DESIGNING SYSTEM BASED ENVIRONMENTAL INSTRUCTION PROGRAM AND EVALUATING ITS EFFECTS ON SEVENTH GRADE STUDENTS

by

Zerrin Doğança

B.S. in Primary Mathematics and Science Education, Boğaziçi University, 2003

M.S. in Environmental Sciences, Boğaziçi Universty, 2007

Submitted to the Institute of Environmental Sciences in partial fulfillment of the requirements for the degree of

Doctor

of

Philosophy

Boğaziçi University

2013

Dedicated to my dear
grandmother Aynişah Güler,
whom I miss so much.

ACKNOWLEDGEMENTS

Writing dissertation was a long journey and a number of people supported me during this challenging and informative journey.

First of all, I would like to thank to my advisor, Assoc. Prof. Ali Kerem Saysel who is very patient, hardworking, and most importantly modest to learn anything about the field science education. I really love to come to his office and make discussions every single week and I hope this academic relationship will continue with several projects and articles.

I would like to express my sincere gratitude to Prof. Dr. Emine Erktin and Assist. Prof. Ebru Z. Muğaloğlu. They are the reasons to choose an academic path for my life. Prof. Dr. Emine Erktin always supported me during this study both academically and emotionally. And, Assist. Prof. Ebru Muğaloğlu assisted me, gave brilliant and practical ideas that helped me to complete this study.

I also owe thanks my jury members Prof. Dr. Yaman Barlas and Prof. Dr. Miray Bekbölet. Their extensive feedback and multidisciplinary point of view have contributed greatly to the quality of this study. I also wish to express my appreciation to Prof. Dr. Esra Macaroğlu for her careful review of the thesis and for sharing her expertise.

This was a field study and a number of people helped me to carry out and complete this research. I owe thanks to Ezgi Talum and Sinem Arslanoğlu to support me about everything in class and at school. They are very nice teachers that motivated my students throughout the study. And, I want to thank Science Teacher Huriye Çelik and School Principal Erhan Ziya Sancar for their endless trust and support to carry on my study at their school.

As a member of Primary Education Department, I feel very lucky to work with precious instructors and friends. Firstly, I want to thank to my friends Serkan Özel and Engin Ader. I really like our conversations and spending time with you. And, my office mates; Işık Sabırlı and Gürsu Aşık. I do not know how I can leave you. I cannot imagine any other people whom I even look forward to eat lunch. With all my heart, I hope to work with you some time in future.

My dear friend Müge Ataman, we have been through a lot and I hope we will experience a lot more together throughout our lives.

I would like to thank my brother Erdal Doğança and my father Murat Doğança who always believe in me and tolerate me while I was feeling stress because of this study. Most importantly, I wish to express my heartfelt thanks my closest friend and my mother Ayşegül Doğança. She is always by my side and encourages going on.

Lastly, I want to express that this study was supported financially by Boğaziçi University Fund with the project number 5729D.

ABSTRACT

The present research aimed to test the effects of an environmental instructional design with systems approach on seventh grade students. The main focus was to examine whether systems approach is a more effective way to teach environmental issues that are dynamic and complex. The research was a quasi-experimental study that enabled to compare performances on general systems thinking skills, competence in dynamic environmental problem solving, and success in standard science achievement tests of subjects from different groups. The sample of the study included 42 seventh grade students (12-14 year old). The same pre, post, and delayed tests were applied to both groups. The control group was taught according to the standard unit plan suggested by the Ministry of Education, while the experimental group was taught the same content with activities including, feedback loops, stock and flow diagrams, behavior over time graphs, and computer modeling. It was found that after one month of systems based environmental instruction, the experimental group performed better on systems thinking skills and dynamic environmental scenarios (DES) tests at .05 significance level. Besides, the effects of the system based intervention were more enduring on performance on DES test for the experimental group, when delayed tests were taken into account. No significance difference was found on science achievement level between the two groups. In addition to quantitative results, interviews resulted in higher levels of feedback thinking skills of the selected respondents from the experimental group.

ÖZET

Bu araştırma, sistem yaklaşımı ile hazırlanan bir çevre eğitimi tasarımının yedinci sınıf öğrencileri üzerindeki etkisini sınamayı amaçlamıştır. Temel olarak, dinamik ve karmaşık çevre konularının sistem yaklaşımıyla daha etkin öğretilip öğretilmeyeceği araştırmak hedeflenmiştir. Çalışmanın yarı-deneysel tasarımı, farklı gruplarda yer alan öğrencilerin genel sistem düşüncesi becerilerindeki, dinamik çevre problemlerini anlamada yetkinliklerindeki ve standart fen testi başarılarındaki olası farklılıkları izlemeye elverişlidir. Çalışmanın örneklemini 12-14 yaşında, aynı devlet okulunda okuyan 42 yedinci sınıf öğrencisi oluşturmaktadır. Aynı ön, son ve gecikmeli son testler tüm katılımcılara uygulanmıştır. Karşılaştırma grubuna, müfredatın önerisi doğrultusunda öğretim yapılırken, deney grubuna aynı çevre içeriği geri besleme döngüleri, stok-akış şemaları, davranış-zaman grafikleri ve dinamik modellerle öğretim yapılmıştır. Bir aylık öğretim sonucunda, Sistem Beceri Testi (SBT) ve Dinamik Çevre Senaryoları Testi'nde (DÇS) iki grup arasındaki farkın.05 düzeyinde manidar olduğu bulunmuştur. Dahası gecikmeli son-test sonuçları incelediğinde, sistem yaklaşımıyla yapılan öğretimin etkisinin deney grubunun DÇS testi performası üzerinde hala etkili olduğu saptanmıştır. Ancak, iki grup arasında var olan SBT testindeki performans farkı, altı ay sonunda yok olmuştur. Nitel sonuçlar incelendiğinde, iki gruptan rasgele seçilen katılımcılardan deney grubunda olanların geri besleme düşüncesi ile ilgili sorulara daha üst seviyede cevaplar verdikleri saptanmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

Symbol/ Abbrevation	Explanation
α	Cronbach alpha coefficient
_del	Delayed test
_exp	Experimental group
_pre	Pre test
_post	Post test
_q	Question
_sug	Suggestion
BOT	Behavior over time
CL	Causal loop question
CLD	Causal loop diagram
DEL	Estimating delay question
DES	Dynamic environmental scenarios
FB:	Feedback thinking
Math_gra:	Mathematics grade
SAT	Science achievement test
Sci_gra	Science grade
SD	Standard deviation
SF	Stock-flow thinking
Sig.	Significance
STELLA	Systems Thinking for Education and Research Software
STRS	Systems thinking required skills
STS	Systems thinking skills

1. INTRODUCTION

Most of today's global problems are dynamic and complex in nature. That is to say, structures of the environment problems change constantly in time and the variables of these problems are nested. Moreover, actions taken for prevention or stabilization of these environmental problems generally create more severe consequences. For instance, introduction of an alien species to fight with pests results in invasion the alien species. Understanding the system structure and dynamic behavior patterns of environmental problems are related to development of systems thinking skills. Sweeney and Sterman (2000) define systems thinking skills as

- identifying stocks (accumulations) and flows (their rates of change),
- identifying delays and estimating their possible effects on a system,
- identifying feedback loops for observed behaviors of a system,
- identifying nonlinearities,
- defining boundaries of both mental and formal models,
- reasoning certain behavior patterns of a system due to interactions of different aspects.

Our natural environment and each ecosystem consist of stocks, flows (e.g. the food webs), delays (e.g. the bioaccumulation of toxic chemicals), feedbacks (e.g: the carbon cycle) and nonlinearities (e.g: species growth in a limited environment). Capra (1998) introduces a new concept- eco-literacy; which means "understanding basic principles of organization of ecosystems and using those principles for creating sustainable human communities" (Capra, 1998; p.3). Stone and Barlow (2005) also explain that "Nature sustains life by creating networks." (p.3). That is to say, ecosystems are systems with various inner and inter-connected parts and people should understand these complex and inter-related natural networks for a sustainable living. In this sense, education gains importance. Education for sustainable living aims to bridge natural and artificial design and sustain nature (Capra, 2005). There are a number of studies and educational practices going on in the name of environmental education, education for sustainability (Armstrong and Impara, 1992; Hungerford and Volt, 1990; Hsu and Roth, 1996; Lane, et. al, 1996;

McKeown-Ice, 2000; Ostman and Parker, 1987; Oweini and Houri; 2006; Plevyak, et. al, 2001; Powers, 2004; Reid and Sa'di, 1997; Şahin, et. al, 2004; Tozlu, 1996; Tuncer, et al. 2005; Wilke, 1985; Yılmaz, et. al., 2002; Zak, 2005; cited in Doğança, 2007). However, only a few numbers of studies address development of systems thinking skills for environmental sustainability education (Grotzer and Basca, 2003; Assaraf and Orion, 2005; Riess and Mischo, 2010).

One of the limitations of the current education system is heavy load of factual knowledge with limited and inadequate connections between fragments of knowledge (Brown, 1992). Another drawback is related to limited school curricula in terms of ecological content (Grotzer and Basca, 2003). So, less time is devoted to teaching ecological subjects that impede formation of understanding dynamic natural systems. Moreover, in most of the cases, teachers think that ecological subjects are simple to understand for students; however Grotzer and Basca (2003) mention several studies on misconceptions about ecological content that students at different grade levels have.

Taking these limitations into account, this research aimed to design an alternative instructional plan for "Human and Environment" science unit with systems approach and to study effects of the systems based intervention on subjects' general systems thinking skills, competence in dynamic environmental problem solving, and achievement in standard science achievement tests. The study included development of specific tests to measure systems thinking skills (STS test) and understanding of dynamic structures of environmental issues (DES test) and a codebook for a semi-structured interview. In addition to development of various instruments, the study included development of various Turkish instructional materials to teach basics of systems dynamics and some dynamic environmental issues.

2. LITERATURE REVIEW

2.1. Systems Thinking and Systems Approach

2.1.1. Understanding of Systems and Systems Thinking

"System" is a collection of interrelated elements that function as a whole (Assaraf & Orion, 2003). Hence, concepts of interrelatedness and cooperation come into prominence within a system. Ackoff (1994) proposes a similar definition: "A system is a whole that cannot be divided into independent parts or subgroups of parts." (p.175). By referring to these two definitions of systems, it can be asserted that systems thinking involves identifying relevant elements and understanding their interconnections within a pre-determined boundary.

To gain a deeper insight on systems thinking, it is appropriate to further examine its definitions in system dynamics literature. Jay Forrester (1994), the founder of the field of system dynamics, defines systems thinking as an approach to identify relevant elements of a system. He argues that system thinking goes beyond emphasizing existence and importance of systems. In his opinion, systems thinking involves "general and superficial awareness of systems" (p. 251). He notes awareness-rising feature of systems thinking and believes that it is possible to gain deeper understanding of complex problems of today's world by means of systems thinking.

Mandinach and Cline (1994) refer systems thinking as a problem-solving strategy that deals with changing components of a dynamic system with the help of models and simulations. In contrast to "laundry list thinking" that is; listing various variables that address a specific issue and creating unidirectional causal relationships, systems thinking presents a circular picture of a complex situation and aims at explaining dynamic behaviors created by a system.

2.1.2. Systems Thinking as a Discipline and Its Practices and Principles

Senge (1990) calls systems thinking as the "fifth discipline". He develops a framework to define and understand how learning organizations should "*shift their minds*" to be able to see interrelationships and changing patterns rather than focusing on "*static snapshots*". Apart from the disciplines of personal mastery, mental models, shared vision, and team learning, systems thinking is presented as a cornerstone that emphasizes all the elements of the developed framework and aims to decrease learning disabilities in organizations.

Senge (1990) examines each discipline in his framework on three levels; namely practices, principles, and essences. Figure 2.1 illustrates different levels and components of each level for systems thinking. Practices are simply conscious efforts performed to gain experience about the "discipline".

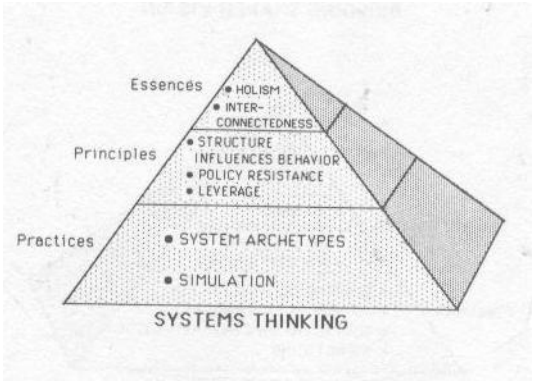

Figure 2.1. Levels of systems thinking proposed by Senge (1990)

Simulations and system archetypes are mentioned as practices in Figure 2.1. Barlas (2002) defines simulations as "a step-by-step operation of the model structure over compressed time" (p.5). Simulations have various functions in different fields. For instance, Sterman (1994) explains simulations as tools that enable users to make experiments on decision-making skills or "refresh" these skills, while Feurzeig and Roberts (1999) refer to simulations as learning and teaching tools that enable users to gain experience in a controlled, relatively simpler world.

System archetypes are generic structures of common behavior patterns. Principles include theories that imply related practices. Senge (1990) compiles common behavior patterns in the dynamic world that we live in and classify them under these titles:

- Balancing process with delay
- Limits to growth
- Shifting the burden
- Eroding goals
- Escalation
- Success to the successful
- Tragedy of the commons
- Fixes that fail
- Growth and underinvestment

Senge (1990) explains each archetype in detail with a focus on learning organizations in his book- The Fifth Discipline. Apart from his explanations, one can relate these archetypes to current global environmental problems. For example, "tragedy of the commons archetype" is typically related to management of natural resources of our planet. At first, people were able to use natural resources without any limitations and got gains. As time passed, resources started to decrease, hence people preferred to intensify their activities to find and consume these resources. Depletion of resources is increased while individual benefits decrease. For instance, The Amazon Rainforest constitutes 7 % of land, is home for 50 % of the species in the Earth and provides 20 % of the earth's oxygen. Now, it is observed that more than 20 % of the Rainforest has been destroyed and it is estimated that the whole forest will disappear within 50 years if no conservation activities take place (Rainforest Alliance, 2013). This situation is an example of the system archetype- "tragedy of the commons". Senge (1990) suggests some managerial practices like educating people who consume the resources or setting regulations preferably by the stakeholders.

In Figure 2.1.; “Structure influences behavior”, “policy resistance” and “leverage” are given as principles. Barlas (2002) defines structure as set of all relationships between variables of a system. Hence, a behavior pattern is expected to be generated, when structure of a system functions over a period of time, Growth, decline, s-shaped growth, growth and decline, and oscillations are generic categories of most dynamic behaviors that are created by certain structures (Barlas, 2002).

Policy resistance is related to dramatic and worse results and unexpected side effects due to attempts of stabilizing a system (Sterman, 2000). Forrester (1971) terms this phenomenon as “counterintuitive behavior of social systems” while Meadows (1982) mentions about policy resistance and delayed interventions due to unexpected results of stabilization attempts (cited in Sterman, 2000). Sterman (2000) gives the example of banning birth control practices in Romania during the Ceausesçu Regime. The government intended to increase birth rate in the country to ensure ethnic identity and national pride. The first years of the prohibition, the birth rate increased sharply; but a sudden decrease was observed in the fourth year of the application and the fall continued throughout the following years. The reasons of the fall in spite of the prohibition were alternative and unhealthy abortion practices, increase in infant and neonatal mortality rate, financial problems of families with higher number of children. The dramatic result of the birth rate policy was almost the same low birth rate in the country with more children living in orphanages with severe conditions.

Leverage, the last principle in the Levels of Systems Thinking pyramid (Figure 2.1), is related to small, well-focused actions to endure improvements rather than obvious actions that work in the short run. High leverage actions are so non-obvious that their effects can only be observed in a long period of time or in a remote place (Senge, 1990). For instance, to solve the problem of hunger in Africa, there are lots of campaigns going on to send food and medicine to the continent, which seems the most obvious solution for the hunger problem. However, the problem has not been solved for years. To invest money on African market would be a high leverage action in this circumstance.

Essences are placed at the top of the Senge's systems thinking pyramid (Figure.2.1). They are experienced naturally in mastery of a discipline. For example, the ability to recognize interconnections in all aspects of life is developed by time, as one gains experience in systems thinking.

2.1.3. Systems Thinking and Dynamic Complexity

In addition to identifying elements and their interconnections within a system, some scientists include dynamic complexity in their systems thinking descriptions. For instance, Sweeney and Sterman (2000) mention about representation and assessment of dynamic complexity in a system by using words and graphs as one of the aspects of systems thinking. Riess and Mischo (2010) also express systems thinking as "the ability to recognize, describe and model complex aspects of reality as systems" (p.707). Moreover, they also focus on awareness of time dimension to model and to make projections for future behaviors of a system.

Sterman (1994) also focuses on dynamic complexity of the world that we live in and points to learning difficulties of people in this dynamically complex environment. Dynamic complexity involves:

2.1.3.1. Stocks. Stocks are simply accumulations within a system and are only changed via their flows. However, they have inertia. In other words, stocks accumulate in a relatively long time and their values do not change suddenly (Barlas, 2002). Hence, it becomes harder to predict behavior of a stock. Sterman and Sweeney (2002) made a research with highly educated group of people on their conceptualization about global warming and CO_2 emissions. There was a common tendency among subjects that average global temperature responds to variations in CO_2 emissions immediately, hence extreme changes in emissions cause a sudden peak or decline in temperature. Another study conducted by Moxnes and Saysel (2009) also reveals improper mental models of the subjects about CO_2 accumulation. They developed a simulation where subjects should control total global emissions of CO_2 to a target for CO_2 stock in the atmosphere. It is found that people tend to overshoot the target for the stock.

2.1.3.2. Time-delays. As the duration between taking action and observing its effects within a system becomes longer, it becomes harder to understand the outcome (Sterman, 2000). For instance, plastic wastes are not decomposed as they are disposed. Indeed, there is a long time delay between disposal (inflow) and decomposition (outflow) of plastic wastes. The continuous inflow of plastic wastes leads to an accumulation of plastic wastes due to delay. Hence, people should develop certain strategies to deal with discharges of plastic wastes.

2.1.3.3. Feedback loops. A problem becomes more complex as number of feedback loops increases or as number of interconnections between variables within a system increases. Barlas (2002) emphasizes the difficulty to construct a mental model that is capable of organizing and working on a number of feedback loops concurrently; to predict dynamic behavior of a system with various feedback loops. For instance, a cat yowls at a doorway and the person at the house feeds the cat. Now, there is available food for cat and the cat stops yowling. But, since there is available food for the cat at the doorway, the number of cats increases and they start yowling to get more food. This relatively simple occasion includes dynamic complexity due to existence of more than one feedback loop, since increasing number of cats is not an intended outcome for the situation.

2.1.3.4. Nonlinearities. In dynamic systems, there are usually a variety of factors that affect decision making and these factors do not have a linear relationship with the stock. This leads to an uncorrelated behavior of stock and its flows, which is difficult to project with a static mental model in mind (Sterman, 2000). Another ecological example about dynamic complexity can be about interrelated relationships of organisms within an ecosystem. If a predator species becomes extinct in an ecosystem, one cannot expect a linear increase in the related prey population.

