerdeki bilgi çizgileri birlikte artıyor.
Her şey artıyor.
Daha fazla CO_2 yayılımı, daha yüksek sıcaklık demektir.

Kod 12: Sıcaklık ve karbon dioksit yayılımı arasındaki pozitif bir ilişkiye (genel anlamda) değinir.
[Not: Bu kod, öğrencilerin 'pozitif ilişki', 'benzer şekil' ya da 'doğru orantılıdır' gibi terminolojiyi kullanımlarını yakalamayı amaçlamaktadır; buna rağmen aşağıdaki örnek yanıt tamamen doğru değildir, burada puan verilebilecek yeterli anlayış düzeyini göstermektedir.
Toplam CO_2 miktarı ve Dünya'nın ortalama sıcaklığı doğru orantılıdır.
Onların benzer bir şekli var, bu da bir ilişkiyi göstermektedir.

Sıfır Puan

Kod 01: Ya (ortalama) sıcaklık ya da karbon dioksit yayılımındaki artışa değinir.
Sıcaklık yukarı fırlamıştır.
CO_2 artıyor.
O, sıcaklıklardaki çarpıcı değişikliği göstermektedir.

Kod 02: İlişkinin doğası hakkında net bir görüş bildirmeden sıcaklık ve karbon dioksit yayılımına değinir.
Karbon dioksit yayılımının (1. grafik) Dünya'nın artan sıcaklığı (2. grafik) üzerinde bir etkisi vardır.
Karbon dioksit Dünya'nın sıcaklığındaki artışın esas nedenidir.

YA DA

Diğer yanıtlar.
Karbon dioksit yayılımı, Dünya'nın ortalama sıcaklığından çok daha fazla artıyor. *[Not: Bu yanıt doğru değildir çünkü, CO_2 yayılımı ve sıcaklıktaki artış düzeyi yanıt olarak görünüyor,[her ikisinin de artmakta olduğu belirtilmiyor.]*
CO_2'in yıllar geçtikçe artışı, Dünya'nın atmosferindeki sıcaklık artışından dolayıdır.
Grafiğin doğrultusu yukarıya doğrudur.
Bir artış vardır.

Kod 99: Boş.

SERA PUANLAMA 3.2

Tam Puan

Kod 2: Grafiklerin her ikisinin birlikte azalmadığı ya da birlikte artmadığı belirli bir bölümüne değinir ve buna uygun gelen açıklamayı verir.

1900–1910 yıllarında (yaklaşık olarak), CO_2 artıyordu, buna karşılık sıcaklık aşağıya iniyordu.
1980–1983 yıllarında karbon dioksit aşağı indi ve sıcaklık arttı.
1800 'lerde sıcaklık hemen hemen aynı kaldı ama birinci grafik tırmanmaya devam etti.
1950 ve 1980 arasında sıcaklık artmadı ama CO_2 arttı.
1940'dan 1975'e kadar sıcaklık yaklaşık aynı kalır ama karbon dioksit yayılımı keskin bir yükselme gösterir.
1860'dan 1900'e kadar karbon dioksit çok az artan bir eğridir, buna karşılık sıcaklık eğrisi çok fazla dalgalanmalar gösterir.
1940'ta sıcaklık 1920'den oldukça fazladır ve onların benzer karbon dioksit yayılımı vardır.

Kısmî Puan

Kod 1: Doğru bir zaman aralığından bahseder, ama hiç açıklama vermez.
1930–1933.
1910 civarında.

Belirli bir yıldan bahseder (bir zaman aralığı değildir), kabul edilebilir bir açıklama verir.

[Not: Eğer açıklama grafiklerden birindeki bir düzensizlik üzerinde odaklanırsa Kod 14 kullanılmalıdır.]
1980'de yayılım seviyesi düşüktür ama, sıcaklık artmaya devam etmiştir.
1910 yılında karbon dioksit arttı ve sıcaklık düştü.

YA DA

Ali'nin sonucunu desteklemeyen bir örnek verir ama, zaman aralığından bahsederken bir hata yapar.
1950 ve 1960 arasında sıcaklık azaldı ve karbon dioksit yayılımı arttı.

Belirli bir zaman aralığından bahsetmeden, iki eğri arasındaki farklılıklara değinir.
Gaz yayılımı azalsa da, bazı yerlerde sıcaklık artar.
İlk başta daha az yayılım vardı ama yine de sıcaklık yüksektir.
Onlar aynı oranda artmazlar.
1. grafikte sürekli bir artış varken, 2. grafikte artış yoktur, o sabit kalır. *[Not: O, 'tamamen' sabit kalır.]*
Çünkü başlangıçta karbon dioksit çok düşükken sıcaklık hâlâ yüksekti.

Grafiklerden birindeki bir düzensizliğe değinir.
Sıcaklık düştüğünde yaklaşık olarak 1910 yılıydı ve belirli bir zaman aralığında bu şekilde devam etti.
İkinci grafikte 1910 yılında Dünya atmosferinin sıcaklığında bir düşüş vardır.

Grafiklerdeki farkı belirtir, ama açıklama zayıftır.
1940'larda sıcaklık çok yüksekti, ama karbon dioksit çok düşüktü. *[Not: Açıklama çok zayıftır, ama belirtilen farklılık açıktır.]*

Sıfır Puan

Kod 0: İki grafiğe özel olarak değinmeden bir eğrideki düzensizliğe değinir.
O, biraz yukarı çıkar ve iner.
O, 1930'da aşağıya inmiştir.