After explaining aspects of dynamic complexity in detail, specific systems thinking skills can be listed from Sweeney and Sterman's perspective (2000). It should be re-emphasized that they define systems thinking as "the ability to represent and assess dynamic complexity both textually and graphically" (p.250). Based on the definition that includes dynamic complexity, they suggest systems thinking skills as:

- Identifying stock and flow relationships,

- Identifying delays and estimating their possible effects on a system,
- Identifying feedback loops for observed behaviors of a system,
- Identifying nonlinearities,
- Defining boundaries of both mental and formal models,
- Reasoning certain behavior patterns of a system due to interactions of different aspects of dynamic complexity.

The next section is related to systemic understanding of life and ecoliteracy. The section includes issues about dynamic complexity within nature, systems thinking skills related to ecological concepts, and structure of environmental sustainability problems.

2.2. Systemic Understanding of Life and Ecoliteracy

Global problems cannot be understood in isolation, because they are interdependent and interconnected. In other words, they are systemic problems. According to Capra (1997), humanity is facing the challenge to deal with global problems and they should change their thinking styles from analytical to contextual thinking. Although traditional view of science supports dividing things into parts and then measuring and quantifying them (i.e. analyzing parts of a whole), contextual thinking does not deal with parts but emphasizes "understanding the context of the whole" (Capra, 2005; p.21). Capra (1997) proposes a shift from analytical to contextual and eventually to environmental thinking. He explains "environmental thinking" in the context that is related to one's environment and emphasizes the importance of environmental thinking to live in a sustainable way in our planet. Capra also identifies systems thinking as contextual because it enables to think in the context of a whole when compared to analytical thinking where parts are identified to understand the whole.

Environmental thinking or ecological point of view stems from the field of ecology. Ecology, which means "household" in Greek, studies the Earth as house of all organisms. The German biologist Ernst Haeckel (1886) defines ecology as "the science of relations between the organism and the surrounding outer world" (Capra, 1997; p.33). In 1920s, ecologists started to study food webs that enable them to consider patterns of life. The shift

in their perspective also promoted systems thinkers to extend their network models in ecological subjects. Also, ecologists preferred to benefit from aspects of systems thinking in their studies. For instance, Eugene Odum (1953) explains ecosystems with simple stock and flow diagrams in his essential book "Fundamentals of Ecology". Moreover, Ludwig von Bertalnffy (1968) includes flows and structure of systems to explain interdependence and interconnections between living organisms and environment:

"The organism is not a static system closed to the outside and always containing the identical components; it is an open system in a (quasi-) steady state… in which material continually enters from and leaves into, the outside environment." (cited in Capra, 1997; p. 48)

Capra (2005) identifies three basic principles on "systemic understanding of life" (p.xiv):

1. "Life's basic pattern of organization is the network.
2. Matter cycles continually through the web of life.
3. All ecological cycles are sustained by the continual flow of energy from the Sun."

Capra (1997) also mentions about "non-linearity" and "feedback loops" in networks. He attributes relationships in an ecosystem to non-linearity and he explains that these relationships are non-linear because they are going in all directions. This explanation is a bit insufficient to explain non-linearity from system dynamics point of view. He (2005) also makes reference to biodiversity; several species within a habitat that have overlapping functions and which are interconnected. The strength of biodiversity comes from complex structures of patterns among species. In addition to non-linearity, there are feedback loops that result from cyclical path of relationships in an ecosystem. Likewise, control mechanisms within an ecosystem are based on negative feedback loops.

The "cyclical nature of ecological processes" results in sustainability within nature. Capra (1997) gives a striking explanation about contradictory features of ecology and economics:

"Nature is cyclical whereas our industrial systems are linear. Our businesses take resources, transform them into products plus waste, and then sell the products to consumers, who discard even more waste when they have consumed the products. Sustainable patterns of production and consumption need to be cyclical, imitating the cyclical processes of nature." (p.299)

This quotation from Capra includes some problems related to the systemic terms he used. From system dynamics point of view, cycles represent behaviors, while feedback loops represent structures of dynamic situations. Hence, comparing cycles and linearity is irrelevant as stated in the quotation above. Although, Capra uses systemic terms differently, his contribution to integration of systems thinking and ecology and his support to environmental sustainability education are undeniable.

To combat with current global problems and to minimize the gap between artificial and natural design, people have to be aware of the principles of ecology that are based on cyclical processes and sustainability within nature. Capra (1998) introduces the term ecoliteracy; that is "understanding basic principles of organization of ecosystems and using those principles for creating sustainable human communities" (p.3) and mentions about education for sustainable living to promote ecoliteracy. Moreover, he is the founder of Center for Ecoliteracy where several educational projects are taking place to enhance ecoliteracy among children.

In the following section, systems approach in education is discussed in terms of content and related practices before going into more detail about education for sustainable living and related practices.

2.3. System Based Approach in Education

Jay Forrester (1996) criticizes that "Education has taught static snapshots of the real world. But the world's problems are dynamic." (p. 6). Brown (1992) argues that content of education does not deal with dynamic situations and does not have the tendency to explain how things change over time. In addition to being discipline oriented rather than being multi-disciplinary, education system transfers heavy load of curriculum to teachers and

then to students with insufficient cooperation among teachers and among students. The consequence is unrelated fragments of knowledge that lasts for a short period of time (Brown, 1992).

Besides, several researchers claim that “children are natural systems thinkers” who are able to identify interrelationships and interdependencies before schooling (Brown and Campione, 1994; Senge et al., 2000; cited in Sweeney and Sterman, 2007). And, there are arguments criticizing school as a malfunctioning system that transforms natural systems thinkers into poor systems thinkers by overemphasizing isolated loads of knowledge (Sweeney and Sterman, 2007).

An alternative to traditional fragmented learning, where students are passive participants of instruction by memorizing fragments of knowledge, is system-based learning. Lyneis and Fox-Melanson (2001) claim that system-based learning promotes student-centered learning and supports development of critical thinking and problem solving skills. Moreover, Stunzt, Lyneis and Richardson (2002) mention about students’ progress in conceptualizing interdependencies, short and long-term decisions, and results of their own actions within a given system (cited in Hopper and Stave, 2008).

2.3.1. Studies Related to System Based Approach in Education

A Turkish study also supports inclusion of system thinking related practices in the current curriculum. Nuhoğlu and Nuhoğlu (2007) made a research on seventh grade students and modified “spring mass systems” topic with aspects of system dynamics for the experimental group. They conclude that students develop some systems thinking skills like identifying cause-effect relationships, drawing graphics and arguing about structure of the system. Moreover, students in the experimental group constructed their own dynamic model on STELLA (Systems Thinking for Education and Research Software) that represents relatively higher level of systems thinking skill.

Sweeney and Sterman (2007) made a research on existing mental models on dynamic structures, obstacles that inhibit understanding of dynamic systems, and effective ways of teaching dynamic problems. They observed that subjects do not have the tendency to

- think in a circular fashion,
- close the loops,
- enlarge boundary of their mental models and include real factors that influence behavior,
- focus on both flows (rather than focus on only inflows).

Based on the systems thinking skills that are proposed by Sweeney and Sterman (2000) and explained in Section 2.1.3 on dynamic complexity, they also suggest some basic (pre-requisite) skills to be taught in schools to improve systems thinking skills of younger generations:

- "interpreting graphs, creating graphs from data;
- telling a story from a graph, creating a graph of behavior over time from a story;
- identifying units of measure;
- basic understanding of probability, logic and algebra" (p.250).

2.3.2. Teaching Environmental Content with Systems Approach

When the issue is teaching environmental content with system based approach, there are a number of researches done in the field of science education. Assaraf & Orion (2005) argue that main goal of science education should be development of skills to conceive environmental problems. As discussed, most environmental problems are dynamic in nature, hence understanding these problems require some systems thinking skills. However, still a few studies are conducted to study systems thinking skills in science education, especially in environmental education.

An important deficiency in teaching environmental sustainability was exposed by Assaraf and Orion (2005). They showed that junior high school students have difficulty to see the interrelationships within water cycle. Moreover, they were unable to identify all elements of water cycle that are relatively non-obvious like groundwater and processes like transpiration and capillarity in plants. Teaching water cycle with system based approach resulted in significant differences in

- identifying components, processes, and interrelated relationships within a system,
- understanding cyclic nature of a system,
- organizing components of a system within a network,
- thinking in a time dimension.

It should be noted that these objectives are parallel with systems thinking skills proposed by Sweeney and Sterman (2000).

Riess and Mischo (2010) applied an alternative design to study the effect of different teaching methods to promote systems thinking in the field of education for sustainable development. They designed a special lesson for promotion of systems thinking, a computer-simulation related to forest ecosystem, and a combination of these two teaching methods. The study was applied to 424 sixth graders in Germany. At the end of their study, they concluded that a combination of two methods resulted in significantly higher achievement scores (related to conceptual understanding) and computer simulation and combined method resulted in increase in justification scores (measured as ability to give right answers simultaneously).

Another relevant study on teaching ecological concepts was conducted by Grotzer and Basca (2003). Although most teachers believe that ecological concepts are simple for students, it is proved by several researches that students have misconceptions about these concepts (Adeniyi, 1985; Barmen, Griffiths, and Okebukola, 1995; Gallegos et al., 1994; Leach et. al, 1996; Munson, 1994; cited in Grotzer and Basca, 2003). For instance, when first graders were asked about living organisms in a forest habitat, they only mentioned animals but skipped plants, insects, and decomposers (Strommen, 1995; cited in Grotzer and Basca, 2003). In other words, these first graders were unable to see the whole picture in terms of a forest habitat. It

was also discussed that most students do not understand decomposition and material cycling and they do not have the tendency to relate these processes. Indeed, decomposition process is an outflow which will eventually enter the material cycling process. Leach et al. (1996) worked with high school students and revealed that they did not conceptualize any change in prey population when size of predator population changed. So, these students have difficulty to see the relation between prey and predator populations and are unable to predict future behavior of a dynamic system. Moreover, they also have difficulty in understanding the directional relationship of carnivores, herbivores, and plants from up to down, but understand the opposite direction namely; plants, herbivores, and carnivores in an energy pyramid (cited in Grotzer and Basca, 2003). Grotzer and Basca (2003) criticize that school curriculum is not equipped to support students to overcome these misconceptions. They argue that direct information is not enough to understand these dynamic processes. Hence, they used causally focused activities and classroom discussion as alternative teaching strategies. The intervention is found to be successful in overcoming the misconceptions related to some dynamic problems in ecology.

3. STATEMENT OF THE PROBLEM AND RESEARCH QUESTIONS

3.1. Statement of the Problem

Most environmental problems are dynamic and complex. So, it is hard to understand their effects, structures, and future behaviors that are all related to systems thinking skills. The problem arises when one examines the current education system to look for a systemic perspective and system based education applications.

Brown (1992) criticizes discipline-oriented, fragmented, factually loaded features of the education system. When the issue becomes teaching environmental sustainability, learning facts about environmental issues is insufficient to "develop sound decision making abilities" related to environment (Assaraf and Orion, 2005). Hence, it becomes harder for students to understand dynamics and complexity of the environment. Moreover, teachers perceive that ecological subjects are simple for students, although students have misconceptions about dynamic ecological problems at different grades (Grotzer and Basca, 2003). Besides, there are limited educational materials (Zaraza and Fisher, 1999) and limited system dynamic measures (Plate, 2010) that support development and systems thinking skills of students. Lastly, Riess and Mischo (2010) emphasize the lack of studies about promotion of systems thinking skills especially for fifth to seventh graders.

3.2. Purpose of the Study and Research Questions

There are increasing numbers of environmental problems in Turkey. In addition to number of problems, complexity of these problems arises. In this respect, environmental education plays the crucial role in understanding dynamics of our environment. However, Science and Technology Curriculum has its own deficiencies in terms of environmental content. The curriculum presents the environmental content in a superficial manner. Besides, dynamic environmental problems are explained with limited representations and there is no emphasis on "change over time".

"Human and Environment" is the unique unit that is solely composed of ecological issues in the Science and Technology Curriculum. It includes the concepts of ecology, population, and habitat. It is the first time that the Science and Technology Curriculum mentions "biodiversity" up to seventh grade. The ultimate goal of the unit is to make students to be able to discuss various environmental issues and make inferences about how environmental problems will affect future of their local environment and the world (TTKB, 2013).

This study aims to design an alternative unit plan to teach a seventh-grade science unit; "Human and Environment" and compare effects of the systems based design with the standard unit plan suggested by the Science and Technology Curriculum. Conceptualizations about ecological systems and systems thinking skills were evaluated before and after teaching the chapter with two different approaches, after a six month-period. Besides, a number of instruments were developed to study possible effects of the alternative unit design on students' system thinking skills, competence in dynamic environmental problem solving, and transfer of the systems thinking skills to familiar and unfamiliar contexts. The unit design and the instruments are expected to be exemplary Turkish teaching materials for environmental sustainability education and systems based education.

By taking into account the problems mentioned based on the literature and purposes of the study, the main motivation is to answer the question;

Whether systems approach provides efficient means to teach dynamic environmental issues to seventh grade students?

To answer this question, a number of research questions are formed:

- Do the subjects already have systems thinking pre-requisite skills?
- Which systems thinking skills do the subjects already have?
- Are there any significant differences between systems thinking skills of the subjects in the experimental and the comparison group right after the interventions and after six-month period?
- Are there any significant differences between conceptualizations of dynamic environmental problems of the subjects in the experimental and the comparison group after the interventions and six-month period?7
- Are there any significant differences between science achievement level on environmental questions of the subjects in the experimental and the comparison group right after the intervention and after six-month-period?
- How do the participants in both groups understand and verbalize the structures of some dynamic environmental questions?
- Are the subjects in the experimental group able to transfer knowledge and skills between two environmental tasks with similar dynamic structure after the intervention and after six-month-period?
- Are the subjects in the experimental group able to analyze more general environmental problems that are dynamic and complex after the intervention and after six-month-period?

4. METHODOLOGY

This chapter includes sample of the study, content and reliability analyses of all quantitative and qualitative instrument, design of the study, content of the intervention programs, and finally the procedure of the overall study. Moreover, the pilot study and its findings are presented in this chapter.

4.1. Sample of the Study

The study took place in a public primary school in Istanbul The school was selected for the practical reasons because it is close to the university. In this respect, sampling method of the study is convenient sampling. There were two seventh grade classes in the school, so one class was selected as the experimental group and the other as the comparison group.. Among 52 seventh grade students, 42 student s attended most of the classes and took all the tests. So, the results were reported over 42 subjects. Ages of the subjects ranged from 12 to 14. The distribution of the sample by sex is summarized on Table 4.1. It is seen that the total number of female and male subjects were equal.

Table 4.1 Distribution of sex between groups

		Sex		Total
		Female	Male	
Groups	Experiment	12	10	22
	Comparison	9	11	20
Total		21	21	42

4.2. Instruments

Systems literature (Sweeney and Sterman, 2007; Nuhoğlu, 2008) emphasizes the lack of instruments to evaluate systems thinking skills of children. Besides, Anderson and Burns (1987) also support developing specifically-designed instruments rather than conducting standardized tests to test proposed objectives of educational programs. A number of instruments together with their parallel forms were developed for this study. Each instrument addressing its content and aim together with reliability analyses are presented in the following sections. In this part, the psychometric properties (namely, validity and reliability) of the instruments are examined in depth.

Validity is defined as "the degree to which a test measure what it supposes to measure" (Gay, et. al., 2006; p.134). As one of the types of validity, content validity is related to how well items on a test cover the target content area. To establish content validity, STS and DES tests with systems content were sent to a specialist on systems dynamics and the content of the tests were approved. Besides, some wording modifications were made based on the feedback from one Turkish teacher and a Turkish Literature graduate.

As the second psychometric property to be mentioned in this part, reliability is defined as "the degree to which a test consistently measures" (Gay, et. al., 2006; p.139). Two types of reliability were studied throughout the research; namely, inter-rater reliability and internal consistency.

To omit subjectivity of a single scorer, all test papers of five randomly selected subjects were assessed by three more raters who had systems background. Then, intra-class reliability coefficients (ICC) have been computed to determine proportions of variance of scores assigned by the raters for each item on the tests (McGraw and Wong, 1996). Based on the scores of the raters, some modifications were made. In the final phase of the inter-rater reliability studies, two raters reached nearly a complete agreement (ICC ranging from .91 - 1.00) on the evaluation criteria, content, and weights of each item on the tests. Descriptive statistics about ICCs are summarized on Table 4.2.

Table 4.2. Descriptive statics for ICCs for the individual test items rated by two scorers

Test	Number of items included	Minimum ICC	Maximum ICC	Mean ICC
DES	9	.73	1	.91
STS_pre	3	1	1	1
STS_post	4	.94	1	.98

As the second type of reliability, internal consistency of each test was studied. Internal consistency is related to how much items on a test are consistent between each other and within the test as a whole. Cronbach alpha is the most recommended method to assess reliability of tests including items more than two scoring values. Kuder-Richardson (KR) 20 is the reliability coefficient for tests with dichotomous items such as multiple choice and true-false questions scoring as one or zero (Gay, et. al., 2006). So, Cronbach alpha coefficients were calculated for STS and DES tests and KR-20 coefficient was calculated for SAT test. Besides, there are parallel forms of the STRS and STS tests, so parallel-reliability coefficients that are related to "assessing the consistency of the results of two tests constructed in the same content domain" (Trochim, n.d.) are presented in the following sections.

There are no universal criteria about interpretation of reliability coefficients. But, as a rule of thumb, the criteria below are applied (George and Mallery, 2005):

$\alpha \geq .9$; excellent
$\alpha \geq .8$; good
$\alpha \geq .7$; acceptable,
$\alpha \geq 6$; questionable
$\alpha \geq .5$; poor
$\alpha < .5$; unacceptable

Based on the rule above, the reliability of the test designed are at acceptable and good ranges, except the fact that STRS test is in questionable range.

4.2.1. Demographic Information Sheet

Demographic Information Sheet, Systems Thinking Pre-required Skills Test (STRS) and Systems Thinking Skills Test (STS) were delivered as one test to gain time during implementation of the tests. Totally, seventh grade students were able to complete the overall test within one lesson period (40 minutes). Demographic Information Sheets were used to gather some personal information about the subjects. Sex, parental education background, age, science and mathematics grades were asked on these sheets (see on Appendix H). This information sheet aids to define the sample in detail and to do further analyses by using different characteristics of the sample and data collected via the other instruments.

4.2.2. Systems Thinking Pre-required Skills Test

Systems Thinking Pre-required Skills Tests (STRS) were implemented three times to the sample as indicated on Table 4.3. The test is composed of four questions with two or more sub-questions for each item. The questions on STRS were designed according to the suggestions by Sterman and Sweeney (2000). They proposed that

- Interpreting graphs, creating graphs from data,
- Telling a story from a graph, creating a graph of behavior over time from a story,
- Identifying units of measure,
- Basic understanding of probability, logic, and algebra are the basic skills for applications of further systems thinking skills.

The question types on STRS are ‘fill in the blanks’ and short-answer type of questions (Appendix I). With respect to expected answers, the questions were convergent questions with only one definite answer (Fadem, 2009). On this test, the subjects were not given any partial credits for their answers. So, there is no need to indicate inter-rater reliability for the test. The maximum possible score on this test is 5.5 points. Cronbach alpha for STRS test is .62 and the correlation coefficient between the two parallel STRS test is found to be .64.