Hiç bir açıklama olmaksızın zayıfça tanımlanan bir zaman aralığına ya da yıla değinir.
Orta bölüm.
1910.

Diğer yanıtlar.
1940'da ortalama sıcaklık arttı, ama karbon dioksit yayılımı artmadı.
1910 civarında sıcaklık arttı ama, gaz yayılımı artmadı.

Kod 9: Boş.

SERA PUANLAMA 3.3

Tam Puan

Kod 11: Güneş'ten gelen enerjiye / radyasyona değinen bir etken verir.
Güneş'in ısıtması ve belki Dünya'nın konumunu değiştirmesi
Dünya'dan geri yansıyan enerji

Kod 12: Doğal bir bileşen ya da potansiyel bir kirletici etkenden söz eder.
Havadaki su buharı
Bulutlar.

Volkanik püskürme gibi şeyler.
Atmosfer kirliliği (gaz, yakıt).
Egzoz gazı miktarı
CFC'ler (Kloroflorokarbonlar).
Arabaların sayısı.
Ozon (havanın bir bileşeni olarak). *[Not: tükenmeye yapılan atıflar için Kod 03'ü kullanınız.]*

Sıfır Puan

Kod 01: Karbon dioksit konsantrasyonunu etkileyen bir nedene değinir.
Yağmur ormanlarının temizlemesi
CO_2 yayılım miktarı.
Fosil yakıtlar.

Kod 02: Özel olmayan bir etkene değinir.
Gübreler.
Spreyler.
Son zamanlarda hava durumunun nasıl olduğu

Kod 03: Diğer doğru olmayan etkenler ya da diğer yanıtlar.
Oksijen miktarı.
Azot.
Ozon tabakasındaki delik de gittikçe daha büyüyor.

Kod 99: Boş.

GİYSİLER PUANLAMA 4.1

Tam Puan

Kod 1: Evet, Evet, Evet, Hayır, sıralama bu şekilde.

Sıfır Puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GİYSİLER PUANLAMA 4.2

Tam Puan

Kod 1: A. Voltmetre.

Sıfır Puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GRAND KANYON (BÜYÜK KANYON) PUANLAMA 5.1

Tam puan

Kod 1: İkisi de doğrudur: Evet, Hayır sırasıyla

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş

GRAND KANYON (BÜYÜK KANYON) PUANLAMA 5.2

Tam puan

Kod 1: D. Kaya çatlaklarında donan su genleşir.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GRAND KANYON (BÜYÜK KANYON) PUANLAMA 5.3

Tam puan

Kod 1: C. O zamanlarda okyanus buraları kaplamıştı, sonra sular eski yerine çekildi.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GÜNEŞTEN KORUYUCULAR PUANLAMA 6.1

Tam puan

Kod 1: D. Mineral yağ ve çinko oksidin ikisi de karşılaştırma için kullanılan birer maddedir.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GÜNEŞTEN KORUYUCULAR PUANLAMA 6.2

Tam puan

Kod 1: A. Güneşten koruyucu maddelerden her birinin koruma gücü diğerlerine kıyasla nasıldır?

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GÜNEŞTEN KORUYUCULAR PUANLAMA 6.3

Tam puan

Kod 1: D Damlalara eşit kalınlık vermek için

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

GÜNEŞTEN KORUYUCULAR PUANLAMA 6.4

Tam puan

Kod 2: A. ZnO'nun bulunduğu yuvarlağın (güneş ışığını engellediği için) koyu gri **ve** M'nin bulunduğu yuvarlağın (mineral yağ çok az güneş ışığı emdiği için) beyaz olduğunu açıklayan yanıtlar.

[Parantez içinde gösterilen ileri düzeydeki açıklamalar (yeterli olsa da) gerekli değildir]

- ZnO gelen güneş ışınlarını engelledi ve M geçmesine izin verdi.
- A'yı seçtim çünkü en açık olanın mineral yağ, en koyu olanın da ZnO olması gerekir.

Kısmi Puan

Kod 1: A. Ya ZnO yuvarlağı **ya da** M yuvarlağı için doğru açıklama yapar, fakat her ikisi için doğru açıklama **yapmaz ve** diğer yuvarlak için de yanlış açıklama yoktur.

- Mineral yağ UV ışınlarına karşı en az direnci gösterir. Bu nedenle diğer maddeler için kağıt beyaz olmayacaktır.
- ZnO hemen hemen tüm ışınları emer ve şekil bunu göstermektedir.

Sıfır puan

Kod 0: Diğer yanıtlar.

- A. ZnO ışığı engeller ve M emer.
- B. ZnO ışığı engeller ve M geçmesine izin verir.

Kod 9: Boş.

MARY MONTAGU PUANLAMA 7.1

Tam puan

Kod 1: B Çocuk felci gibi virüslerin sebep olduğu hastalıklar.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

MARY MONTAGU PUANLAMA 7.2

Tam puan

Kod 1: B. Vücudun, bu tür bakterileri çoğalmadan önce öldürecek antikorlar yapmış olması

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

MARY MONTAGU PUANLAMA 7.3

Tam puan

Kod 1: Genç yaşta olanların ve /ya da yaşlıların diğer insanlardan daha zayıf bağışıklık sistemi olduğundan bahseden yanıtlar ya da benzeri.