4.2.3. Systems Thinking Skills Test

Systems Thinking Skills Tests (STS) were applied three times during the research (Appendix I). The questions were designed in a way that a natural system thinker can answer the questions without having any knowledge about the field specific systems terms like stocks, flows, feedback loops, etc The systems literature includes some specifically designed systems tasks (like Bathtub Task by Sweeney and Sterman (2000), Department Store Task by Sterman (2002)) but, there is deficiency of a test with a number of questions on systems thinking skills. Besides, STS is the first systems test designed in Turkish. STS includes questions modified from famous systems tasks with some original extensions. Each question with references from literature and the addressed systems thinking skill are explained in this section.

The feedback thinking question is modified and translated from the "Systems-Based Inquiry Protocol" designed by Sweeney and Sterman (2007). The question includes two independent situations that could be combined with a reinforcing feedback loop and a balancing feedback loop, individually. The situations were given as in a "fill in blanks" format and the subjects were expected to complete the sentences with the words; "increase" or "decrease". In the second sub-part of the question, it was asked whether these two incidents have something different or in common in terms of "increase" and "decrease" phrases. Subjects were supposed to feel the sense of the "loop" by starting and finishing with the same variable in the incidents.

Delay question was developed by the researchers. A graph of number of participants in a course was given. These participants were expected to graduate in two months after their registration. The subjects were asked to draw the graph of graduates' numbers over time and to compare the two graphs. The expected answer for comparison is to mention about the two-month delay period.

There are two questions addressing to stock-flow thinking skill on STS tests. The first question was asked in a "true/false" format. This question was inspired by the Federal Deficit Task designed by Ossimitz (2002). The theme and content of the question was simplified for the STS test. On this STS question, subjects were expected to calculate the

depth of a boy's pocket money, which is rather a more familiar matter than federal debt for children.

The second stock-flow thinking question was also inspired by the "Arrivals and Departures in the Alpenhotelwork" task designed by Ossimitz (2002). The theme was shifted to number of passengers in a bus, which is a more familiar context for Turkish students.

The maximum possible score on STS test is 11.5 points. For the internal consistency, Cronbach Alpha value is computed as .73 and Spearman-Brown coefficient is computed as .69 for the parallel forms of STS.

4.2.4. Dynamic Environmental Scenarios Test

Dynamic Environmental Scenarios Test (DES) was implemented as immediate and delayed post-tests. This instrument has two purposes in this design. Firstly, it aims to reveal whether the system-based and conventional instructions would result in any differences between the groups. Secondly, this instrument was designed to assess different levels of transfer of learning for application of systems thinking skills on dynamic and complex structured environmental issues. Some tasks included exactly the same content taught in interventions (near transfer), while some tasks involved more complex and general issues than the tasks in the instructions (far transfer) (Perkins and Salomon, 1992).

"Dynamic Environmental Scenarios" (DES) test includes original questions related to local environmental problems. All the questions were developed by the researcher. There are five different environmental scenarios; two of the scenarios are about unfamiliar environmental subjects (ie. subjects not taught during the interventions); construction of the third bridge and collection of wastes, while the other environmental subjects; population dynamics and bioaccumulation were familiar environmental topics for the subjects (see Appendix J). Stock-flow thinking, behavior over time, feedback loops, leverage, delay, identifying variables of a system, and modeling constitute the systems content of this instrument. All the questions on this test are open-ended questions. And, giving partial credits to some responses is a contradictory issue. Hence, inter-rater

reliability for DES is an important point to clarify (Table 4.2).The possible maximum score on DES is 27 points and Cronbach alpha value for internal consistency is .74.

4.2.5. Science Achievement Test

Science Achievement Tests (SAT) were applied twice as immediate and delayed post-tests during the study. The questions were modified from two science test-books (Güvender, 2009 and Oran, 2008) and re-designed for this study. SAT was prepared by taking into account the objectives listed in Science and Technology Curriculum. The test served as a standard test to assess students' achievement after the "Human and the Environment" unit (see Appendix K). SAT contains a variety of questioning styles; multiple-choice, short-answer, matching types of questions. All the questions are convergent questions, so only KR-20 coefficient is computed as .81 for the 20 multiple choice questions on the test.

4.2.6. Student Interviews

To get a deeper understanding of subjects' responses on the various environmental issues and some other dynamic issues addressed in the quantitative instruments, the same questions were asked with some probing in the semi-structured interviews. It should be noted that "probing does not mean prompting" (Bernard and Ryan, 2010, p.31). The interviewer was able to give prompts in a variety of ways such as

- Being silent for a while to let an interviewee think on the question (silent probe)
- Repeating one's last words and encouraging the interviewee to continue (echo probe)
- Asking a question in an explanatory manner rather than asking terse questions (long question probe)

Probes given during interviews should be pre-planned and identical to make comparisons between subjects and groups in experimental studies. In this study, an interview guide including all suggested prompts and explanations were delivered (Appendix L). The interview questions were open-ended type of questions (see on

Appendix M) and all the three types of probes listed above were used during the interviews.

Conducting interviews has some other advantages over written tests besides giving probes. The risk of subjects' diverse writing skills can be manipulated in mixed method researches by providing subjects another expression format. Another risk is related to the limited effort of respondents for completion of written tests (Patton, 1983). The effort of subjects to respond questions is generally higher during interviews in the presence of appropriate probes.

The interviews were semi-structure in the sense that there were a set of questions to be asked and an interview guide with exemplary probing. The interviewer was flexible and was able to make changes in order and details about the questions. But, the questions as well as the probes were similar to make comparisons. According to Bernard and Ryan (2010), semi-structure interviews are appropriate for the respondents who cannot be interviewed in a formal manner. Hence, conducting semi-structured interviews was a good choice for a sample consisting of 12-14 year old teenagers, who find it difficult to express themselves in written order.

Interviews were conducted within two week-period after the instructions were completed. To be fair to both groups and to collect data as much as possible, nearly half of the groups (10 respondents from each group) were selected randomly to participate to the interviews. The range of duration of the interviews was 10-21 minutes. Interview questions were selected among the open-ended questions from STS and DES tests. To evaluate participants' responses, a codebook with designated levels for each question and exemplary participants' responses from the pilot study was designed (Appendix N). In terms of expected answers; the questions were divergent type questions having more than one possible answer. Hence, inter-rater reliability gains importance in the presence of these divergent questions.

The interview questions were taken from STS and DES tests and inter-rater reliability analyses were done for items on these tests. Moreover, to study on inter-rater reliability of the codebook, two researchers assessed an interview with respect to the codes and levels mentioned on the codebook independently. Only one slight disagreement was

detected in the bluefish question and it was ended with an agreement between the researchers.

4.3. Design of the Study

The design of the research is quasi-experimental, in which there are variables that are under control in the field (classroom) and in the absence of random assignment. Gribbons and Herman (1997) supports quasi-experimental research designs in studies where effects of certain educational programs are evaluated and when it is not plausible to deliver random assignments of subjects. There were pre and post-tests (immediate and delayed) throughout the study to measure and evaluate current status and progress of the subjects. By referring to the existence of pre- and post-tests, it can be said that the present study is a "non-equivalent control group, pre- and post-test design" (Gribbons and Herman, 1997).

Conducting pre-tests to both groups enabled the researcher to test equivalence of the groups in terms of the skills to be measured. For practical reasons, the subjects were not placed randomly to the groups, because the study took place in a public school during science lessons. The school contains two seventh grade classes and one class was selected as an experimental group, while the other as a comparison group for the study. The school does not have a policy to classify students according to their achievement levels. In other words, the classes are heterogeneous in terms of student achievement. Hence, it was expected to have somewhat similar variety in terms of subjects' pre-requisite skills and system thinking skills. This expectation was tested via pre-tests that were conducted to both groups.

Immediate and delayed post tests were also implemented right after and six months after the interventions. These post-tests enabled to compare differences between groups after the interventions. Moreover, they were also useful to make within group comparisons from the time the pre-tests implemented to post-test implementation. Delayed post-tests were helpful to study long term, enduring effects of the interventions.

In addition to quantitative tests, interviews were conducted to randomly selected participants from each group. After the instructions and implementation of post-tests, interviews were conducted by asking selected open-ended questions from the post-tests. Interviews were conducted two weeks after the instructions, so these interviews could also be accepted as qualitative post-tests.

Mixed method of data collection including both quantitative and qualitative aspects was adopted in the study. This research method enabled to enrich data collected, to compare data collected in different techniques and to present results in various different formats. The methodology can be categorized as "Explanatory Mixed Method Design". This design implies an emphasis on quantitative data collection. Firstly, quantitative data was collected and analyzed. Then, based on the results of the quantitative part, a qualitative phase was conducted. Qualitative data was collected to support understanding and to explain quantitative data in depth (Gay, Mills, and Airasian, 2007). Sequential procedures rather than concurrent procedures were implemented in the study. These procedures permit to support findings of one measure with the following other measure on the design (Creswell, 2003). Naturally, there could be some parallel or contradictory findings when using each measure and these comparisons would deepen the study itself.

The sequence of the instruments and the design of the overall research is presented on Table 4.3. It is important to note that each group had the same tests at the same period during the study. Upper case letters A and B stand for the parallel or alternate forms of the tests on Table 4.3.

Table 4.3. Research design of the study

Experimental Group	**Comparison Group**
Pre-tests (1 hour) -Demographic Information Sheet -Pre-requisite Skills Test (A) -Systems Thinking Skill Test (A)	Pre-tests (1hour) -Demographic Information Sheet -Pre-requisite Skills Test (A) -Systems Thinking Skill Test (A)
Introduction to system dynamics (3 hours)	Meeting with students (1 hour)
System Based Instruction (15 hours)	Conventional Instruction (16 hours)
Immediate Post-tests (3 hours) --Pre-requisite Skills Test (B) -Systems Thinking Skill Test (B) -Science Achievement Test -Dynamic Environmental Scenarios	Immediate Post-tests (3 hours) --Pre-requisite Skills Test (B) Systems Thinking Skill Test (B) -Science Achievement Test -Dynamic Environmental Scenarios
Interviews (with randomly selected 10 subjects)	Interviews (with randomly selected 10 subjects)
Delayed Post-tests (3 hours) -Pre-requisite Skills Test (B)- -Systems Thinking Skill Test (B) -Science Achievement Test -Dynamic Environmental Scenarios	Delayed Post-tests (3 hours) --Pre-requisite Skills Test (B) -Systems Thinking Skill Test (B) -Science Achievement Test -Dynamic Environmental Scenarios

4.4. Instruction Programs

In this study, two different instructional programs were implemented to two different groups. Both groups were taught by the researcher together to eliminate any possible effects caused by teacher differences. The independent variable for this study was designated as the teaching approaches. The systems based instructional program comprises extensive use of questioning, classroom discussion, and emphasis on teaching various representation techniques for any data as teaching approach. The conventional instruction program incorporates questioning and lecturing.

One of the original components of this research is the designed system-based intervention program for the experimental group. To summarize and to compare contents and activities in the programs, Table 4.4 was formed. It should be noted that the numbers in the parentheses indicate the number of class hours spent for each activity. Another notification about the table is that the number of class hours separated for each group adds up to 14 hours. The following sections include more detailed information about the content of the intervention programs. The original Turkish names of the activities are presented in parentheses in italic fonts in the next sections and the original activity sheets are presented in Appendices. The common instructional activities directly taken from the Science and Technology Lesson Book are placed on Appendix S.

Table 4.4. Comparison of the two unit designs for the experimental and comparison groups

Subject Matter (hours spent with exp. group / hours spent with comp. group)	**Experimental Group**	**Comparison Group**
Introduction to the unit and warm-up (1/1)	"Relating Environmental Concepts Activity" "Habitat Activity" Definition of ecosystems	"Environmental Activity" "Habitat Activity" Definition of ecosystems
Teaching ecology related concepts (1/1)	Ecosystem, species, population, habitat Ecosystem is a system itself. Examples of different ecosystems	Ecosystem, species, population, habitat Examples of different ecosystems
Ecosystems (4/5)	"Population Activity" Trees in a Forest Examining Ecosystems Activity (with CLDs)	Watching documentary Examining Ecosystems Activity
Food chains and food webs (1/1)	"Whom eats Who?" Activity (with negative feeadback loops) "Differences on Two Close Islands" Activity	"Kim Kimi Yer? Etkinliği" "Relationships between Living Organisms"
Biodiversity (2/2)	Presentation about Biodiversity and Life Watching documentary	Presentation about Biodiversity and Life Watching documentary
Environmental Problems (2+3/2+2)	Introduction to environmental problems Bioaccumulation Modeling Activity"	Introduction to environmental problems Bioaccumulation Group presentations related to biodiversity

4.4.1. Systems-Based Instruction Program

The systems-based instruction program aims to enhance systems thinking skills of the subjects to gain a more holistic point of view towards dynamic environmental issues. The focus is on environmental content, because the intervention is designed for the Human and environment unit. However, the instructions started with teaching basic system dynamics concepts in three instruction hours as stated on Table 4.3.

Introductory system dynamics classes include teaching about systems in general, stock-flow diagrams, feedback loops, construction of simple models, and causal loop diagrams. The detailed lesson contents are summarized on Appendix A. "Identification of Stocks and Flows", "Problems with STELLA", and "Feedback Loop Practices" activities are taken from "Road Maps-A Guide to Learning Systems Dynamics" (MIT,1997) and modified with more familiar terms and concepts for Turkish students. The story "Be Nice to Spiders" was modified for teaching feedback loops by Linda Booth Sweeney (Waters Foundation, n.d.) and the story and the lesson design were modified for the introductory lessons. All the introductory activities are placed as Appendix B.

After the introductory lessons, the systems-based "Human and Environment" unit starts a contest "Relating Environmental Concepts Contest" [*Çevresel Kavramları İlişkilendirme Yarışması*] on making interconnections between previously learnt environmental terms. "Environmental Activity" [*Çevre Etkinliği*] was taken from the Science and Technology Lesson Book (Tunç, et. al., 2011) and modified for the systems-based intervention program by making and presenting the interconnections in a loop fashion (Appendix C). The winners of the contest were designated according to the number of the loops and the variables in the loops. Then, the exactly the same activity "Habitat Activity" [*Yaşam Alanları Etkinliği*] (Appendix S) with the comparison group was implemented. The definition of "ecosystems" was presented with a reference to systems that were taught in the introductory lessons. Introduction of the totally the same environmental concepts (ecosystem, population, habitat, ecology, species) was presented with an emphasis on systems and changes over time (i.e. dynamic natural systems).

The Science and Technology Curriculum suggests to examine some ecosystems by taking into account some external and internal environmental factors. However, there are no structured population dynamics activities with any graphs and any other form of data. For systems-based intervention, population dynamics is an important dynamic issue to teach. An activity, which is taken from Road Maps (MIT; 1997) and modified, is called "Population Activity" [*Popülasyon Etkinliği*] (Appendix D). Another population dynamics related activity, which is modified from the book Shape of Change (Quaden , Ticotsky, and Lyneis, 2008), is called "Trees in a Forest" [*Ormandaki Ağaçlar Etkinliği*] (Appendix E). In this activity, subjects were expected to construct a simple stock-flow diagram about a tree population in a forest. After examining the change in tree population, a classroom discussion took place about comparing planting seeds and cutting trees. Seeds belong to another stock; seed stock. And, it was emphasized that planting seeds actually does not replace cutting mature trees.

"Examining Ecosystems Activity" [*Ekosistem İnceleme Etkinliği*] was a common classroom activity that took place in the lesson book (Tunç et. al., 2011). The activity (Appendix S) was structured with presenting two informative chapters about two ecosystems (desert and tundra) from the book "Mountains and Deserts" (Chesire, 2007). The subjects were expected to combine information on one of the ecosystems and present the ecosystem with geographic information, climatic features, and organisms on a poster. The difference of this activity from the one for the conventional instruction is that subjects from the experimental group were expected state interrelations between organisms on causal loop diagrams. This activity formed a basis for food chain, which was the next subject in the program.

Introduction of food chains and food webs were explained based on the information from the lesson book. Then, the common activity "Whom eats whom?" [*Kim Kimi Yer?*] (Appendix S) was done with causal loop diagrams. The activity "Differences on Two Close Islands" [*İki Yakın Adadaki Farklılılar*] was taken from the book Shape of Change (Quaden et. al, 2008) and modified (Appendix F). The cover story was shortened and simplified with the help of a Turkish Literature graduate. This activity includes a larger causal loop diagram containing eight variables. The subjects were supported while drawing the huge causal loop diagram.

Teaching biodiversity part was common for both groups. The same presentation prepared by the researcher was presented. And, Planet Earth Special Episode on Saving Species was watched by the two groups. This documentary is concurrent with the objectives about saving species in the curriculum.

The last part of the unit is related to environmental problems. A presentation on the environmental problems mentioned in the lesson book was presented to both groups Then, an entirely new environmental problem; bioaccumulation was presented to the both groups. The experimental group was expected to construct a dynamic model about bioaccumulation problem in a village near to an old, closed mercury mine. The cover story was a real incident that took place in Clear Lake, California (Giusti, 2009). The modeling activity includes steps like identifying the variables, drawing simple stock-flow diagrams, specifying units, writing equations, and constructing final models on STELLA program. The activity took place in computer laboratory of the school. The subjects had difficulty in specifying units. The researcher and the project assistant supported them to find the correct units. After the support for specifying units, they were able to write equations. The most crucial part of the activity involves classroom discussions about the models. Question 9 about changes in variables and discussing the new possible graphs and Question 10 about additional variables that could be included in the model were very supportive to maintain discussion in the classroom (Appendix G).

4.4.2. Conventional Instruction Program

The conventional instruction program is restricted to the unit plan for “Human and Environment” unit suggested by the Science and Technology Curriculum. The lesson activities and content knowledge were taken from both the lesson book and Student Workbook. The overall content of unit on the lesson book is placed at the Appendix S.

“Human and Environment” unit includes four subparts as introduction of ecological terminology, food chains and webs, biodiversity, and environmental problems as stated on Table 4.3. The activities specified in quotation marks on Table 4.4 were directly taken from both the lesson book and student workbook. One of the supplements to the unit content is to watch Planet Earth Documentaries. During teaching about ecosystems, the

comparison group watched two episodes of the documentary on “Rain Forests” and “Marine Life”. This additional activity was placed because the experimental group was working on two population dynamics related activities. Hence, watching two shortened episodes of Planet Earth documentary enables to equate the number of lesson hours spent on the interventions. The episodes were chosen with the contents that are not related to the instruments. The special episode on “Saving Species” Planet Earth Series was a common activity and is directly related to the objectives of the unit. The final supplement to the unit is related to teaching bioaccumulation as an environmental problem. The curriculum suggests teaching a number of environmental problems in this part of the unit like air pollution, acid rain, and deforestation. Bioaccumulation is a dynamic and complex environmental problem by definition. Teaching bioaccumulation with exactly the same presentation and worksheets do not disrupt the balance of the environmental content of the two interventions. Besides, DES includes a question on bioaccumulation and all the subjects were able to answer this question with the teaching on bioaccumulation.

4.5. Procedure of the Overall Study

This section includes the overall procedures that have been followed during the study. Firstly, all the steps of the study are summarized in the “History of the Study” sub-section and then, the pilot study together with the findings is presented.

4.5.1. History of the Study

The research started in Spring Semester, 2010. The proposal was presented to the jury members in June, 2010. The proposal included a research proposal on designing systems-based trainings for both student and teachers. The jury suggested going on the research with only students for the ease and sustainability of the research. The proposal was revised and sent to the jury members on September, 2010.

The lesson materials and the instruments were prepared for a pilot study. In December 2010, the study was accepted as a Scientific Research Project (BAP) by the university. One project assistant was assigned to accompany lessons during applications and to support data entry. A pilot study was organized to test practicability and validity of the developed lesson materials and the instruments. The pilot study took place in a public primary school in March to April, 2011. During the pilot study, the researcher and the project assistant entered the science classes in the absence of the science teacher. The pilot study took place in two seventh grade classes of the school. There was no comparison group, so all the students were treated with systems-based intervention program. After getting feedback from the pilot study, revisions on materials and instruments were made. It was planned to organize the experimental study at a different school at the end of April, 2011. But on the last week of the study, most of the students quit coming to school due to SBS exam on June 4th, 2011.

The cancelation of the experimental study could be regarded as a waste of one academic year. Because, "Human and the Environment" science unit is taught in the second academic term according to Science and Technology Curriculum. Only small modifications such as exchange of units within one academic term are possible. But, it is nearly impossible to change the placement of one unit across the terms. However, the cancelation provides extra time to study on reliability of the instruments. The instruments were conducted to 130 students from two private primary schools and reliability analyses were done on this new sample. To maintain inter-rater reliability of the instruments with open-ended questions, four raters scored the instruments and reliability analyses among these raters were done. Besides, the system content of the lesson materials and the instruments were controlled by an expert in systems dynamics field and the language used was controlled by one Turkish teacher and one Turkish Linguistic expert to maintain validity of the overall study.

The complete experimental study with all the modifications took place in March, 2012. The researcher and the project assistant had entered the science classes for one month in the absence of the science teacher in two classes assigned as experimental and comparison group. The post-tests were delivered in the following science lessons after the instructions had been completed. After one week, interviews were started. Twenty

interviews were finished within two week-period. The researcher and the project assistant conducted the interviews individually and randomly. After six months of the completion of the study, delayed post-tests (the same alternated forms with the post-tests) were conducted to the same sample in October, 2012.

Five conference proceedings were written from this study. Two of these proceedings were presented at national level, while the others were presented at international level. Two journal articles are being planning to submit; one on quantitative results of the study and one on qualitative findings of the study.

4.5.2. Pilot Study

The pilot study had been applied to 51 seventh grade students in a public primary school in Rumelihisarüstü, Istanbul. The age range of the subjects was 12-14 years old. There were 25 female and 26 male subjects in the sample.

STRS_pre and STS_pre were conducted before the treatment and STS_post, DES, and SAT were conducted as post-tests. The descriptive statistics of the tests are presented on Table 4.5. It should be noted that the number of subjects taking STRS and STS tests were equated to 35 to enable to do further pairwise analyses by eliminating the ones that took either pre or post-test. On the other hand, the number of subjects taking DES and SAT were not manipulated.

Table 4.5. Descriptive statistics for all tests during the pilot study

	N	Mean	Std. Deviation
STRS_pre	35	3.07	1.09
STRS_post	35	3.16	.90
STS_pre	35	1.94	1.35
STS_post	35	2.94	1.62
DES	46	9.60	4.08
SAT	51	58.26	15.24

Table 4.6 includes pair wise t-test results. It could be concluded that there was a statistically significance difference between STS scores of the sample after the treatment (t (35) = -3.667, p = 0.01; d =.67). However, no significant difference was found on STRS tests at .05 significance level (t (32) = -.485, p = .631; d = .09).

Table 4.6. Pair-wise t-test results for STRS and STS tests during the pilot study

	t	df	Sig. (2-tailed)
STRS_pre – STRS_post	-.485	34	.631
STS_pre – STS_post	-3.667	34	.001

Pilot study was helpful to test practicality of the lesson materials and instruments. Assessing open-ended question was also helpful to insert new criteria for evaluation of students' responses. For instance, an unexpected response came from the subjects about suggestions for saving bluefish population: To freeze genetic materials of bluefish and then make them breed in laboratory settings. The subjects were probably inspired with the documentary "Planet Earth" that they watched during the intervention. These kinds of responses were included in the codebook and answer keys of the instruments. After assessment of DES, it was found that

- 21 subjects were able to complete the bridge-traffic loop.
- 11 subjects suggested alternative solutions rather than bridge construction.
- three subjects mentioned about sustainability of bluefish population.

The interviews were conducted three weeks after the pilot study had been completed. Eight students were selected in accordance with their varying performance throughout the intervention. The range for the duration of interviews is 15-27 minutes. Table 4.7 is a frequency table for each level assigned for every question in the interviews. Conducting interviews provide tremendous feedback about possible student responses, amount and type of probing, and evaluation of responses.

Table 4.7. Frequency table for assigned levels of the interview questions during the pilot study

Interview Question	Causal Loop Thinking			Estimating Delay			Stock-Flow Thi.		Bluefish Question			Bluefish Question Sug.			Third Bridge Question			Third Bridge Question Sug.		
Level	0	1	2	0	1	2	0	1	0	1	2	0	1	2	0	1	2	0	1	2
Fre.	1	3	4	0	4	4	0	7	3	3	2	5	3	0	5	2	1	4	4	0

The most important gains from the pilot study could be listed as:

- Cognitive load of the intervention program should be lessened. During the pilot study, it has been concluded that the program includes lots of activities in a limited amount of time. Hence, some activities and part of the program were omitted.
- It was observed that demands of the instruments should be lessened due to time restrictions.
- It was also found that most of the seventh graders are able to write equations, draw and discuss graphs, but they are unable to identify and combine units.

When preliminary results of the pilot study are taken into account, it is concluded that even one month of systems intervention seemed to improve system thinking skills significantly. This was a motivation to improve and continue the experimental study and compare the further results with the control group. However, it should be mentioned that most of the responses in the dynamic environmental scenarios test include static and open-loop thinking. In other words, the subjects were able to connect variables in a cause and effect relation, but they are unable to perceive events as on-going processes in a loop fashion. Hence, it was decided that closed loop thinking with concrete examples related to the content of the unit should have been emphasized more during the systems based instructions.

Another important inference of the pilot study is the importance of qualitative data collection. It was observed that the subjects tend to express themselves more during the interviews; they made longer sentences with more details compared to the shorter phrases in the open-ended questions on the written tests.

5. QUANTITATIVE RESULTS

This section is devoted to detailed quantitative analyses. Firstly, demographic characteristic of the sample are presented. Then, analyses on between groups and repeated measures bases are exhibited. Subparts of the tests are presented and analyzed separately. And, finally quantitative analysis of interview responses are shown.

In this study, both the experimental (N = 22) and the comparison group (N = 20) included less than 30 subjects. For such cases, data must be checked whether it is normally distributed to apply parametric tests. Otherwise, non-parametric tests should be applied (Huck, 2012). Razali and Wah (2011) suggest application of Shapiro-Wilk test for testing normality. If a significant difference is found on Shapiro-Wilk test, Mann-Whitney U test can be applied for independent groups and Wilcoxon Signed Rank test can be applied for paired samples as non-parametric tests.

In addition to statistical differences, observed differences are important to mention in scientific studies. Computation of effect size gives extra information about the degree of the impact or the magnitude of the difference as small, medium, or large. It is assumed that Cohen d values between .2-.49, .5-.79, and above .8 indicate small, medium, and large effect sizes, respectively (Huck, 2012). Throughout the study, computed significance values are supplemented with effect size values, which inform about magnitude of the differences. It should be noted that effect size values are presented together with both statistically significant and non-significant results, because effect size values give extra information rather than statistical differences. Zientek et. al. (2012) strongly support reporting effect size values in quantitative studies: “Effect sizes help researchers determine the importance of observed effects and facilitate meaningful comparisons of findings across studies.” (p.286).

5.1. Demographic Data

The demographic data about the sample include information about distribution of parental education levels and science and mathematics grades. Table 5.1 and Table 5.2 include distributions of mothers' and fathers' education levels, respectively. The parental education of level of the experimental group seemed to be higher at the first glance, but parental education was not designated as a significant covariate with ANCOVA test.

Table 5.1. Distribution of mothers' educational levels between groups

		Mothers' Education Level					
		Pri. Sch. Leaving	Pri Sch.	Middle Sch.	High Sch.	University	Total
Group	Experiment	1	7	6	4	4	22
	Comparison	2	7	5	6	0	20
Total		3	14	11	10	4	42

Table 5.2. Distribution of fathers' educational levels between groups

		Fathers' Education Level					
		Pri. Sch. Leaving	Pri Sch.	Middle Sch.	High Sch.	University	Total
Group	Experiment	0	7	4	9	2	22
	Comparison	5	3	6	6	0	20
Total		5	10	10	15	2	42

Distributions of science and mathematics grades are presented on Table 5.3 and Table 5.4, respectively. There seemed to be some variations in mathematics and science grades across groups. In the following sections, these covariates are taken under control with ANCOVA tests.

Table 5.3. Distribution of science grades between groups

		Science Grades					
		1	2	3	4	5	Total
Group	Experiment	0	0	6	10	6	22
	Comparison	1	5	4	6	4	20
Total		1	5	10	16	10	42

Table 5.4. Distribution of mathematics grades between groups

		Mathematics Grades					
		1	2	3	4	5	Total
Group	Experiment	0	2	9	5	6	22
	Comparison	3	5	4	4	4	20
Total		3	7	13	9	10	42

5.2. Correlational Analysis

To look for any associations between the demographic information, a correlation matrix was constructed. This matrix includes pooled experimental and comparison groups' data (demographic data and pre-test scores) before the interventions took place (Table 5.5). The correlation matrix is important to identify possible covariates that might affect analyses of some results. These covariates are used in ANCOVA tests to control pre-treatment effects in the following sections.

Table 5.5. Correlation matrix for possible covariates

		sci_gra	math_gra	mother_edu	father_edu	STRS_pre	STS_pre
sci_gra	Pearson Cor.	1	.795**	.049	.117	.537**	.417**
	Sig. (2-tailed)		.000	.757	.462	.000	.006
	N			42	42	42	42
math_gra	Pearson Cor.			.119	.164	.544**	.425**
	Sig. (2-tailed)			.452	.300	.000	.005
	N			42	42	42	42
mother_edu	Pearson Cor.				.361*	.026	-.014
	Sig. (2-tailed)				.019	.871	.929
	N				42	42	42
father_edu	Pearson Cor.				1	.007	.091
	Sig. (2-tailed)					.964	.568
	N					42	42
STRS_pre	Pearson Cor.					1	.534**
	Sig. (2-tailed)						.000
	N						42
STS_pre	Pearson Cor.						1
	Sig. (2-tailed)						
	N						

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

It can be seen that science and mathematics grades are highly correlated with STRS_pre and STS_pre tests at .01 significance level. Besides, these variables are also correlated with each other at .01 significance level. Based on this correlation matrix, it is decided that parental education will not be included in further quantitative analyses.

5.3. Between Groups Comparisons

The current study is a quasi-experimental research including two groups to be studied and to be compared. Both the experimental and the comparison groups took the same tests, so there are several possible comparisons between these groups. Table 5.6 includes a summary of the descriptive statistics including mean and standard deviation values and the inferential statistics including statistical differences, effect size values, and their corresponding categories. On Table 5.6, the tests with statistically significant differences are represented in bold characters. This table enables to see the big picture in between groups comparisons in the study. It can be seen that there are statistically significant differences between groups on STS_post, DES and DES_del tests at the first glance. After showing all comparisons between the two groups, each measurement was examined under several subcategories in depth. During the analyses of each test, ANCOVA test was applied to the measures with statistically significant differences among the groups.

ANCOVA test enables a researcher to control *"pretreatment group differences"* by taking into account any covariates (Huck, 2012). Application of ANCOVA test is more appropriate rather than selecting subjects with identical characteristics to control variables. There are the risks of reducing statistical power and generalizability of the results when subjects are selected and matched for the sake of control of covariates. ANCOVA is an appropriate method analysis with the inclusion of all subjects from each group. Partial eta squared (η^2) values are reported together with the ANCOVA results to show "the percentage of the variability in the dependent variable that is explained by the grouping variable" (Huck, 2012; p.223). Partial eta squared values between .01-.05, .06-.13, and above .14 refer to small, medium, and large variability, respectively.

Table 5.6. Descriptive and inferential statistics for all the tests with respect to groups

	Group	N	Mean	Std. Dev.	Sta. Sig..	Effect Size	Effect Size Category
STRS_pre	Exp.	22	2.32	.91	.753	-.01	Very Small
	Com.	20	2.33	1.04			
STRS_post	Exp.	22	3.11	.84	.665	.14	Very Small
	Com.	20	2.98	1.02			
STRS_del	Exp.	22	3.11	.94	.649	.19	Very Small
	Com.	20	2.93	.98			
STS_pre	Exp.	22	4.84	2.37	.276	.34	Small
	Com.	20	4.00	2.56			
STS_post	Exp.	22	6.84	1.96	.009	.84	Large
	Com.	20	4.78	2.88			
STS_del	Exp.	22	5.80	2.14	.587	.17	Very Small
	Com.	20	5.43	2.24			
DES	Exp.	22	12.48	4.93	.004	.94	Large
	Com.	20	8.00	4.59			
DES_del	Exp.	22	10.50	3.28	.042	.64	Medium
	Com.	20	8.08	4.18			
SAT	Exp.	22	64.86	17.32	.757	.19	Very Small
	Com.	20	61.25	20.97			
SAT_del	Exp.	22	57.41	20.53	.486	.22	Small
	Com.	20	53.10	19.06			

5.3.1. Between Group Comparisons on STRS tests

Systems thinking require skills (STRS) tests were applied as pre, post, and delayed tests together with STS tests. The test includes four questions; mostly based on mathematical skills. Results from the Shapiro-Wilk test indicate that the distributions of STRS tests were not normal (Table 5.7). So, Mann-Whitney U-test was applied for testing the difference between the groups. DeBruine (2011) suggests that descriptive statistics including mean rank and sum of ranks should be reported rather than mean and standard deviations in non-parametric tests. Table 5.8 includes descriptive statistics and inferential statistics about STRS tests.

Table 5.7. Normality test results for STRS scores

	Group	Shapiro-Wilk		
		Statistic	df	Sig.
STRS_pre	experiment	.861	22	.005
	comparison	.902	20	.045
STRS _post	experiment	.821	22	.001
	comparison	.910	20	.064
STRS _del	experiment	.840	22	.002
	comparison	.815	20	.001

Table 5.8. Descriptive statistics of STRS tests and Mann-Whitney U test results for STRS tests between groups

	Group	N	Mean Rank	Sum of Ranks	Mann-Whitney U Value	Asymp. Sig (2-tailed)
STRS_pre	experiment	22	22.05	485.00	208.00	.753
	comparison	20	20.90	418.00		
STRS_post	experiment	22	22.25	489.50	203.50	.665
	comparison	20	20.68	413.50		
STRS_del	experiment	22	22.27	490.00	203.00	.649
	comparison	20	20.65	413.00		

It can be seen in Table 5.8 that there are no significant differences between groups at .05 significance level. This is an expected case, since none of the interventions include any content that is totally related to the mathematical skills addressed on STRS tests.

However, this test plays an important role as a covariate for STS and DES tests. This role was examined with applications of ANCOVA tests on the corresponding instruments.

Table 5.9 is constructed to give a general idea about the items on the STRS tests and their frequencies and corresponding significance values after Chi-Square tests between groups. Performances of the subjects are categorized as having no credits, partial credits, and full credits and percentages are reported accordingly. The question about creating behavior over time (BOT) graphs is categorized as no credits vs. full credits. Based on the frequencies of the subjects on Table 5.9, it is seen that the subjects were able to draw and interpret graphs. However, they had difficulties in conversion of units and writing equations and the difficulties had persisted even after the interventions.

Table 5.9. Performances of the subjects from both groups on STRS questions

		STRS_pre			STRS_post			STRS_del		
STRS Questions	Groups	No Credits	Partial Credits	Full Credits	No Credits	Partial Credits	Full Credits	No Credits	Partial Credits	Full Credits
Unit Conversion	Exp	20	2	0	18	4	0	16	6	0
	Comp	16	4	0	14	6	0	13	5	2
	Sig. (2-tailed)	.313			.369			.251		
Writing & Solving Equations	Exp	19	1	2	12	1	9	14	1	7
	Comp	16	1	3	15	0	5	16	0	4
	Sig. (2-tailed)	.834			.541			.279		
Interpreting graphs	Exp	0	12	10	0	4	18	0	1	21
	Comp	0	12	8	0	3	17	0	2	18
		.721			.900			.493		
Creating BOT graphs	Exp	1	X	21	0	X	22	1	X	21
	Comp	3	X	17	1	X	19	2	X	18
		.249			.288			.493		

5.3.2. Between Group Comparisons on STS

The STS scores are presented and compared on Table 5.6. The distributions of STS test scores are tested with Shapiro-Wilk test is applied for all the STS distributions and the distributions were found to be normal ($p > .05$). To determine any differences between

groups on STS tests, independent t-test was applied (Table 5.10). A statistically significant difference was found between the comparison and the experimental group on STS_ post (t (40) = 2.74, p =.009, d = .84). It seems that there is no significant difference between the groups on the STS_pre and STS_del tests at .05 significance level. These comparisons are based on group means. To examine these test performances in depth, the tests are also compared on question or category bases.

Table 5.10. Independent samples t-test results for STS tests

	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
STS_pre	1.104	40	.276	.84	.762	-.699	2.380
STS_post	2.742	40	.009	2.07	.753	.543	3.589
STS_del	.548	40	.587	.37	.676	-.996	1.737

To test whether another variable (covariate) plays a role in the statistically significant difference of STS_post test scores, ANCOVA test was applied. Among the possible covariates (sci_gra, math_gra, strs_pre, and sts_pre), only the covariates math_gra and STRS_pre are significantly related to STS_post scores (F (1, 39) = 25.99, p < .001 and F (1, 39) = 6.26, p =.017, respectively). From Table 5.11, it is understood that STS_post scores are still statistically different between the groups when the significantly related covariates (math_gra and STRS_pre scores) are taken into account.

Table 5.11. ANCOVA test results on STS_post test scores with math_gra and STRS_pre covariates

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared (η^2)
Math_gra	95.153	1	95.153	25.991	.000	.40
Group (math_gra)	15.378	1	15.378	4.201	.047	.10
STRS_pre	32.918	1	32.92	6.26	.017	.14
Group (STRS_pre)	44.987	1	44.99	8.56	.006	.18

STS test includes four questions. These four questions also contain sub-questions (see Appendix I). Among these four questions, two questions are related to stock-flow thinking, one is related to feedback thinking, and the remaining question is related to identifying delays. The categories identified as 0, 1, and 2 correspond to "no points at all", "partial points", and "full points", respectively.