Puanlama Notu : Verilen neden ya da nedenler genel olarak herkesi değil de-özellikle genç yaşta olanlar ve yaşlı insanları işaret etmek zorunda. Aynı zamanda yanıtlar, bu insanların diğer insanlara göre daha zayıf bir bağışıklık sisteminin olduğunu dolaylı ya da doğrudan belirtmelidir-genel olarak, sadece "daha zayıf" demekle yetinmemelidir.

- Bu insanların hastalıklara karşı daha az dayanıklılığı vardır.
- Genç yaşta olanlar ve yaşlılar diğerleri kadar hastalıklarla baş edemez.
- Gribe daha çok yakalanma olasılıkları vardır.
- Gribe yakalanırlarsa bu insanlardaki etkiler daha kötü olabilir.
- Çünkü küçük çocukların ve yaşlı insanların organizmaları daha zayıftır.
- Yaşlı insanlar daha kolay hasta olurlar.

Sıfır puan

Kod 0: Diğer yanıtlar.

- Böylece gribe yakalanmazlar.
- Onlar daha zayıftır.
- Onların gribe karşı savaşta yardıma ihtiyaçları vardır.

Kod 9: Boş.

ASİT YAĞMURU PUANLAMA 8.1

Tam puan

Kod 2: Duman çıkaran herhangi bir otomobil, fabrika atıkları, petrol ya da kömür gibi fosil yakıtların *yakılması*, yanardağlardan çıkan gazlar ya da benzer şeyler.

- Kömür ve gaz yakma.
- Fabrika ya da sanayi alanlarındaki kirlenmeden meydana gelen havadaki oksitler.
- Yanardağlar.
- Elektrik santrallerinden çıkan duman ["*Elektrik santrallerinin"* fosil yakıtları yakan elektrik santrallerini de içerdiği kabul edilir.]
- Kükürt ve azot içeren maddelerin yanması ile oluşurlar.

Kısmi puan

Kod 1: Kirliliğin doğru kaynaklarını kapsadığı kadar yanlış kaynaklarını da kapsayan yanıtlar

- Fosil yakıtları ve nükleer elektrik santralleri.[Nükleer elektrik santralleri asit yağmuru kaynağı değildir]
- Ozon'dan, atmosferden ve göktaşlarından dünyaya gelen oksitler. Aynı zamanda fosil yakıtlarının yanması
- " Kirlilikten" bahseden fakat asit yağmuruna anlamlı bir neden oluşturan kirlilik kaynağını vermeyen yanıtlar.
- Kirlilik
- Genel olarak çevre, yaşadığımız atmosfer, örneğin, kirlilik
- Gaz hâline çevirme, kirlilik, ateşler, sigara ["Gaz hâline çevirmenin" ne anlama geldiği açık değil, "ateşler" yeterince belirli değil, sigara içilmesi asit yağmurunun anlamlı bir nedeni değil]
- Nükleer elektrik santrallerindeki gibi kirlilik

Puanlama Not : Kod 1 için sadece "kirlilik"'ten bahsedilmesi yeterli.Bunun yanında verilecek herhangi bir örnek, sadece yanıtın Kod 2'yi hak edip etmediğine karar vermek için değerlendirilmelidir.

Sıfır puan

Kod 0: Diğer yanıtlar, "kirlilik"'ten bahsetmeyen *ve* asit yağmurunun anlamlı bir nedenini içermeyen yanıtlar da dahil olmak üzere.

- Plastiklerden yayılırlar.
- Havanın doğal bileşenleridir.
- Sigaralar.
- Kömür ve petrol (yeterince belirgin değil-yanmadan bahsetmiyor)
- Nükleer elektrik santralleri
- Endüstriyel atıklar. (yeterince belirgin değil)

Kod 9: Boş.

ASİT YAĞMURU PUANLAMA 8.2

Tam puan

Kod 1: A. 2,0 gramdan daha az

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

ASİT YAĞMURU PUANLAMA 8.3

Tam puan

Kod 2: Sirke ve mermer testi ile karşılaştırmak ve bu suretle tepkinin oluşması için

asidin(sirke) gerekli olduğunu göstermek.

- Yağmur suyu da asit yağmuru gibi bu tepkimeye neden olması için asidik olmak zorunda.
- Mermer parçalarındaki delikleri oluşturan diğer sebeplerin var olup olmadığını görme.
- Çünkü bu, su yansız olduğu için, mermer parçalarının herhangi bir sıvıyla tepkimeye girmediğini gösterir.

Kısmi puan

Kod 1: Sirke ve mermer testi ile karşılaştırmak için, fakat tepkimenin oluşması için asidin(sirke) gerekli olduğu açıkça gösterilmemiştir.

- Başka bir test tüpüyle karşılaştırmak
- Mermer parçalarının saf su içinde değişip değişmediğini görmek
- Öğrenciler bu basamağı, normal yağmurda kalan mermere ne olduğunu görmek için dahil etti.
- Çünkü damıtılmış su asit değildir.
- Kontrol etmek için.
- Normal su ve asidik su (sirke) arasındaki farkı görmek için

Sıfır puan

Kod 0: Diğer yanıtlar.

- Damıtılmış suyun bir asit olmadığını görmek.

Kod 9: Boş.