The descriptive statistics for the scores on stock-flow thinking (SF) questions are summarized on Table 5.12. To compare SF scores on pre, post and delayed tests with respect to the groups, independent t-test was applied (since all the distributions are normal). It is concluded that there are no significant differences on SF scores between the groups on the three tests at .05 significance level (Table 5.12).

Table 5.12. Descriptive statistics and independent t-test results for SF scores

	Group	N	Mean	SD	t	Sig.
SF_pre	Exp.	22	2.30	1.00	1.903	.064
	Comp.	20	1.70	1.03		
SF_post	Exp.	22	2.61	1.11	1.744	.089
	Comp.	20	1.98	1.26		
SF_del	Exp.	22	2.43	1.48	1.357	.182
	Comp.	20	1.88	1.13		

For deeper examination of SF scores and to quantify the number of subjects who were able to solve SF questions partially and completely and who could not solve at all, a category based comparison was performed. There are two questions related to the stock-flow thinking skills, so the questions were categorized and analyzed separately. Table 5.13 indicates frequencies of categories for the first and the second stock-flow thinking questions (SF_q1 and SF_q2), respectively. To look for any significant differences between categories of the groups, Chi-Square test is applied and the significance levels are also reported on Table 5.14. It is observed that there are no statistically significant differences between categories of stock-flow questions between the groups at .05 significance level.

Table 5.13. Frequencies for SF_q1 and SF_q2 categories among groups at each STS test and corresponding significance values

		SF_q1 categories				
						Sig. Level
		0	1	2	Total	
SF_q1_ pre	Experiment	17	5	0	22	
	Comparison	11	7	2	20	.196
SF _q1_ post	Experiment	2	16	4	22	
	Comparison	3	12	5	20	.718
SF _q1_ del	Experiment	4	11	7	22	
	Comparison	6	9	5	20	.782
SF _q2_ pre	Experiment	9	6	7	22	
	Comparison	9	8	3	20	.416
SF _q2_ post	Experiment	10	2	10	22	
	Comparison	10	6	4	20	.136
SF _q2_ del	Experiment	11	4	7	22	
	Comparison	12	6	2	20	.224

There is one feedback thinking (FB) question on the STS tests and descriptive statistics about the question are summarized on Table 5.14. As the distributions of three FB scores are normal, independent t-test was applied. It seems that there are no significant differences on FB scores between the groups at .05 significance level (Table 5.14).

Table 5.14.Descriptive statistics and independent t-test results for FB scores

	Group	N	Mean	SD	t	Sig.
FB_pre	Exp.	22	2.00	1.07	1.910	.063
	Comp.	20	1.35	1.14		
FB_post	Exp.	22	2.77	1.23	1.453	.154
	Comp.	20	2.20	1.32		
FB_del	Exp.	22	2.18	.59	.133	.895
	Comp.	20	2.15	.93		

To quantify the number of subjects who were able to solve FB questions partially and completely and who were not able to solve at all, a category based comparison was done. Table 5.15 summarizes frequencies of FB categories among groups on STS tests. Besides, significance level values after applying Chi Square test are reported on the table. No significant difference was detected between categories at .05 significance level.

Table 5.15. Frequencies for FB categories among groups at each STS test and corresponding significance values

		FB categories				
						Sig. Level
		0	1	2	Total	
FB_pre	Experiment	9	8	5	22	
	Comparison	14	2	4	20	.090
FB_post	Experiment	4	7	11	22	
	Comparison	8	4	8	20	.326
FB_del	Experiment	1	16	5	22	
	Comparison	4	12	4	20	.372

The descriptive statistics for the scores on the delay (DEL) question are summarized on Table 5.16. Again, all the distributions about the DEL question scores seem to be normal. Hence, independent t-test was applied to look for any significant differences between the groups. Different from other categories, mean of DEL scores of the comparison group are higher than the experimental group on both pre-test and delayed test. However, it is found that there are no significant differences on delay scores between the groups on the three tests at .05 significance level (Table 5.16).

Table 5.16. Descriptive statistics and independent t-test results for the DEL question scores

	Group	N	Mean	SD	t	Sig.
DEL_ pre	Experiment	22	.50	.91	-.899	.374
	Comparison	20	.80	1.24		
DEL_post	Experiment	22	1.14	1.28	1.127	.266
	Comparison	20	.70	1.22		
DEL _ del	Experiment	22	1.18	.96	-.607	.547
	Comparison	20	1.40	1.35		

The frequencies of delay categories are summarized on Table 5.17. To look for any significant differences among the DEL categories, Chi-Square test was applied. The only significance difference detected was among STS_del DEL categories and expected values for each group were reported on Table 5.17 to understand the source of the significance. The significance stems from relatively high number of subjects in the experimental group who were able to answer the DEL question "partially". It can be said that the difference is in favor of the experimental group ($\chi 2$ (1, *42*) = 6.78, p = .03). This result implies that learning about "identification of delay" is retained when the numbers of subjects, who reached category 1 and 2 and remained at category 0, are compared between the groups.

Table 5.17. Frequencies for DEL categories among groups at each STS test and corresponding significance values

		DEL categories				
		0	1	2	Total	Sig. Level
DEL_ pre	Experiment	16	5	1	22	
	Comparison	13	3	4	20	.365
DEL_ post	Experiment	11	6	5	22	
	Comparison	14	2	4	20	.323
DEL_ del	Experiment	6	14	2	22	
	Expected	7.9	10	4.2	*22*	
	Comparison	9	5	6	20	.032
	Expected	7.1	9	3.8	*20*	

5.3.3. Between Group Comparisons on DES

The DES scores were analyzed and compared on Table 5.6. The DES test includes five questions and each question contains at least two sub-questions (see Appendix J). To test whether any pre-determined covariates affect the statistically significance difference of DES and DES_del tests, ANCOVA test was applied. Among the possible covariates, sci_gra, math_gra, and STRS_pre are significantly related to DES scores ($F(1, 39) = 15.01$, $p = .000$; $F(1, 39) = 6.27$, $p = .017$; $F(1, 39) = 9.98$, $p = .003$, respectively). From Table 5.18, it can be concluded that DES scores are still significantly different between the groups when the significantly related covariates are taken into account.

Table 5.18. ANCOVA test results on DES scores with sci_gra, math_gra, and STRS_pre covariates

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared (η^2)
sci_gra	254.06	1	254.06	15.08	.000	.28
Group (sci_gra)	76.73	1	76.73	4.55	.039	.11
math_gra	126.17	1	126.17	6.27	.017	.14
Group (math_gra)	122.63	1	122.63	6.1	.018	.14
STRS_pre	185.72	1	185.72	9.98	.003	.20
Group (STRS_pre)	211.42	1	211.42	11.36	.002	.23

ANCOVA tests were applied to control pre-treatment effects on DES_del test. Sci_gra and STRS_pre are found to be significantly related to DES_del scores (F (1, 39) = 5.64, p =.023; F (1, 39) = 11.12, p =.002, respectively). It can be seen that DES_del scores are still significantly different between the groups when STRS_pre scores are taken into account (Table 5.19). However, the statistical difference between groups on DES_del diminishes when science grades are controlled with ANCOVA test (Table 5.19).

Table 5.19. ANCOVA test results on DES_del scores with sci_gra and STRS_pre covariates

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared (η^2)
sci_gra	70.52	1	70.52	5.64	.023	.13
Group (sci_gra)	23.22	1	23.22	1.86	.181	.05
STRS_pre	123.82	1	123.82	11.12	.002	.22
Group (STRS_pre)	62.23	1	62.23	5.59	.023	.13

The DES test contains two parts; familiar and unfamiliar parts. The familiar part includes questions on the subjects that were taught during the interventions. The questions on the familiar part (DES_f) are on population dynamics and bioaccumulation that were covered in both classes. The unfamiliar part (DES_unf) questions are about third bridge construction in Istanbul and waste management and these issues were not taught in both classes. The maximum possible scores for DES_f and DES_unf are 18 and 7, respectively.

Shapiro-Wilk test was applied for the four distributions of parts of DES and DES_del tests. It was found that DES_unf and and DES_unf_del scores are not normally distributed ($p < .05$). Hence, independent t-test was applied for DES_f and DES_f_del scores (Table 5.20) and Mann Whitney U test was applied for DES_unf and DES_unf_del scores (Table 5.21). The descriptive statistics about familiar parts of DES and DES_del tests and unfamiliar parts of the two DES tests are also summarized on Table 5.20 and Table 5.21, respectively.

Table 5.20. Descriptive statistics and independent t-test results for DES_f and DES_f_del scores

	Group	N	Mean	Std. Dev.	t	Sig.
DES_f	Exp.	22	8.25	3.280	2.617	.012
	Comp.	20	5.60	3.275	.	
DES_f_del	Exp.	22	6.25	2.434	1.741	.089
	Comp.	20	4.88	2.685		

Table 5.21. Descriptive statistics and Mann-Whitney U test results for DES_unf and DES_unf_del scores

	Group	N	Mean Rank	Sum of Ranks	Mann-Whit. U	Asymp. Sig.
DES_unf	Exp.	22	26.20	576.50	116.50	.008
	Comp.	20	16.33	326.50		
DES_unf_del	Exp.	22	28.70	631.50	61.50	.000
	Comp.	20	13.58	271.50		

The most remarkable differences take place in terms of DES scores. It is found that there are statistically significant differences between groups in terms both their DES scores (t (40) = 3.04, p=.004, d = .94) and their scores on both familiar (t (40) = 2.62, p = .012, d = .81) and unfamiliar questions (U = 116.50, p = .008, d = .85). Moreover, there is also statistically significance difference between groups on DES_del test (t (40) = 2.10, p = .042, d = .64). When DES_del_f scores are tested with respect to groups, the result is significant at .10 significant level (t (40) = 1.74, p = .89, d = .54). The most striking result

is that the experimental group got significantly higher scores on DES_del_unf than the comparison group ($U = 61.5$, $p < .001$, $d = 1.63$). The main motivation of the study was to examine any possible transfer of systems thinking skills to dynamic environmental questions that were both taught and were not taught. These conclusions seem to support this main motivation.

The DES questions were also analyzed separately to look for effects of the single questions and to make comparisons between the groups. Table 5.22 shows the descriptive statistics about the DES and DES_del items. To examine scores of each single question on DES and DES_del tests, Shapiro-Wilk test was applied for checking normality. None of the score distributions of the questions are normal. Table 5.22 also includes Mann Whitney U test values for each question. It is found that the subjects in the experimental group scored significantly higher than the subjects in the comparison group on the third bridge question (q2), bioaccumulation and modeling question (q4), and population dynamics (q5) questions on the DES test at .05 significance level. No significant differences are reported between groups on DES_del test when the questions are examined one by one.

Table 5.22. Descriptive statistics and Mann-Whitney U test results for scores on des and DES_del questions

	Group	N	Mean Rank	Sum of Ranks	Mann-Whit. U	Asymp. Sig.
DES_q1	Exp.	22	23.20	510.50	182.50	.307
	Comp.	20	19.63	392.50		
DES_q2	Exp.	22	25.41	559.00	134.00	.011
	Comp.	20	17.20	344.00		
DES_q3	Exp.	22	20.59	453.00	200.00	.559
	Comp.	20	22.50	450.00		
DES_q4	Exp.	22	24.55	540.00	153.00	.035
	Comp.	20	18.15	363.00		
DES_q5	Exp.	22	24.77	545.00	148.00	.042
	Comp.	20	17.90	358.00		
DES_del_q1	Exp.	22	24.18	532.00	161.00	.112
	Comp.	20	18.55	371.00		
DES_del_q2	Exp.	22	22.50	495.00	198.00	.521
	Comp.	20	20.40	408.00		
DES_del_q3	Exp.	22	23.32	513.00	180.00	.236
	Comp.	20	19.50	390.00		
DES_del_q4	Exp.	22	23.27	512.00	181.00	.105
	Comp.	20	19.55	391.00		
DES_del_q5	Exp.	22	22.95	505.00	188.00	.355
	Comp.	20	19.90	398.00		

To study DES and DES_del questions deeply, categories were assigned for each question. The first DES question is related to bluefish population and the corresponding STS for this question are stock-flow thinking, feedback thinking, and estimating system behavior. Comparison of the experimental and the comparison groups with respect to the assigned categories on DES and DES_del tests and Chi-Square test results are summarized on Table 5.23. No significant difference is detected with Chi-Square test at .05 significance level.

Table 5.23. Frequencies for categories of q1 among groups at DES and DES_del tests and corresponding significance values

		Categories				
		0	1	2	Total	Sig. Level
DES_q1	Experiment	9	3	10	22	
	Comparison	10	5	5	20	.341
DES_del_q1	Experiment	8	4	10	22	
	Comparison	10	7	3	20	.094

The second DES question is related to construction of third bridge in Istanbul and the corresponding STS for this question is feedback thinking. Comparison of the experimental and the comparison groups with respect to the assigned categories on DES and DES_del tests and corresponding significance values are placed on Table 5.24. It is found that there are significantly more people who were able to complete the traffic loop (Category 1) in the experimental group compared to the comparison group on the DES test when expected values are compared.

Table 5.24. Frequencies for categories of q2 among groups at DES and DES_del tests and corresponding significance values

			Categories			
			0	1	Total	Sig. Level
DES_q2	Experiment	Count	9	13	22	
	Expected		13.1	8.9	*22*	
	Comp.	Count	13	4	20	.010
	Expected		11.9	8.9	*20*	
DES_del_q2	Experiment		11	11	22	
	Comparison		12	8	20	.516

The third DES question is about waste management. The systems thinking skill related to this question is identification of delay. Table 5.25 consists of frequencies of categories on DES and DES_del tests across groups and significance values resulted from Chi-Square test. No significant difference is detected with Chi-Square test at .05 significance level.

Table 5.25. Frequencies for categories of q3 among groups at DES and DES_del tests and corresponding significance values

		Categories			
		0	1	Total	Sig. Level
DES_q3	Experiment	13	9	22	
	Comparison	10	10	20	.554
DES_del_q3	Experiment	7	15	22	
	Comparison	10	10	20	.231

The corresponding topic for the fourth DES question is about bioaccumulation and the question is related to modeling. Table 5.26 consists of frequencies of categories for each group on DES and DES_del tests and significance values.

Table 5.26. Frequencies for categories of q4 among groups at DES and DES_del tests and corresponding significance values

			Categories			
			0	1	Total	Sig. Level
DES_q4	Experiment	Count	12	10	22	
		Expected	15.2	6.8	22	
	Comp.	Count	17	3	20	.033
		Expected	13.8	6.2	20	
DES_del_q4	Experiment		17	5	22	
	Comparison		19	1	20	.187

The last question of DES is related to fish population in an aquarium. The corresponding STS for this question is stock-flow thinking. The frequencies of categories for the fifth question for each group on DES and DES_del tests and their significance values are summarized on Table 5.27.

Table 5.27. Frequencies for categories of q5 among groups at DES and DES_del tests and corresponding significance values

		Categories				
		0	1	2	Total	Sig. Level
DES_q5	Experiment	6	2	14	22	
	Comparison	12	1	7	20	.090
DES_del_q5	Experiment	11	4	7	22	
	Comparison	14	0	6	20	.138

Chi Square test was applied for the five dynamic environmental problems stated above. Results indicate that there are statistically significant differences between categories for the experimental and the control groups on the third bridge construction ($\chi^2(1, 42) = 6.64$, $p = .01$) and bioaccumulation ($\chi^2(1, 42) = 4.55$, $p = .03$) questions on DES at .05 significance level. This finding seems to be almost parallel with the significant difference on analyses of scores of single DES items between the groups. No statistically significant differences are detected between categories for the groups on the DES_del test at .05 significance level.

The first three DES questions also include a second part. In the second sub-question of these items, suggestions for the corresponding dynamic environmental problem were asked. Sound suggestions related to the corresponding STS were accepted as category 1. Other irrelevant and ungrounded suggestions were identified as category 0. The frequencies of categories for the suggestions on the three questions on DES and DES_del tests across the groups are shown on Table 5.28. When Chi Square test was applied for the categories of suggestions, no statistically significant difference is found between the experimental and the comparison group in terms of category frequencies of the suggestions they gave.

Table 5.28. Frequencies for categories of suggestions among groups on DES and DES_del tests and corresponding significance values

	Categories			
	0	1	Total	Sig. Level
DES_q1_s				
Experiment	5	17	22	.065
Comparison	10	10	20	
DES_del_q1_s				
Experiment	3	19	22	.152
Comparison	7	13	20	
DES_q2_s				
Experiment	9	13	22	.217
Comparison	12	8	20	
DES_del_q2_s				
Experiment	7	15	22	.129
Comparison	11	9	20	
DES_q3_s				
Experiment	3	19	22	.152
Comparison	7	13	20	
DES_del_q3_s				
Experiment	4	18	22	.216
Comparison	7	13	20	

5.3.4.. SAT Test Scores between Groups

The SAT scores were analyzed and compared on Table 5.6. From this table, it can be concluded that the SAT scores and SAT delay scores were not statistically different for the groups after getting two different instructions. This is an expected situation because the questions on the SAT are factual (knowledge-based) and convergent questions (questions with finite sets of answers) [see on Appendix K]. Indeed, science text books do not include any divergent (open-ended) questions that lead learners to go beyond factual knowledge about the "Human and Environment" unit that was chosen for this study.

5.4. Repeated Measures Statistics

The research design of the current study also enables to evaluate development within groups as the study proceeds. Descriptive statistics including mean and standard deviations are summarized on Table 5.6. To show the big picture about any performance changes of the groups, inferential statistics of the groups including significance levels and effects sizes are presented for the experimental and the comparison group on Table 5.29 and Table 5.30, respectively. To study within group effects, Paired Sample t-test and Wilcoxon Matched Pairs tests are applied. In addition to paired samples t-test and the corresponding non-parametric test, Analysis of variance (ANOVA) tests enable to compare more than two distributions at a time. One-way repeated measure ANOVA test enables to compare distributions of a particular variable belonging to the same participants at different stages. As number of applications of t-test increase, there is a tendency to make Type 1 error; that is, concluding that there is statistical significance although there is not (Huck, 2012).

Table 5.29. Inferential statistics to compare test mean scores of the experimental group

Comparison	Normal Distri.	Statistical Test	Statistical Sig.	Effect Size	Effect Size Cat.
STRS_pre-STRS_post	No	Wilcoxon Signed Rank Test	.001	.91	Large
STRS_pre-STRS_del	No	Wilcoxon Signed Rank Test	.004	.86	Large
STRS_post-STRS_del	No	Wilcoxon Signed Rank Test	.968	.00	Very Small
STS_pre-STS_post	Yes	Paired Sample t-test	.009	.92	Large
STS_pre- STS_del	Yes	Paired Sample t-test	.157	.42	Small
STS_post-STS_del	Yes	Paired Sample t-test	.005	.51	Medium
DES-DES_del	Yes	Paired Sample t-test	.108	.47	Small
SAT-SAT_del	Yes	Paired Sample t-test	.238	.39	Small

Table 5.30. Inferential statistics to compare test mean scores of the comparison group

Comparison	Normal Distri.	Statistical Test	Statistical Sig.	Effect Size	Effect Size Category
STRS_pre-STRS_post	No	Wilcoxon Signed Rank Test	.005	.63	Medium
STRS_pre-STRS_del	No	Wilcoxon Signed Rank Test	.005	.59	Medium
STRS_post-STRS_del	No	Wilcoxon Signed Rank Test	.672	.05	Very Small
STS_pre-STS_post	Yes	Paired Sample t-test	.178	-.29	Small
STS_pre- STS_del	Yes	Paired Sample t-test	.035	.59	Medium
STS_post-STS_del	Yes	Paired Sample t-test	.263	-.25	Small
DES-DES_del	Yes	Paired Sample t-test	.919	.01	Very Small
SAT-SAT_del	Yes	Wilcoxon Signed Rank Test	.305	.41	Small

Violation of sphericity is a thread when conducting Repeated Measures ANOVA test. Sphericity is related to equivalence of "variance of the differences between all combinations" of the groups (Laerd Statistics, 2013). To test sphericity, Mauchly's Test of Sphericity should be used. Violation of sphericity is determined by the significance level of the test. If *p* value is greater than .05, sphericity holds. Otherwise, the final decision of ANOVA are given with Greenhouse-Geisser test values.

5.4.1. Repeated Measures Statistics about the STRS Tests

STRS tests were analyzed in a repeated measure fashion to understand within group dynamics. Figure 5.1 is the plot that summarizes the within groups development of STRS scores both for the experiment and the comparison groups. It can be seen that the trends for both groups seem to be similar. It was reported on Table 5.6 that there is no statistically significant difference between STRS scores of the groups. Hence, the groups seem to develop those prerequisite skills for applying systems thinking skills during the course of the study.

When Repeated Measure ANOVA Test was applied for the experimental group STRS scores, Mauchly's test indicated that the assumption of sphericity holds (χ^2 (2) =.904 , p = .366), therefore sphericity assumed F-ratio is reported. There was a significant effect of systems based intervention on STRS scores for the experimental group; $F(2, 40)$ = 10.319, $p < .001$.

To understand which test pairs result in significant difference on STRS tests, Bonferroni post-hoc test was applied (Table 5.30). It was found that the main difference for the experimental group lays on the difference between STRS_pre and STRS_post and the difference remains significant even on the STRS_del (STRS_pre and STRS_del).

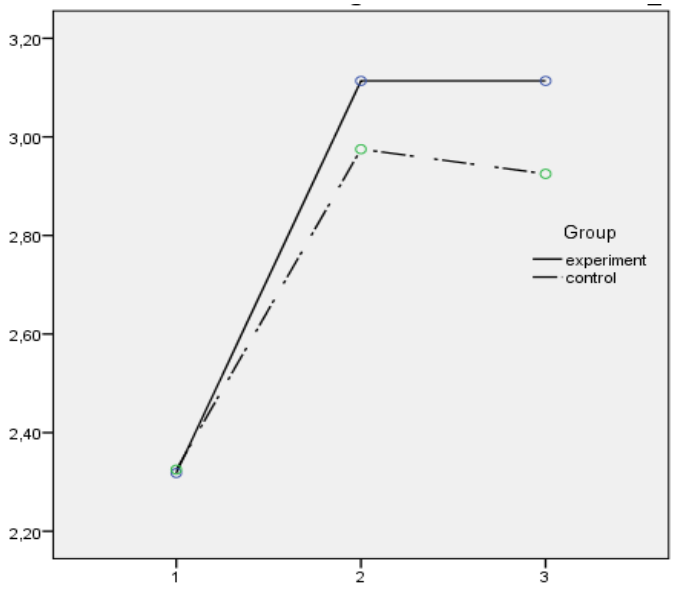

Figure 5.1. Profile plot for STRS scores of the two groups

Table 5.31. Pair-wise comparisons of STRS test scores of the experimental group

(I) strs	(J) strs	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	-.79*	.176	.001	-1.25	-.33
	3	-.79*	.229	.007	-1.39	-.200
2	1	.79*	.176	.001	.34	1.25
	3	.00	.197	1.000	-.51	.51
3	1	.79*	.229	.007	.20	1.39
	2	.00	.197	1.000	-.51	.51

Repeated Measures ANOVA test was also applied to the comparison group. It is found that sphericity holds (χ^2 (2) = .853 , p = .239), therefore sphericity assumed F-ratio is reported. A statistically significant difference was observed on STRS scores of the comparison group after the conventional instruction ($F(2, 40) = 9.845$, $p < .001$). Parallel with the results of the experimental group, the main difference of STRS scores relies on the difference on STRS_pre and STRS_post tests and the difference remains after six months (STRS_del) based on the Bonferroni post-hoc test results (Table 5.32).

Table 5.32. Pair-wise comparisons of STRS test scores of the comparison group

(I) strs	(J) strs	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	-.65*	.185	.007	-1.14	-.16
	3	-.60*	.169	.006	-1.04	-.18
2	1	.65*	.185	.007	.16	1.14
	3	.05	.130	1.000	-.29	.39
3	1	.60*	.169	.006	.16	1.04
	2	-.05	.130	1.000	-.39	.29

5.4.2. Repeated Measures Statistics about the STS Tests

To compare STS scores between and within groups, Figure 5.2 was constructed. This plot is useful to compare and conclude what has been going on about STS scores of the both groups. The highest peak on the plot is observed as the mean STS score of the experimental group right after the treatment. However, there is also a decline of STS scores of the experimental group six months after the treatment. An increase on STS mean scores has also been observed for the comparison group. Overall, the STS scores of the experimental group are always higher than the control group and the delay mean score is still higher, but the difference on the delay tests has not been reported as statistically significant at .05 significance level.

To compare STS scores of each group before, right after the intervention and after six-month period, One-Way Repeated Measure ANOVA Test was applied. When Repeated Measure ANOVA Test was applied for the experimental group STS scores, Mauchly's test indicated that the assumption of sphericity is violated (χ^2 (2) = 11.94 , p = .003), therefore Greenhouse-Geisser F-ratio is reported. There was a significant effect of systems based intervention on sts scores of the experimental group (F(2, 40) = 23.418, p = .014).

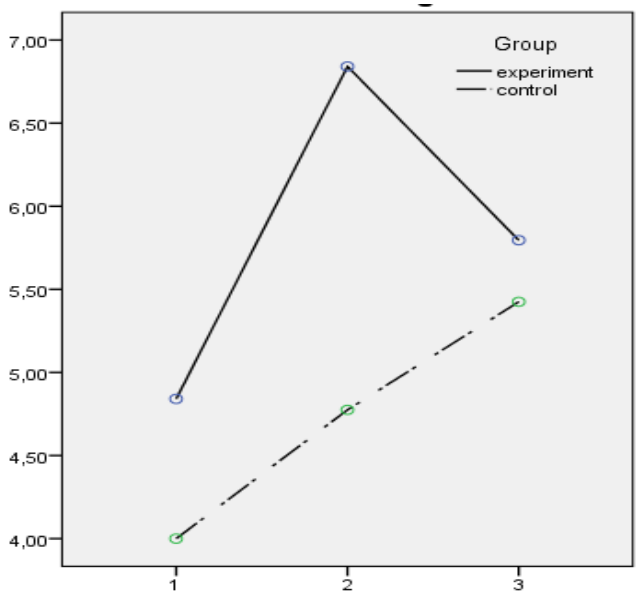

Figure 5.2. Profile plot for STS scores of the two groups

Table 5.33 gives more detailed information about which means differ significantly based on the results of the Bonferroni post-hoc test. It is found that the significant effect of STS mostly stems from the difference between STS_pre and STS_post test scores of the experimental group. However, a statistically significant decrease in mean scores was observed between STS_post and STS_del tests. But, the overall impact of the intervention is still significant in terms of STS scores.

Table 5.33. Paired comparison table of the STS scores of the experimental group

(I) sts	(J) sts	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	-2.00*	.69	.026	-3.8	-.20
	3	-.95	.65	.470	-2.64	.73
2	1	2.00*	.69	.026	.20	3.8
	3	1.04*	.34	.016	.17	1.92
3	1	.95	.65	.470	-.73	2.64
	2	-1.04*	.34	.016	-1.92	-.17

Repeated Measure ANOVA test was applied to comparison group for examining STS scores at each stage of the study. Mauchly's test indicated that the assumption of sphericity holds, χ^2 (2) = .97, p = .79. It is found that the mean scores of the comparison group for the STS tests are not significantly different (F(2, 38) = 3.00, p = .062).

For a more detailed analysis of STS scores, the skills addressed on STS test were examined individually. A summary of Repeated Measure ANOVA test was represented on Table 5.34 indicating each F value and significance value corresponding to the addressed skills for the experimental group. It is observed that there are statistically significant differences on feedback thinking and delay scores of the experimental group at. 05 significance level. To decide on which test(s) determine this difference, Bonferroni post-hoc test results for FB and DEL are presented on Table 5.35 and 5.36, respectively. The pairwise comparisons indicate that the differences lay on the differences between pre and post test scores on the assigned systems thinking skills.

Table 5.34. F and significance values for STS of the experimental group

	F	Sig.
SF	.608	.549
FB	5.767	.006
DEL	4.161	.022

Table 5.35. Paired comparison table of the FB scores of the experimental group.

(I) fb	(J) fb	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	-.773*	.254	.019	-1.434	-.111
	3	-.182	.204	1.000	-.713	.349
2	1	.773*	.254	.019	.111	1.434
	3	.591	.252	.087	-.064	1.246
3	1	.182	.204	1.000	-.349	.713
	2	-.591	.252	.087	-1.246	.064

Table 5.36. Paired comparison table of the DEL scores of the experimental group

(I) delay	(J) delay	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	-.636*	.224	.029	-1.218	-.055
	3	-.682	.274	.064	-1.395	.032
2	1	.636*	.224	.029	.055	1.218
	3	-.045	.290	1.000	-.801	.710
3	1	.682	.274	.064	-.032	1.395
	2	.045	.290	1.000	-.710	.801

To compare performance on each systems thinking skill question between groups and within groups, the Figures 5.3, 5.4, and 5.5 are given. Stock flow thinking scores seem to be parallel with STS scores on Figure 5.2. The changes on SF scores are also parallel between groups, but the experimental tend to get higher SF scores on each STS test.

Feedback thinking scores also show a similar trend with SF scores up to the delayed test. The most apparent difference observed about FB scores is that the difference between the two groups seems to diminish on STS_del test. DEL scores of the comparison group have a distinctive behavior, especially on STS_del test. Although the difference is not reported as statistically significant at .05 between the groups and within the comparison group, the incline of delay scores on STS_del seem to be striking.

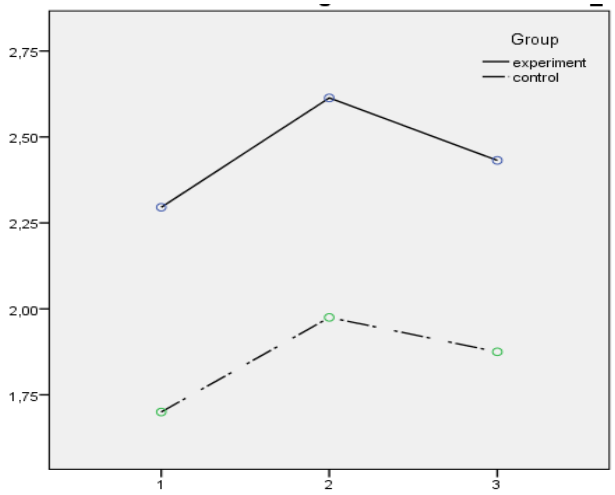

Figure. 5.3. Profile plot for SF scores of the two groups on three STS tests

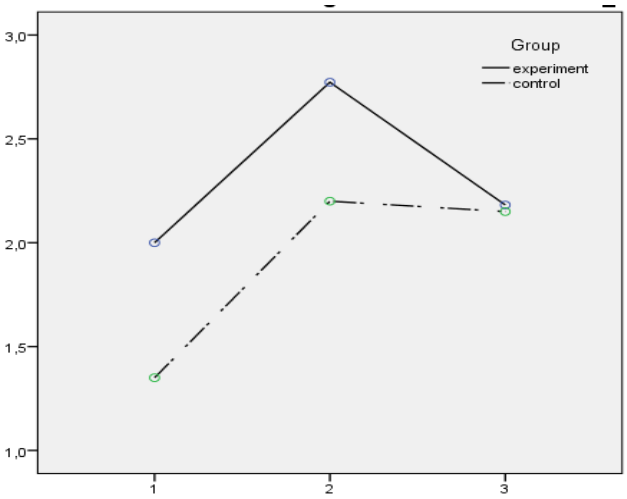

Figure. 5.4. Profile plot for FB scores of the two groups on three STS tests

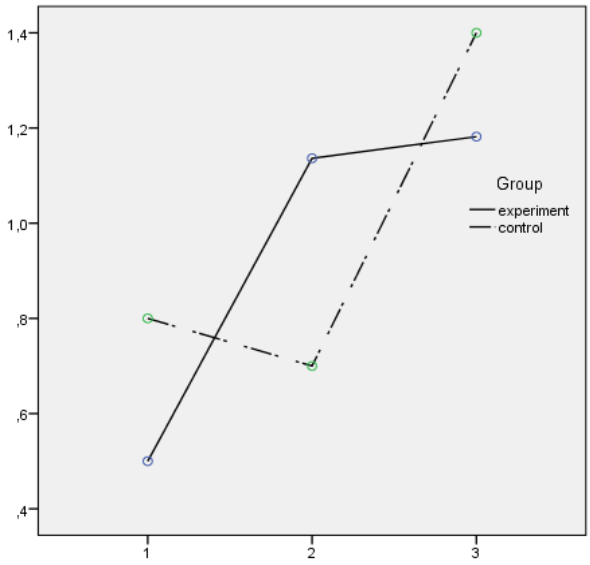

Figure 5.5. Profile plot of DEL scores of the two groups on three STS tests

Repeated Measures ANOVA test was applied to the comparison group to compare scores of each systems thinking skill. ANOVA findings are summarized on Table 5.37. Similar with the experimental group, there are statistically significant differences on FB and DEL scores at .05 significance level. To decide on which test(s) determine this difference, Bonferroni post-hoc test results for FB and DEL are presented on Tables 5.38 and 5.39, respectively. For FB scores, the difference lays between both pre and post tests and pre and delayed tests. An interesting finding about DEL scores is that although the overall comparisons between three STS test show a significance difference, it is found that there are no statistical differences on pairwise comparisons (Table 5.39).

Table 5.37. F and significance values for STS of the comparison group

	F	Sig.
SF	.434	.651
FB	5.151	.010
DEL	3.599	.037

Table 5.38. Paired comparison table of the FB scores of the comparison group

(I) fb	(J) fb	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	-.850*	.284	.022	-1.594	-.106
	3	-.800*	.296	.042	-1.576	-.024
2	1	.850*	.284	.022	.106	1.594
	3	.050	.312	1.000	-.769	.869
3	1	.800*	.296	.042	.024	1.576
	2	-.050	.312	1.000	-.869	.769

Table 5.39. Paired comparison table of the DEL scores of the comparison group

(I) delay	(J) delay	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
1	2	.100	.250	1.000	-.557	.757
	3	-.600	.311	.207	-1.417	.217
2	1	-.100	.250	1.000	-.757	.557
	3	-.700	.282	.068	-1.440	.040
3	1	.600	.311	.207	-.217	1.417
	2	.700	.282	.068	-.040	1.440

5.4.3. Repeated Measure Statistics about the DES Tests

Descriptive information about DES and DES_del was reported on Table 5.6. Figure 5.6 is a profile plot showing the trend of DES and DES_del scores of the two groups. The decrease of the experimental group's scores on the DES_del seems to be striking, but it is not reported as statistically significant at .05 significance level. The DES scores of the control group do not seem to change after six month interval. Table 5.40 includes F and corresponding significance values after application ANOVA test to both groups.

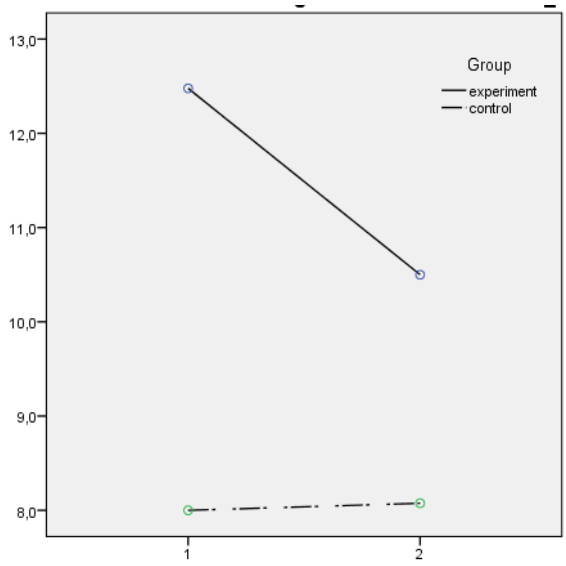

Figure 5.6. Profile plot for DES and DES_del scores of the two groups

Table 5.40. F and significance values for DES and DES_del tests of the two groups

	Group	F	Sig.
DES-DES_del	Exp.	2.81	.108
DES-DES_del	Comp.	.11	.919

To study on familiar and unfamiliar parts of DES and DES_del tests, repeated measure ANOVA test is applied. Figures 5.7 and 5.8 are the plot profiles to summarize the changes of familiar and unfamiliar scores, respectively on DES and DES_del tests for both groups. Table 5.41 and Table 5.42 include F and significance values of DES and DES_del tests for the experimental and the comparison group, respectively. It can be seen that there is a significant decline of scores on familiar items on des tests for the experimental score ($F(1, 21) = 6.71$, $p = .017$).