BEDEN EĞİTİMİ HAREKETLERİ PUANLAMA 9.1

Tam puan

Kod 1: Üçü de doğrudur: Evet, Hayır, Evet sırasıyla.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

BEDEN EĞİTİMİ HAREKETLERİ PUANLAMA 9.2

Tam puan

Kod 1: İkisi de doğrudur: Evet, Hayır sırasıyla.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

BEDEN EĞİTİMİ HAREKETLERİ PUANLAMA 9.3

Tam puan

Kod 11: Artan karbon dioksit seviyesini düşürmek **ve** vücudunuza daha çok oksijen sağlamak için.["Karbon dioksit" veya "Oksijen"'in yerine "Hava" kabul edilemez]

- Egzersiz yaptığınızda; vücudunuz daha fazla oksijene ihtiyaç duyar ve daha fazla karbon dioksit üretir. Nefes almak bunu gerçekleştirir.
- Hızlı nefes alıp verme, çok miktarda oksijenin kana geçmesini ve çok miktarda karbon dioksitin vücuttan atılmasını sağlar.

Kod 12: Artan karbon dioksit düzeyini vücudunuzdan atmak veya vücuda daha çok oksijen sağlamak, fakat ikisi birden değil. ["Karbon dioksit" veya "Oksijen"'in yerine "Hava" kabul edilemez]

- Çünkü oluşan karbon dioksitten kurtulmak zorundayız.
- Çünkü kasların oksijene ihtiyacı vardır. [Beden eğitimi yaparken (kaslarınızı kullanarak) vücudunuzun daha fazla oksijene gerek duyacağını belirtiyor]
- Çünkü beden eğitimi hareketleri oksijen harcar.
- Daha sık nefes alıp verirsiniz çünkü akciğerlerinize daha fazla oksijen alırsınız (Zayıf açıklama fakat çok fazla oksijen sağlandığı kabul ediliyor.)
- Çok fazla miktarda enerji kullandığınız için vücudunuz aldığı havanın iki veya üç katına gereksinim duyar. Aynı zamanda vücudunuzdaki karbon dioksiti atmaya gereksinim duyar. [2. cümle için Kod 12-Vücudunuzdan normal zamandan (daha fazla karbon dioksit atmak zorunda olduğunu içeriyor. Birinci cümle ikinci ile çelişkili değil; ama tek başına olsaydı Kod 01'lik olurdu.]

Sıfır puan

Kod 01: Diğer yanıtlar.

- Akciğere daha fazla hava almak
- Çünkü kaslar daha fazla enerji tüketir.(yeterince belirgin değil)
- Çünkü kalbiniz daha fazla çarpar.
- Vücudunuzun oksijene ihtiyacı vardır.(Daha fazla oksijene ihtiyacı olduğundan bahsetmiyor.)

Kod 99: Boş.

GENETİK YAPILARI DEĞİŞTİRİLEN TARIM ÜRÜNLERİ PUANLAMA 10.1

Tam puan

Kod 1: İkisi de doğrudur: Hayır, Evet sırasıyladır.

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş

GENETİK YAPILARI DEĞİŞTİRİLEN TARIM ÜRÜNLERİ PUANLAMA 10.2

Tam puan

Kod 1: D Mısırın değişik yetiştirme koşullarda nasıl büyüyeceğini görmek için

Sıfır puan

Kod 0: Diğer yanıtlar.

Kod 9: Boş.

Kaynak: Eğitim Araştırma Geliştirme Daire Başkanlığı (2006), *PISA Türkiye,* http://earged.meb.gov.tr/pisa/dil/tr/pisa2006.html

REFERENCES

Acar, O., 2008, PISA sonuçları ışığında Türkiye' nin rekabet gücünün değerlendirilmesi. Retrieved from internet November 10, 2008, http://www.tepav.org.tr/tur/admin/dosyabul/upload/Politika_notu_ozan_acar.pdf

American Educational Research Association (AERA)Standards for Educational and Psychological Testing, 1999, *Standards for educational and psychological testing.* WashingtonDC: American Educational Research Association.

Agin, M., 1974, Education for scientific Literacy: A conceptual frame of reference and some applications. *Science Education*, 58(3), 403-415.

Allalouf, A., R. K. Hambleton and S. G. Sireci, 1999, Identifying the causes of DIF in translated items. *Journal of Educational Measurement*, *36(3),* 185-198.

Angoff, W. H., 1988, Validity: An Evolving Concept. In H. Wainer and H. I. Braun(Ed.). Hillsdale*, Test Validity,* Vol.14*,* pp. 19-32. NJ: Lawrence Erlbaum.

Aşkar P. and S. Olkun, 2005, PISA 2003 Sonuçları Açısından Okullarda Bilgi ve İletişim Teknolojileri Kullanımı, Retrieved January, 21, 2008, http://www.ejer.com.tr/pdfler/tr/1315108165.pdf

American Psychological Association (APA), 1966, Standards for Educational and Psychological Tests and Manuals, Washington, DC: American Psychological Association.

António N. and T. Yariv-Mashal, 2006, Comparative research in education: a mode of governance or a historical journey? Comparative Education (electronic) Retrieved from internet on September, 19, 2008, http://www.cesnova.fcsh.unl.pt/DOCS/Novoa.pdf

Banerjee, J. and S. Luoma, 1997,Qualitative approaches to test validation. In Clapham, C. and Corson, D., editors*, Encyclopedia of language and education*. Vol. 7, pp.275-287, Language testing and assessment. Dordrecht: Kluwer Academic.

Baykal A., 2008, *Eğitimde Ölçme ve Aranan Nitelikler.* In Türkiye Özel Okullar Birliği, VII. Ortaöğretimde Yeni Arayışlar Sempozyumu (Ed), pp. 93-135, Antalya.

Beaton, A. E., T. N. Postlethwaite, K. N. Ross, D. Spearritt, and R. M. Wolf, 1999, *The benefits and limitations of international educational achievement studies*. Paris: International Institute for Educational Planning/UNESCO.

Benson, J., 1987, Detecting bias in affective scales. *Educational and Psychological Measurement,* 47, 55–67.

Berberoğlu G., 2004, Türk bakış açısından PISA araştırma sonuçları. Retrieved May, 13, 2008, from http://www.konrad.org.tr/Egitimturk/07girayberberoglu.pdf

Bingham W.V., 1937, *Aptitudes and Aptitude Testing*. New York, London : Pub. for the National Occupational Conference by Harper & Brothers

Birenbaum M., 2007, Evaluating the assessment: sources of evidence for quality assurance. *Studies in Educational Evaluation,* 33, 29–49.

Bodin, A., 2005, What does PISA really assess? What it doesn't A French view. Report prepared for Joint Finnish-French conference "Teaching mathematics: Beyond the PISA survey".

Bonnet, G., 2002, Reflections in a critical eye: on the pitfalls on international assessment. *Assessment in Education* 9, 387-400

Boyatzis, R., 1998, *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Brandwein, P. W., 2007, Science talent in the young expressed within ecologies of achievement. National Research Center on the Gifted and Talented, Retrieved January, 11, 2008, from http://www.gifted.uconn.edu/nrcgt/reports/rbdm9510/rdm9510.pdf

Burns, N. and S. K., Grove, 1993, *The Practice of Nursing Research: Conduct, Critique, and Utilization* (ed. 2). W.B. Saunders, Philadelphia.

Bybee R. W., 1997, *Toward An Understanding Of Scientific Literacy*, (In Advancing Standards for Science and Mathematics Education: Views From the Field). The American Association for the Advancement of Science, Washington, DC.

Cattell, R. B., 1946, Description and measurement of personality. New York: World Book Company.

Colvin, R. L., 1996, November 21, *Global study finds U.S. students weak in math*, Los Angeles Times.

Cronbach, L. J., and P. E. Meehl, 1955, Construct Valıdıty In Psychologıcal Tests, *Psychological Bulletin*, *52*, pp.281-302.

Çavaş, B., and T. Kesercioğlu, 2004, *Fen Eğitiminin Uygunluğu: Rose Projesi* , VI. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi, Marmara University, Atatürk Education Faculty, İstanbul.

Daly, J., A. Kellehear and M. Gliksman, 1997, *The public health researcher: A methodological approach*. Melbourne, Australia: Oxford University Press.

Dindar, H., 2000, Gazi Üniversitesi, Gazi Eğitim Fakültesi Derg., 20*(1),* 145-148, Ankara

Dohn, N. B., 2007, Knowledge and skills for PISA – assessing the assessment. *Journal of Philosophy of Education*, *41 (1)*, 1-16.

Dorans, N. J., and P.W. Holland, 1993, DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* , pp. 35-66. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Downing, S. M., and T. M. Haladyna, 1997, Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education,10,* 61–82.

Downing, S. M., and T. M. Haladyna, (Eds.).,2006, *Handbook on test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Durant, J. R., 1993, What is scientific literacy? In J. R. Durant & J. Gregory (Eds.), Science and culture in Europe, pp. 129–137, London: Science Museum.

Egelund, N., 2008, The value of international comparative studies of achievement - a Danish perspective. *Assessment in Education: Principles, Policy & Practice*, 15*(3),* pp. 245 – 251.

Ercikan K., 2002b, Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing,* 5*(1),* 23-35.

Ercikan, K. and T. McCreith, 2002, *Effects of Adaptations on Comparability of Test Items and Test Scores*. In D. Robitaille & A. Beaton (Eds.) Secondary Analysis of the TIMSS Results; A Synthesis of Current Research, pp. 391-407, Dordrecht, the Netherlands, Kluwer Academic Publishers.

Ercikan, K. , 2006, *Developments in Assessment of Student Learning and Achievement*. In P.A. Alexander and P. H. Winne (Eds.), American Psychological Association, Division 15, Handbook of Educational Psychology, 2nd edition, pp. 929-953, Lawrence Erlbaum Associates.

Ercikan, K. and N. Alper, 2008, *Cultural Studies of Science Eduction*, Adaptation of instructional materials: a commentary on the research on adaptations of Who Polluted the Potomac. Received February 8, 2008, from Sprininger Science+Business Media.com

Eşme. İ., 2008, January 9. *PISA 2006 sonuçları ve Türkiye'de fen eğitimi*, Radikal, p.24

Ferrara, S., 2007, *Educational Measurement: Issues and Practice Toward a Psychology of Large-Scale Educational Achievement Testing: Some Features and Capabilities*, Springer London.

Fertig, M., 2003, Who is to blame? The determinants of German students' achievement in the PISA 2000 study. Bonn, Institute for the study of labor**:** Retrieved March, 16, 2008, from http://www.iza.org/publications/dps/

Freouire, P., 2006, *Working with Qualitative Variables*. From Product Description to Cost: A Practical Approach, Building a Specific Model, Vol.2, pp. 117-124.