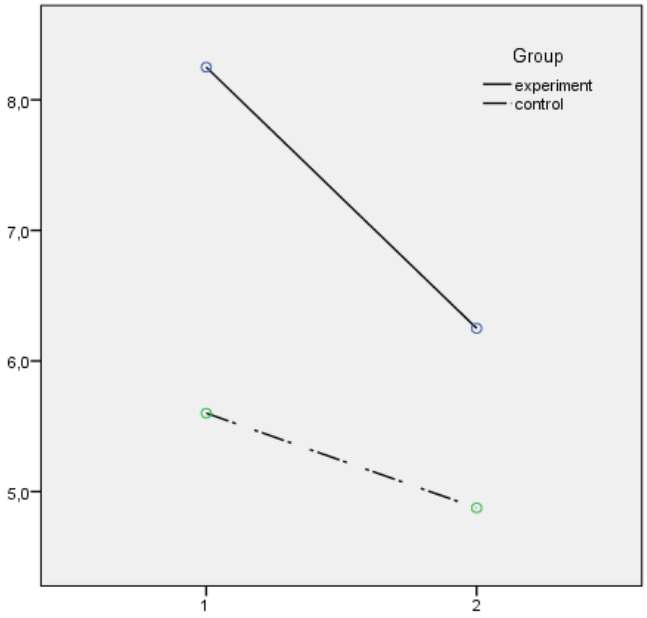

Figure 5.7. Profile plot for DES_f and DES_del_f for the two groups

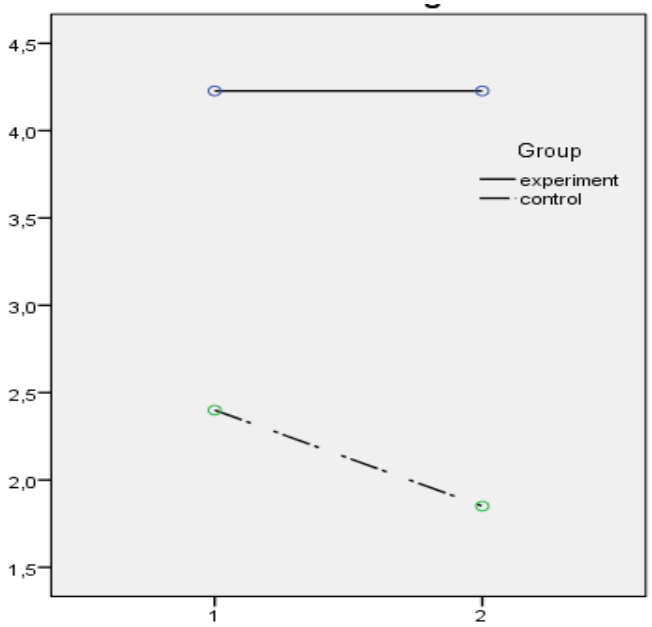

Figure 5.8. Profile plot for DES_unf and DES_del_unf for the two groups

Table 5.41. F and significance values for familiar and unfamiliar parts of DES and DES_del tests of the experimental group

	F	Sig.
DES_f and DES_f_del	6.71	.017
DES_unf-DES_unf_del	.000	1.000

Table 5.42. F and significance values for familiar and unfamiliar parts of DES and DES_del tests of the comparison group

	F	Sig.
DES_f- DES_f_del	1.66	.213
DES_unf-DES_unf_del	1.95	.179

ANOVA analyses result in statistically significant decrease from DES_f to DES_f_del scores for the experimental group. However, this decrease does not affect the overall statistical difference of DES scores of the experimental group when compared to the comparison group (Table 5.6). Another important finding is that the experimental group scores did not change on DES_unf_del .The stability of the experimental group's scores on unfamiliar part can be explained as the main reason of the still significant difference on DES_del in favor of the experimental group.

To examine the DES and DES_del tests deeply, a question-based analysis for each group was carried on. Paired sample t-test is convenient for the comparisons of DES tests at two different stages. Table 5.43 represents paired-sample t-test results for the experimental group. The only statistical significance difference stems for the decline of scores on the fish population in an aquarium question (q5).

Paired sample t- test results for the comparison group are reported on Table 5.44. No statistically significant difference was observed on question-based comparisons of the DES tests for the comparison group. This finding is an expected case when Figure 5.6 is taken into account.

Table 5.43 Effect of time on DES scores of the experimental group

	t	df	Sig.	Mean Difference	Std. Error Mean
DES_q1-DES_del_q1	-.153	21	.880	-.04	.298
DES_q2-DES_del_q2	.624	21	.540	.09	.146
DES_q3-DES_del_q3	-1.667	21	.110	-.27	.164
DES_q4-DES_del_q4	1.555	21	.135	.22	.146
DES_q5-DES_del_q5	2.238	21	.036	.54	.244

Table 5.44. Effect of time on DES scores of the comparison group

	t	df	Sig.	Mean Difference	Std. Error Mean
DES_q1-DES_del_q1	.462	19	.649	.10	.216
DES_q2-DES_del_q2	-1.710	19	.104	-.20	.117
DES_q3-DES_del_q3	.000	19	1.000	.00	.126
DES_q4-DES_del_q4	1.453	19	.163	.10	.069
DES_q5-DES_del_q5	.547	19	.591	.15	.274

5.4.4. Within Subject Statistics about the SAT Tests

To see the big picture about SAT tests, Figure 5.9 is a profile plot showing the trends of SAT and SAT_del scores of the two groups. SAT and SAT_del scores tests were analyzed on Tables 5.29 and 5.30 and no statistically significant difference were found within both groups at .05 significance level. The experimental SAT scores seem to be higher but the differences are not reported as statistically significant at .05 significance level.

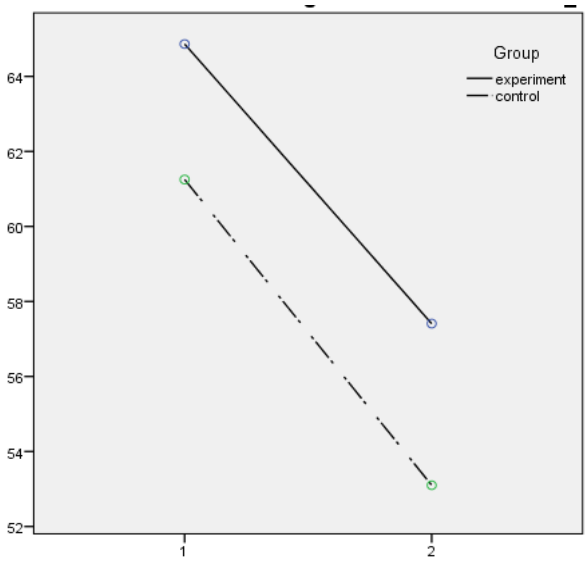

Figure 5.9. Profile plot for SAT and SAT_del scores of the two groups

5.5. Quantitative Analyses of the Interview Responses

The interview responses are analyzed deeply on the "Qualitative Results" Chapter. This section is devoted for quantitative analyses with frequencies of the interview responses. Table 5.45 is a frequency table consisting of the codes specified on the codebook (Appendix N) and frequencies of the levels for each code.

Table 5.45. Categories of the interview questions and frequencies of each category for each group

Groups	Causal Loop Thinking			Estimating Delay			Stock-Flow Thinking		Bluefish Question			Bluefish Question-Suggestion			Third Bridge Question			Third Bridge Question-Suggestion		
	0	1	2	0	1	2	0	1	0	1	2	0	1	2	0	1	2	0	1	2
Exp	0	2	8	2	4	4	2	8	1	1	8	1	5	4	1	3	6	7	2	1
Com	6	0	4	4	1	5	5	5	5	1	4	2	4	4	4	5	1	4	3	3

To compare frequencies among the groups, Chi Square test was applied. To summarize and compare chi-square test results, Pearson chi-square, significance, and Cramer V values are placed on Table 5.46. Cramer V test values are related to strength of association. Significance of Cramer V tests is identical to significance of Pearson chi-square, so significance is reported once on the table. From Table 5.46, it can be seen that the observed values of causal loop thinking and feedback thinking category values differ statistically significant from the expected values (χ^2 (2,20) = 9.33, p = .011 and χ^2 (2,20) = 6.23, p = .036, respectively). It should be noted that the differences are in favor of the experimental group (Table 5.45).

Table 5.46. Summary of Chi Square test for frequencies of interview categories among groups

	Pearson Chi-Square	Asymp. Sig.	Cramer's V
Causal Loop Thinking	9.333	.011*	.683
Delay	2.578	.386	.359
Identification of Stocks & Flows	1.978	.175	.314
Stock-Flow Thinking	4.000	.141	.447
Stock-Flow Thinking_Suggestions	.000	1.000	.000
Feedback Thinking	6.623	.036*	.590
Feedback Thinking_Suggestions	.148	1.000	.008

The statistically significant skills; causal loop thinking and feedback thinking are closely related to each other. Furthermore, these skills were found to be significant on the written STS and DES tests at .05 significance level. These parallel findings support the reliability of the measures and validate the choice of "mixed method" as a data collection method.

6. QUALITATIVE RESULTS

Mixed method of data collection enables to collect both quantitative and qualitative data. As mentioned in the “Methodology” Chapter, this study includes “Explanatory Mixed Method Design”. Based on the characteristics of the “Explanatory Mixed Method Design”, “Systems Thinking Skill Test” (STS) both as pre and post-tests in alternate forms, “Dynamic Environmental Scenarios” (DES), and “Science Achievement Test” (SAT) were applied firstly as quantitative data collection instruments. The first two tests include some open-ended questions where the subjects should reflect their way of thinking about some dynamic environmental phenomena or non-environmental topics that have dynamic characteristics. Although the subjects would like to express their thoughts in the courses during the experimental study, the problem is that their written responses were extremely short; with a few words in most of the cases. Hence, interviews with a limited number of students were helpful to get more insight about their thoughts on dynamic issues and any further effects of the system-based intervention. This chapter includes detailed qualitative analyses of the responses to the interview questions. Before going deep into the qualitative analyses, demographic information about the randomly selected respondents in an anonymous format are presented on Table 6.1.

Table 6.1. Demographic information about the interviewees

Subject Number	Group	Sex	Science Grade	Mathematics Grade
Subject #1	Exp.	Female	3	4
Subject #2	Exp.	Male	4	5
Subject #3	Exp.	Female	5	4
Subject #4	Exp.	Male	5	5
Subject #5	Exp.	Female	4	3
Subject #6	Exp.	Female	5	5
Subject #7	Exp.	Male	4	3
Subject #8	Exp.	Male	4	5
Subject #9	Exp.	Male	3	3
Subject #10	Exp.	Female	3	2
Subject #11	Com.	Male	5	4
Subject #12	Com.	Male	1	1
Subject #13	Com.	Female	5	5
Subject #14	Com.	Male	3	1
Subject #15	Com.	Male	4	5
Subject #16	Com.	Female	2	2
Subject #17	Com.	Female	4	4
Subject #18	Com.	Male	3	3
Subject #19	Com.	Male	2	1
Subject #20	Com.	Female	2	3

6.1. Qualitative Analyses

The qualitative analyses are based on the codebook designed by the researcher (Appendix N). In this section, each interview question (on Appendix M) is explained deeply by referring to all the levels specified on the codebook and giving exemplary subject responses from each group if possible. To be more explicit, there will be

- explanation about frequencies of for each level in each group,
- examples of
    - ✓ typical correct responses
    - ✓ typical false responses
    - ✓ frequently mentioned concepts and phrases,
    - ✓ excellent responses
    - ✓ unexpected and/or irrelevant responses
- comparison between quantitative and qualitative data for the corresponding interview question,
- comparison of mathematics and science grades and levels of respondents for each interview question.

6.1.1. Causal Loop Thinking Question

The first interview question (on Appendix J) was originally taken from the STS test and its code is “causal loop thinking”. Three levels were identified for this question. Level 0 refers to expressing invalid or deficient interrelationships. Six respondents from the comparison group were assigned to Level 0, while none of the respondents from the experimental group were assigned to Level 0. These respondents tend to express invalid interrelationships between the number of chickens and chicks and their responses are examples of typical false responses. For instance, Subject # 13 (female, from the comparison group) explained that:

As number of chickens increases, number of chicks increases. As the number of chicks increases, number of chickens will be constant. Because, I do not think that number of chicks will have any effects on the increase of number of chickens.

Subject #12 (male, from the comparison group) had somewhat similar opinion about the chick and chicken interrelationship:

As number of chickens increases, number of chicks does not increase. As the number of chicks increases, number of chickens does not increase, so it decreases.

Level 1 responses for causal loop thinking include correct interpretation of the interrelationships in an open-loop style. Two respondents from the experimental group expressed correct interrelationships, while none of the respondents from the comparison group were assigned to Level 1 for this particular question.

In addition to correct interpretation of the interrelationships, Level 2 responses consist of comparisons between the two relationships (eating-hunger and chicken-chick) and explanations in a loop fashion. Eight respondents from the experimental group and four respondents from the comparison group reached Level 2 for causal loop thinking code. When Level 2 responses were examined deeply, three types of responses were identified. One type of Level 2 responses included completely correct usage of the system terminology. Two respondents from the experimental group mentioned about loops and called each loop reinforcing and balancing, correctly and their distinctive responses could be classified as excellent responses. For instance, Subject #4 (male, from the experimental group) constructed the interrelationships and named the loops:

"As the number of chickens increases, number of chicks increases. As the number of chicks increases, number of chickens increases. This is a reinforcing loop. The balancing loop is that; as we get hungrier, we eat more. As we eat, our hunger lessens."

Another type of Level 2 responses included explanations based on direct and inverse proportions. These responses based on proportions had not been expected before preparation of the STS questions, but similar responses were got during the pilot study. Two interviewees from the experimental group gave this type of responses. For instance, Subject #2 (male, from the experimental group) compared the relationships as:

"They are not similar because the first one [eating-hunger] is inversely proportional and the second one [chicken-chick] is in direct proportion."

The third type of Level 2 responses included identification of the sentences with the phrases "increasing" and "decreasing" and relevant decisions. As the most frequent response, four respondents from each group gave this type of response. Subject #3 (female, from the experimental group) expressed the interrelationships correctly. She skipped the first part of the two interrelationships and focused on the second part when asked about similarity or difference between the interrelationships:

"The difference is 'as we eat, hunger decreases, as number of chicks increase, number of chickens increases."

Subject #11 (male, from the comparison group) clearly expressed the difference:

"The difference is that there are increases in both [statements of the first interrelationship], but there is only one increase and a decrease here [statements of the second interrelationship]."

Interviewing participants on the causal loop question provides more data than asking them the same question on the written STS test. It was found that three subjects from the experimental group and five subjects from the comparison group did not answer the third part of the causal loop question that demands an explanation about any existing differences or similarities between the two interrelationships on the question. However, all the respondents answered this question during the

interviews. When quantitative and qualitative data on this question were compared, it was found that 50 % of the subjects from the experimental group were able to reach Level 2 on STS_post test, while 80 % of the respondents from the experimental group gave Level 2 responses during the interviews. The percentage (40 %) did not change for the comparison group for both STS_post and the interviews. Besides, statistically significant results were found on Chi-Square test with respondents' frequencies across groups (Table 5.46). No significant results were found on Chi-Square test with frequencies of STS_post test FB question frequencies (Table 5.15).

To examine any relationships between science and mathematics grades and responses to the causal loop question, levels and grades were compared. No regular patterns for both groups were identified between grades and levels of the responses to the causal loop question. For instance, two underachievers in science and mathematics from the comparison group were able to reach Level 2 and one high achiever from the experimental was among the limited Level 1 respondents for the causal loop question.

6.1.2. Estimating Delay Question

The estimating delay question in the interviews was a question from the STS test, (Appendix J). The question includes a graph of number of participants attending a course versus time. The critical information was that participants in the question were graduated two months after their registration. The respondents were asked to comment on the number of graduates versus time graph. Respondents, who misinterpreted the graduate graph, were assigned to Level 0 for the delay code. Two respondents from the experimental group and four respondents from the comparison group were assigned to Level 0. The Level 0 responses included graphs that were totally the same as the registration graph or graphs including accumulation of registered participants irrespective of months.

Partially correct graphs were accepted as Level 1 responses for the delay code. Four respondents from the experimental group and one respondent from the comparison group were assigned to Level 1. As an example of a partially correct graph, Subject #14 (male,

from the comparison group) explained the corresponding values of the graduate graph for the first five months, correctly, but he said that:

"There were no students registered on those months [November and December] so, there would be no graduates."

To be more explicit, this respondent realized the delays in the first months, but could not apply the delay concept for all the values on the graph. The remaining four respondents explained the graph in a correct way but had difficulty to compare and contrast the two graphs. For instance, Subject #6 (female, from the experimental group) drew the graph correctly but could not interpret the graph and gave an unexpected answer with a sense of accumulation:

"As students graduate, the number of registered people decreases."

The respondents, who explained the graduate graph with no deficiencies and clarified the difference between the two graphs, were assigned to Level 2. Four respondents from the experimental group and five respondents from the comparison group were able to reach Level 2. As one of the Level 2 respondents, Subject #2 (male, from the experimental group) explained the difference between the two graphs as follows:

"I can say that the difference is time. As the similarity, the graphs are totally the same. I mean, if 15 people are registered on June, then 15 people have to be graduated on August."

A similar response comes from Subject #17 (female, from the comparison group). The response below can be accepted as one of the excellent answers that touch on both the similarity and difference between the two graphs:

"The shape of the graph is the same but, people change in time. I mean their corresponding time changes. It was 20 people on July, but then, it was 0 [referring the graduate graph]."

In quantitative analyses, it was observed that scores and frequencies of the estimating delay question showed a rather different distribution than the other questions. The comparison group performed better than the experimental group on DEL question on STS_pre and STS_del tests, but the differences were not reported as statistically significant at .05 significance level. On STS_post test, 50 % of the experimental group and 70 % of the comparison group were assigned to Level 0 for the DEL question. Besides, it was reported that five subjects from the comparison group did not answer this question on STS_post test. The situation of Level 0 responses was more pleasant in the interviews: Only 20 % of the respondents in the experimental group and 40 % of the comparison group were assigned to Level 0. During the interviews, it was found that 80 % of the respondents from the experimental group and 60 % of the respondents from the comparison group were able to estimate delay, since they were able to reach Level 1 or Level 2 on delay question.

Comparison of grades and responses yields a different pattern for the delay question. It was found that nine respondents, who were able to reach Level 2 for the delay question, had high science and mathematics grades (grades ≥ 3). This finding is important in the sense that the particular question depends on previous science and mathematics achievement rather than the newly acquired skills during the intervention.

6.1.3. Stock-Flow Thinking Question

The third interview question taken from STS test is the only question that includes two levels in the codebook (Appendix J). The question is about identification of accumulations and calculating stock-flow dynamics. Level 0 implies that the accumulation is ignored, while Level 1 implies the perception of accumulation. Two respondents from the experimental group and five respondents from the control group were assigned to Level 0. Level 0 responses, which were easily recognized, were based on the static calculations rather than dynamic calculations that consider change over time. Evidently in typical false responses, number of passengers in the train were calculated as the difference of people getting on and off at each stop without considering the passengers already inside the train.

Five respondents from the comparison group and eight respondents from the experimental group were classified to Level 1 for stock-flow thinking code. These respondents explicitly included the passengers already inside the train to their calculations on the number of passengers at each stop. For example, Subject #1 (female, from the experimental group) calculated the number of passengers in the train at the second stop in her excellent response:

"15 people got on the train at the first stop… Later, 10 more people got on. But, five people got off. 25 minus five is 20."

This question is rather explicit; the respondents were expected to identify stocks and only two levels were assigned. No unexpected responses were caught during the interviews and the most frequent responses were the Level 1 responses with the sense of accumulation in their explanations about calculations. There were no blank answers for this question on STS_post test for both groups. It was found that 50% of the subjects from the experimental group and 65% of the subjects from the comparison group did not mention accumulation of passengers in the bus across the bus stops on STS_post test. On the other hand, 80% of the respondents from the experimental group and 50% of the comparison group were able to identify stocks during the interviews.

Another important issue about this question is that among the seven Level 0 respondents, five of them (four from the comparison and one from the experimental group) had very low mathematics and science grades. It can be deduced that previous academic achievement plays a crucial role to solve this question.

6.1.4. Bluefish Population Question

The fourth question (on Appendix J) includes environmental content and was taken from Dynamic Environmental Scenarios (DES) test. The skills corresponded to this question are stock-flow thinking, feedback thinking, and estimating behavior of a system. The question is related to the fate of bluefish population if fishermen would continue to fish juvenile bluefish. The question was asked in three steps as fate of juvenile, mature and

overall bluefish population and quotations from the subjects are accompanied with the interviewer questions at each step. Respondents, whose answers relied on a possible increase, steady still or any other irrelevant conditions about the fate of bluefish population, were assigned to Level 0. One respondent from the experimental group and five respondents from the comparison group were allocated to this level for stock-flow thinking code. Among those Level 0 subjects, Subject #19 (male, from the comparison group) explained the fate of the bluefish population as follows:

"Subject #19: Fishermen fish [juvenile bluefish]. Less fish remained. This is why juvenile bluefish [population] decreased.