Frey B.B., S. Petersen , L. M. Edwards, J. T. Pedrotti and V. Peyton, 2007, Item-Writing Rules: Collective Wisdom. March, 16, 2008 from http://people.ku.edu/~bfrey/itemwritingrules.pdf

Gipps, C., and P. Murphy, 1994, *A fair test? Assessment, achievement and equity*, Buckingham: Open University Press

Goodwin, L. D. and N. L. Leech, 2003, The meaning of validity in the new standards for educational and psychological testing: implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36(3),* 181-91

Goldstein, H., 2004, International comparisons of student attainment: some issues arising from the PISA study. In: Assessment in Education – Principles, Policy, and Practice(Ed), pp.319-330.

Grisay, A., 2003, Translation procedures in OECD/PISA 2000 international assessment, *Language Testing,* 20*(2),* 225-240.

Guilford, J. P., 1954, *Psychometric Methods*, New Delhi : Tata McGraw-Hill Publishing

Güven, S., 2001, Sınıf öğretmenlerin Ölçme ve Değerlendirmede Kullandıkları Yöntem ve Tekniklerin Belirlenmesi, *X.Ulusal Eğitim Bilimleri Kongresi Bildiri Kitabı,* pp. 413-423, Bolu İzzet Baysal Üniversitesi Eğitim Fakültesi, Bolu.

Hofmann, T.S., 2006, PISA According to PISA – Does PISA Keep What It Promises? Retrieved April, 25, 2008, from http://www.univie.ac.at/pisaaccordingtopisa/introduction_pisaaccordingtopisa.pdf

Hunter, J. E. and F. L. Schmidt, 1990, *Methods of meta-analysis: Correcting error and bias in research findings*. Newsbury Park: Sage Publications.

Kellegan, T. and V.Greaney, 2001, *Using Assessment to Improvethe Quality of Education*, Paris: UNESCO IJEP.

Kirsch, I., J.D. Long, D. Lafontaine and J. McQueen, 2002, *Reading for Change: performance and engagement across countries.* Paris, OECD

Knain, E. and A. Turmo, 2003, Self-regulated learning Lie S et al (Eds) *Northern Lights on PISA*, University of Oslo, Norway, pp. 101-112.

Konak A., 2008, Merhaba. *Cito Eğitim Kuram ve Uygulama Dergisi.*

Kress, G., 2003, *Literacy in the new media age*, London: Routledge

Kunter, M., and J. Baumert, 2006, Linking TIMSS to research on learning and instruction: A reanalysis of the German TIMSS and TIMSS video data. In S. J. Howie & T. Plomp (Eds.), Learning mathematics and science: Lessons learned from TIMSS, pp. 335–351, London: Routledge.

Kuzel, A.J., and R. C.Like, 1991, *Standards of trustworthiness for qualitative studies in primary care*. In P.G. Norton, M. Stewart, F. Tudiver, M.J. Bass, and E.V. Dunn (eds) Primary Care Research: Traditional and Innovative Approaches. Newbury Park, California: Sage, pp.138-158.

Lange J., 2006, PISA: promises, problems and possibilities, What are PISA and TIMSS? What do they tell us?. Retrieved July, 24, 2009, from http://www.icm2006.org/proceedings/Vol_III/contents/ICM_Vol_3_80.pdf

Laugksch, R. C., and P. E. Spargo, 1996, Construction of a paper-and-pencil Test of Basic Scientific Literacy based on selected literacy goals recommended by the American Association for the Advancement of Science. *Public Understanding of Science, 5(4)*, 331–359.

Lemke, J. L., 1990, *Talking science: Language, learning, and values*. Norwood, NJ: Ablex Publishing Corporation.

Lingens, H., 2005, PISA in Germany: A Search for Causes and Evolving Answers. *International Handbook on Globalisation, Education and Policy Research Global Pedagogies and Policies*, Vol.3, Springer Netherlands

Linn, R.L., 2002, The measurement of student achievement in international studies. In Porter, A.C., Gamoran A. (Eds.), *Methodological advances in cross-national surveys of educational achievement*, pp. 27–57, Washington, DC: National Academy Press.

Mayer, D. P., 1999, Measuring instructional practice: Can policy makers trust survey data? *Educational Evaluation and Policy Analysis, 21(1),* 29–45.

Mehrens, W. A., 1997, The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16(2)*, 16-18.

Messick, S., 1989, Validity. In R.L. Linn (Ed.), in *Educational Measurement* (3rd. edition, page 13-103). New York: American Council on Education and Macmillan.

Messick, S., 1990, *Validity of test interpretation and use* (Rep. No. ETS-RR-90-11). Princeton, NJ: Educational Testing Service.

Messick, S., 1994, The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-24.

Miles, M. B. and A. M. Huberman, 1994, *Qualitative Data Analysis*, 2 (Ed.), Thousand Oaks, CA: Sage Publications.

Ministry of Education (MEB). Yeni ilköğretim müfredatı tanıtımı. Retrieved November 19, 2008, from http://digm.meb.gov.tr/uaorgutler/OECD.doc

Ministry of Education (MEB), Orta Öğretim Programları Geliştirme, Retrieved July, 22, 2009, from http://oop.meb.gov.tr/tr/index.php?view=article&catid=67%3Aorta-oeretim-programlar-gelitirme&id=210%3Aorta-oeretim-programlar-gelitirme&format=pdf&option=com_content&Itemid=31

Mullis, I. V. S. and M. O. Martin, 2006, TIMSS in perspective: lessons learned from IEA's four decades of international mathematics assessments. IEA New Online. Retrieved November, 22, 2008, from http://www.iea.nl/fileadmin/user_upload/IRC2006/Brookings_Institution_Program/Mullis___Martin.pdf

Mullis, I.V.S., M.O. Martin, E.J. Gonzales, K.D.,Gregory, R.A, Garden, R.A.,K.M O'Connor, S.J. Chrostowski and T.A Smith, 2000, *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Murat, F. and Rocher T, 2004, *On the Methods used for international assessments of educational competencies*, In J.H. Moskowitz & M. Stephens (Eds.), Comparing Learning outcomes, pp.197-210, London; Routledge Farmer.

Nardi E., 2008, Cultural biases: a non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice 15 (3)*, 259–266.

Nóvoa, A. and T. Yariv-Marshal, 2003, Comparative Research in Education: a mode of governance or a historical journey? *Comparative Education, 39(4)*, 423-438

OECD, 1999, Measuring Student Knowledge and Skills. A New Framework for Assessment, OECD, Paris

OECD, 2001, Knowledge and Skills for Life. The first Results from PISA 2000. Organization for Economic Co-operation and Development.

OECD, 2002, Reading for Change. Performance and Engagement across Countries. Results from PISA 2000, OECD, Paris

OECD, 2003, The PISA Assessment Framework. Mathematics, Reading, Science and Problem Solving Knowledge and Skills, OECD, Paris.

OECD, 2007, Assessing Scientific Reading and Mathematical Literacy. A Framework for PISA 2006, OECD, Paris.

OECD, 2008, The Story so Far; PISA 2000-2006. Retrieved September 20, 2009, from http://www.oecd.org/dataoecd/51/27/37474503.pdf

O'Halloran, K. L., 2000, Classroom discourse in mathematics: A multisemiotic analysis. *Linguistics and Education, 10(3)*, 359-388.

Oliver M., 2005, 'Automatic Identification of English Multi-Word Units', In: C. Cosme, C. Gouverneur, F. Meunier (eds), *Phraseology 2005: The many faces of Phraseolog,* 261-264.

Olivery, E. M., 2007, *Analysis of Construct Comparability in the Program For Internatıonal Student Assessment, Problem-Solving Measure*, Unpublished doctoral dissertation, The Unıversity of British Columbia, Canada.

Osterlind, S. J., 1983, *Test item bias*. Newbury Park: Sage Publications.

Osterlind, S. J., 1998, *Constructing test items: multiple-choice, constructed-response, performance, and other formats.* (2ed.). Boston, MA: Kluwer Academic Publishers.

Owen, E., 2001, Educational indicators. In Leimu, K., Linnakyla P., RVI L. (Eds.), *Merging national and international interests in educational system evaluation.* pp. 41–50. Jyva¨skyla¨, Finland. University of Jyva¨skyla¨, Institute for Educational Research.

Patton, M. Q., 1990, *Qualitative evaluation and research methods,* SAGE Publications, Newbury Park London New Delhi.

Popham, J. W., 1997, What's wrong and what's right with rubric. *Educational Leadership. 55 (2)*, 12.

Pope, C., S. Ziebland and N. May, 1990, Qualitative research at Health care: Analysis Qualitative Data. *British Medical Journal*, 320, pp.114-116.

Porter, A. C., 2002, Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31(7)*, 3–14.

Prais, S., 2003, Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education,* 29, 139-163.

Prais, S.J., 2007, England: Poor survey response and no sampling of teaching groups *Oxford Review of Education* Vol. 33, pp.134-144.

Psalidas, A., C. Apostolopoulos and V. Hatzinikita, 2007, Investigating Factors Affecting students' performance to PISA Science items, *International Journal of Engineering Science and Technology Review 1*, 90-97.

Rice, P., and D. Ezzy, 1999, *Qualitative research methods: A health focus*. Melbourne: Oxford University Press.

Robitaille, D.F. and R.A. Garden, 1989, *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics.* Oxford: Pergamon Press.

Rochex J. H., 2006, Social, Methodological, and Theoretical Issues Regarding Assessment Analysis of PISA 2000 Literacy Tests Retrieved May 12, 2009, from http://rre.sagepub.com/cgi/reprint/30/1/163

Romainville, M., 2002, On the appropriate use of PISA. La Revue Nouvelle (March-April 2002).

Sartori, M., Pasini, M., 2007, Quality and Quantity in Test Validity: How Can We Be Sure That Psychological Tests Measure What They Have to? *Qual. Quant. 41*, 359–374.

Savran Z. N., 2004, PISA Projesinin Türk eğitim Sistemi Açısından Değerlendirilmesi, Retrieved July, 28, 2008, from http://www.tebd.gazi.edu.tr/arsiv/2004_cilt2/sayi_4/397-412.pdf

Schwab, D.P., 1980, Construct validity in organizational behavior. *Res. Organizational Behavior,* 2, 3–43.

Shamos, M. H., 1995, *The myth of scientific literacy.* New Brunswick, NJ: Rutgers University Press.

Shepard, L. A., 1987, The case for bias in tests of achievement and scholastic aptitude. In Modgil, S. & Modgil, C. (Eds.), Arthur Jensen: Consensus and Controversy. London: Falmer Press.

Shepard, L.A., 1997. The centrality of test use and consequences for test validity, *Educational Measurement: Issues and Practice. 16(2)*, 13, 24.

Simola, H., 2005, The Finnish Miracle of PISA: Histoical and Sociological remarks on techaing and teacher education, *Comparative Education*, *,41(4),* 455-470.

Sipps, G. J., and R. A. Alexander, 1987, The multifactorial nature of extraversion introversion in the Myers–Briggs Type Indicator and Eysenck Personality Inventory. *Educational and Psychological Measurement, 48*, 445–451

Sireci S. G., 1997, Problems and issues in linking assessment across languages. *Educational Measurement: Issues and Practice, 16(1),* 12- 20.

Sireci S. G., 2004, Evaluating Construct Equivalence, *Educational Measurement: Issues and Practice. 15(4)*, 57-60.

Sjøberg, S., 2007, Pupils' experiences and interests relating to science and technology: Some results from a comparative study in 21 countries. Retrieved May, 2009, from http://folk.uio.no/sveinsj/SLOC%20Sjoberg%20paper.pdf

Stedman, L.C., 1997. International achievement differences: An assessment of a new perspective. *Educational Researcher, 26(3),* 4–15.

Strauss, A. and J. Corbin, 2000, *Basics of Qualitative Research*, Newbury Park:Sage.

Şahin İ., 2007, Assessment of New Turkish Curriculum for Grade 1 to 5. *Elementary Education Online*, 6*(2),* 284-304

Teddlie, C., and A. Tashakkori, 2003, *Major issues and controversies in the use of mixed methods in the social and behavioral sciences*, In A. Tashakkori and Tedlie (eds) Handbook of Mixed methods in Socia and Behavioral research. Thousands Oaks,CA: Sage.

TEMPO DERGİSİ, 2005, Türkiye; PISA sonuçları. Retrieved August 21, 2008, from http://74.125.77.132/search?q=cache:JOOS0kEDMyMJ:tempodergisi.com.tr/toplum_politika/07307+M%C3%BCfredat+de%C4%9Fi%C5%9Fikli%C4%9Fi+ve+PISA+sonu%C3%A7lar%C4%B1&hl=tr&ct=clnk&cd=10&gl=tr

Thomas, J., 1993*, Doing critical ethnography.* Newbury Park, CA: Sage

Verschaffel, L., E. De Corte and I. Borghart, 1997, Pre-service teachers' conceptions and beliefs about the role of real-world knowledge in mathematical modeling of school world problems. *Learning and Instruction*, *7*, 339-359.

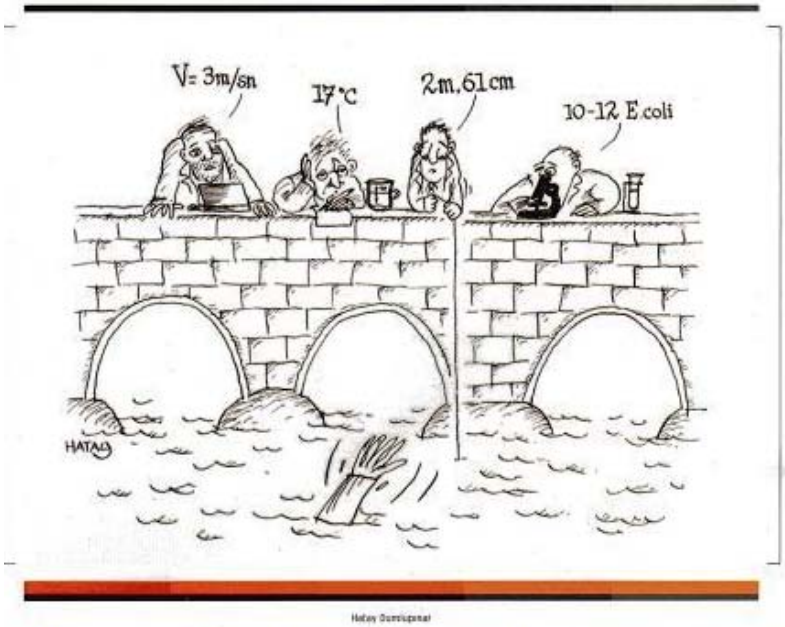
Visser L. Y., 2007, Developing the scientific disposition In formal learning contexts: Applications of problem-oriented Learning, Second Advanced International Colloquium on Building the Scientific Mind (BtSM2007) Vancouver, Canada. Retrieved July 30, 2008 from http://www.learndev.org/dl/BtSM2007/YusraVisser.pdf

Wuttke, J., 2003, Uncertainties and Bias in PISA, Retrieved May 21, 2008, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1159042

Wuttke, J., 2007, PISA according to PISA. Does PISA keep what it promises? In Hopmann, Brinek, Retzl (ed), pp. 241-263, Wien.

Yıldırım, A. and H. Şimşek, 2005, *Sosyal bilimlerde nitel araştırma yöntemleri (5.ed)* , Ankara: Seçkin Yayıncılık.

"Ölçüm ve Günlük Hayatımız"



Dumlupınar, Hatay. (2007). Bülten. TÜBİTAK, Haziran, 2007, Sayı:66, s.20.