Interviewer: What will happen to mature bluefish population?

Subject #19: Mature bluefish [population] increases.

Interviewer: How did you come up with this answer?

Subject #19: Fishermen always fish juvenile bluefish; they do not fish mature bluefish. This is why, it [mature bluefish population] increases.

Interviewer: What about the overall bluefish population?

Subject #19: They stand still.

Interviewer: Why do you think so?

Subject # 19: Fishermen always fish juvenile fish and sometimes mature fish."

Subject #19 considered juvenile, mature, and overall bluefish populations as separate stocks and ignored the flows from one stock to another. This was a typical false response; five out of six Level 0 responses did not include any associations between the stocks and flows on the question.

Subject #1 (female, from the experimental group) gave an unexpected response. The response was unexpected because she mentioned about an increase in overall bluefish population in spite of disappearing of juvenile bluefish. Her response was transcribed as follows:

"Subject #1: There will be a decrease in juvenile fish.

Interviewer: What about the mature ones?

Subject #1: There will be no change in mature ones.

Interviewer: Hımm, no change. And, what about the overall population?

Subject#1: As juvenile fish disappear, mature ones reproduce eventually and the overall population will increase."

Respondents, who were able to predict that juvenile, mature, and overall bluefish population would decrease, were allocated to Level 1 for the bluefish question. One subject from each group was allocated to Level 1. These subjects mentioned decrease in all these bluefish populations, but did not explain the reason for declines explicitly. For instance, when the reason for the decline of the populations was asked, Subject #5 (female, from the experimental group) gave an answer which was the information given in the question itself:

"Because fisherman fish juvenile bluefish."

Respondents at Level 2 were expected to go beyond the predictions about the fate of bluefish populations and should have been able to give sound reasons about the changes of the populations. Eight subjects from the experimental group and four subjects from the comparison group were able to reach Level 2. Subject #3 (female, from the experimental group) seemed to construct the connections (via flows) between corresponding populations (stocks) and explained the underlying reasons of decline of the populations in a relatively longer time frame:

"Subject #3: Juvenile bluefish population will decrease. Reproduction will also decline. Why? Because, the juveniles are supposed to grow and give births. That's why it [juvenile bluefish population] will decrease.

Interviewer: You said it would decrease. What about mature bluefish population? How will the mature bluefish population change?

Subject #3: Since, fishermen fish juvenile bluefish, there will be no reproduction in the future and eventually, no mature bluefish and no bluefish at all."

Subject #7 (male, from the experimental group) gave a similar answer:

"There will be a decrease in juvenile bluefish population due to fishing. There will also be a decrease in mature bluefish population, because juvenile bluefish could not grow. Due to decreases in the two populations, the overall bluefish population will decline."

It is seen that excellent responses at Level 2 included the growth (into the mature fish population stock) and reproduction flows (into the juvenile bluefish population). These responses also included rational projections about the overall bluefish population like extinction of the species.

On the DES test, 45% of the subjects in the experimental group and 50% of the subjects in the comparison group were assigned to Level 0 on the bluefish population question. Besides, it was found that two subjects from the experimental group and six subjects from the comparison group did not answer the question at all. The results in the interviews were more striking: Only 10% (one respondent) of the experimental group was assigned to Level 0, while the percentage (50%) was still the same for the comparison group.

Level 2 distributions between groups on DES test and interviews seem to be closer. On DES test, 12 subjects from the experimental group and five subjects from the comparison group reached Level 2. During interviews, eight respondents from the experimental group and four respondents from the comparison groups reached Level 2 for the bluefish population.

To examine the bluefish population question in depth, the last criterion is to compare science and mathematics grades and response level of the participants from each group. Among the five respondents from the comparison groups at Level 0, four of the respondents were very low achievers. No regular pattern was identified for the experimental group because eight out of ten respondents were able to reach Level 2 independent of their academic achievement at mathematics and science courses.

6.1.5. Suggestions about Bluefish Population

The fifth question (on Appendix J) constitutes the second part of the bluefish population question from the DES test. This question is associated to suggestions for the declining bluefish population. It was expected that these suggestions should be somehow related to the fishing size of bluefish; the information given on the fourth question. Giving superficial suggestions with no foundations or with no relations to the real problem are accepted as Level 0 for this question. One respondent from the experimental group and two respondents from the comparison group gave Level 0 suggestions. For instance, two of these respondents mentioned about saving sea. When they were asked about a second suggestion, they both suggested to forbid fishing but did not give any criteria about the fishing restriction. Subject #17 (female, from the comparison group) made some other suggestions with no exact foundations:

"I will make announcements, I will tell the Ministry and the associated people that these fish have not reproduced yet. I will try to warn people about the situation.

[She was asked for another suggestion.]

I will charge some people and educate them about the things that should be done."

Suggestions based on amount of fishing or size of fishing were classified as Level 1 for this question. Five respondents from the experimental group and four respondents from the comparison group were assigned to Level 1 according to the suggestions they made for bluefish population. For example, Subject #9 (male, from the experimental group) had a somewhat creative suggestion for the bluefish population:

"Fishermen should be supposed to measure length of fish with a special mechanism and should throw away the ones shorter than 20 cm."

Subject #20 (female, from the comparison group) also focused on length of fish in her suggestion indirectly:

"Fishermen should prefer to fish middle-aged fish rather than young fish and new-born fish."

Besides the typical true suggestions on amount of fishing and length of fishing, three suggestions about cloning and DNA replication were regarded as Level 1 suggestions. These unexpected suggestions had not been placed on the codebook before the treatments. However, it seemed that after watching the special episode of the Planet Earth Series- Saving Species, the respondents also mentioned about cloning and DNA replication as a treatment method for saving species. For instance, Subject #8 (male, from the experimental group) seemed to assume that bluefish population would be extinct eventually and made the following suggestion for the bluefish population:

"One can get DNA of these fish, use them in the future and make them alive."

This respondent mentioned about protecting DNA molecules in a cold environment and producing new breeds with those genetic materials as shown in the Planet Earth documentary.

The suggestions categorized as Level 2 should be sound and should include explanations of their own rationale in a loop fashion. Four respondents from each group gave appropriate suggestions including the cyclic explanations that were classified as Level 2 suggestion. For example, Subject #4 (male, from the experimental group) explained his excellent suggestion with the relevant information from the question itself:

"I think, the limit of fish size for fishing has to be increased up to 25 cm. If the fishing limit becomes 25 cm, then there will be more reproduction and more fish eventually. And, fishermen will be able to fish, eventually."

Subject #7(male, from the experimental group) explained his excellent suggestion very clearly and the way he answered this question implies an understanding of stocks and flows:

"It would be wiser to fish mature bluefish, because as juvenile bluefish grow up to 25 cm, they are able to lay eggs. After that moment, even one gets hunted; around five off-springs will remain."

Subject #14 (male, from the comparison group) also made an alternative suggestion based on stock-flow thinking:

"There should be a fishing ban during reproduction season of bluefish… This ban will enable bluefish to reproduce."

Quantitative and qualitative measures on suggestions for the bluefish population do not seem to have parallel results. The quantitative measure was more distinctive across groups for this question: 10 subjects from the experimental group and 13 subjects from the comparison group were assigned to Level 0 on the DES test. During the interviews, one respondent from the experimental group and two respondents from the comparison group were assigned to Level 0. On DES test, seven subjects from the experimental group and three subjects from the comparison group were assigned to Level 2 for this question. The situation was rather different on the interviews: Four respondents from each group were able to reach Level 2 by making one sound suggestion and its relevant explanation. It could be deduced that the subjects learnt from the DES test since the interviews took place one or two weeks after the application of the post-tests and probing during the interviews seem to be supportive. The results were different from the other open-ended interview questions, because the cognitive demand for this type of "suggestion" question was relatively low compared to other open-ended questions and the question was not discriminative.

The responses for the bluefish population question and the corresponding suggestion question were not associated. That is to say, the respondents, who were able to conceptualize the stock-flow structure on the bluefish population question, were unable to give sound suggestions and vice versa. As expected, an explanatory trend could not be identified between achievement on science and mathematics and levels the respondents reached on suggestion for the bluefish question.

6.1.6. Third Bridge Question

The sixth question (on Appendix J) is about a current environmental problem in Istanbul. The question was taken from the DES test. Respondents were expected to explain a sentence that includes a feedback loop itself: "Every bridge creates its own traffic." Respondents, who insisted on the opinion that there would be less traffic with a new bridge or explained the issue with irrelevant variables, were assigned to Level 0. One respondent from the experimental group and four respondents from the comparison group were allocated to Level 0. For instance, Subject #5 (female, from the experimental group) mentioned about irrelevant aspects as sea pollution and wastes in her response:

"If that bridge [3rd bridge] is constructed, the houses at that area will be demolished and the [construction] wastes will be thrown away to the sea. Then, sea will be polluted. That's why; the third bridge should not be constructed."

Subject #18 (male, from the comparison group) gave a typical false answer and insisted on construction of a new bridge:

"For example, there is a bridge in Bosphorus. How can I tell you? Suppose you are in Bebek and want to go to Kadıköy. There is traffic on the way. But, if the bridge is contrasted there, there will be less traffic."

Subject #18's response is a typical one, because people tend to think that a new bridge would lead to less traffic. This response implies that the respondent thinks in short terms and is unable to include various variables into one's own conceptual frame about the traffic problem in Istanbul.

The respondents, who were against the third bridge to solve the traffic problem, were assigned to Level 1 for feedback thinking code. Three respondents from the experimental group and five respondents from the comparison group were allocated to Level 1. For instance, Subject #2 (male, from the experimental group) mentioned about the possible traffic jam on the third bridge, but he included additional and irrelevant variables

to his explanation. It seems that he could not close the feedback loop with the inclusion of new variables:

"There are already two bridges that connect Europe and Asia in İstanbul and there is so much traffic on these bridges every day. To open the 3rd Bridge means to consume more cars, to experience more traffic, to feel more bored. People get more bored."

Subject #16 (female, from the comparison group) explained the possible traffic jam on the third bridge in a more explicit manner:

"The 3rd Bridge is planned to be constructed. I think, as the scientists claim every bridge creates its own traffic even though you construct three or ten bridges."

Subject #16 gave a frequent response in the sense that some respondents tend to repeat the information given on the question rather than to explain one's own thoughts about the issue addressed.

Variables mentioned in the interview question should be organized in a feedback loop and the sentence "Every bridge creates its own traffic." should be explained accordingly for Level 2 explanations. Six respondents from the experimental group and one respondent from the comparison group were able to reach Level 2 for feedback thinking code. As an example for Level 2 explanation, Subject #4 (male, from the experimental group) included a number of variables and explained the issue in a loop fashion:

"People may think that after the construction of a new bridge, the bridge will be empty [free of traffic]. And, everybody will prefer this bridge. Because of that, there will be a crowd on the bridge. I mean, as mentioned here, people living around will surge into this bridge. The bridge will also lead to new residential areas for people. These people will also prefer this bridge….By this way; the traffic will increase on the bridge."

Subject #4's response is an example of an excellent answer for this question because it starts and ends with traffic. One could identify increase of the traffic as one follows the variables mentioned (number of people living around, new residential areas, and those people preferring the new bridge) on the feedback loop proposed in the response.

Subject #6 (female, from the experimental group) made a more clear explanation and her organization of the traffic feedback loop can easily be observed:

"I think there will be more residential areas around due to construction of the new bridge. Because of that, there will be more cars and there will be more traffic on the bridges."

On the DES test, 59% of the subjects from the experimental group and 20% of the subjects from the comparison group were able to close the feedback loop on traffic. It should be noted that all the subjects answered the third bridge question on DES test. The situation was quite similar with the interview results: 60% of the respondents from the experimental group and 10% of the respondents from the comparison group were able reach Level 2 on the third bridge question.

To examine any relationships between science and mathematics grades and responses to the third bridge question, levels and grades were compared. For instance, the only respondent at Level 0 from the experimental group was rather a successful student in mathematics and science courses and Level 2 respondents from the experimental group include both high and medium achievers in science and mathematics. Level 0 respondents from the comparison group also include high and medium achievers. Hence, no regular pattern could be identified between grades and levels of the responses to the third bridge problem across the groups.

6.1.7. Suggestions about Traffic Problem in Istanbul

The final interview question (on Appendix #) constitutes the second part of the 3rd Bridge question from the DES test. This question is associated to suggestions for alternative solutions for traffic problem in Istanbul rather than constructing a new bridge. The solutions related to increasing supply like construction of new highways and roads, widening the existing roads were classified as Level 0 suggestions. Seven respondents from the experimental group and four respondents from the control group seemed to make Level 0 suggestions about the traffic problem in Istanbul. The most frequent suggestions for this question were related to constructing new roads, highways or enlarging the existing roads. For instance, Subject #2 (male, from the experimental group) made some interesting but irrelevant and unexpected suggestions when considering the nature of the traffic problem itself:

"Subject #2: I will construct a road like Istanbul Tram Line or I will construct a bridge on the surface of the sea [Marmara Sea].

Interviewer: Why do you think you should construct a bridge right on the surface of the sea?

Subject #2: I think, it is more rational to construct a tunnel under the sea. The wastes are already discharged to sea. By this way, sea sand would be treated and transportation would be easier."

Subject #6 (female, from the experimental group) made another suggestion based on increasing of supply demand:

"There are lots of trees in crossroads. After all, they [trees] are already harmed due to exhaust gases. I will minimize crossroads and enlarge roads."

Level 1 suggestions were classified as suggestions related to decreasing demand of travelling with private cars and transportation on high ways. Two respondents from the experimental group and three respondents from the comparison group made Level 1 suggestions for the traffic problem in Istanbul. For instance, Subject #1 (female, from the experimental group) suggested that:

"People mostly prefer to drive their own cars. If public transportation is preferred, there will be less traffic… If people need to cross to the other side, they can prefer ferry."

Subject #12 (male, from the comparison group) made another suggestion to limit demand of traveling with private cars:

"You know, people give money to cross the bridge. I will raise entry prices to the bridge[s]. …Some people will not prefer to come [cross the bridge by car] after the raise."

For Level 2 suggestions, suggestions about decreasing demand of travelling with private cars and transportation on high ways should be accompanied with a valid explanation of the suggestion. One respondent from the experimental group and three respondents from the comparison group were able to make Level 2 suggestions for the traffic problem in Istanbul. Subject #4 (male, from the experimental group) included the criterion; reducing demand on cars to his suggestion:

"I can increase number of tram services and "metrobus" stops. I will increase quantity of public transportation. If it [quantity of public transportation] will increase, people with no cars will feel comfortable. After this improvement, more people will prefer public transportation. By this way, traffic will be lessened."

Subject #18 (male, from the comparison group) gave two sound suggestions with relevant explanations about the traffic problem and his response implied that he had some background information about traffic related issues:

"Heavy vehicles travel very slowly and occupy more space. These vehicles should travel at night or after 8 pm when people go to home from work… Until 7 am, when people go to work.

[As a second suggestion] People should prefer marine transportation like sea bus. Hence, traffic on land will lessen."

Suggestions for the traffic problem in Istanbul were evaluated differently on the DES test. Two suggestions for the traffic problem were asked on the DES test, but one suggestion with an appropriate explanation was enough to reach Level 2 on the interviews. The quantitative measure was more distinctive across groups for this question: Nine subjects from the experimental group and 12 subjects from the comparison group were assigned to Level 0 on DES test. During the interviews, seven respondents from the experimental group and four respondents from the comparison group were assigned to Level 0. On DES test, 13 respondents from the experimental group and eight respondents from the comparison group were able to give two sound suggestions with relevant explanations and they were assigned to Level 2 for this question. The situation was rather different on the interviews: Four respondents from each group were able to reach Level 2 by making one sound suggestion and its relevant explanation. This assessment of this question was rather surprising, because this is the only interview question that the comparison group performed better than the experimental group. It was observed that some respondents from the experimental group were able to conceptualize the phrase "Every bridge creates its own traffic." on a loop fashion, but they were unable to make relevant suggestions with their previous explanations. On the other hand, there were some respondents from the comparison group, who could not explain the phrase on the previous question but were able to give sound suggestions like frequent use of mass transportation and some traffic bans. Like the suggestion question on bluefish population, no explanatory trend could be identified between science and mathematics grades of the respondents and their assigned levels for this suggestion question.

7. DISCUSSION AND CONCLUSION

This chapter is composed of three parts. The discussion part is an overview of the results together with explanations and discussion of the findings of the current study. The second part is devoted to limitations, justifications of the limitations and suggestions for further studies. Finally, the conclusion part includes some general results of the study, provides answers to research questions, and some final comments on the implications of the study.

7.1. Discussion

The sample of the study was selected conveniently. The study took place in a public school that is close to Boğaziçi University, in Istanbul. The demographic data revealed that parents of the subjects are not highly educated (only 5% of the fathers and 10% of the mothers are university graduates) although the school is situated nearby the university. The school principal informed the researcher that most of the students in the school belong to families with low-economic status.

The school had two seventh grade classes and all the seventh grade students in the school attended this study. For practical and administrative issues, it would be inappropriate to re-organize classes to enable random selection of students to groups. So, all the students stayed in their own classes. In other words, the researcher controlled the treatments and the overall process of the study, but the subjects had already been assigned to groups (classes) (Black, 1999). This is why the study is classified as quasi-experimental study in the absence of random assignment of the subjects.

In this circumstance, the important issue is *"Were the groups equivalent at the beginning?"* It was found that there was no statistically significant difference on STRS_pre and STS_pre test scores between groups at .05 significance level (Table 5.8). To control and to guarantee pre-treatment differences between groups, ANCOVA test was applied in between groups analyses. STRS_pre and STS_pre data together with the demographic data

were presented in a correlation matrix (Table 5.5) and the covariates of the study were determined accordingly. It was found that mother and father education levels were not correlated with the other variables. Another finding is that science and mathematics grades were highly related to STRS_pre and STS_pre scores at .01 significance level. So, these four variables were selected as covariates that are included in further ANCOVA analyses.

Systems thinking required skills (STRS) tests were applied three times to the whole sample. In “Between Groups Statistics” section, it was concluded that there was no statistically significant difference between the experimental and the comparison group on STRS_pre, STRS_post, STRS_del tests at .05 significance level (Table 5.8). The interesting point is that both groups developed these required skills throughout the study as observed statistical differences between STRS_pre and _post and STRS_pre and_del test scores for both of the groups (Table 5.31 and Table 5.32). It can be clearly seen on Figure 5.1 that the STRS mean score trends are very similar for both groups. Since the groups exhibited parallel developments in terms of STRS, it can be deduced that these developments were independent from the different treatments applied throughout this study. One possible explanation is about the same mathematics instruction that the subjects took by the same mathematics teacher. Seventh grade students learn about graphs and data analyses at the second academic term. This mathematics unit is totally related to STRS test and this instruction might have affected their scores on the STRS tests.

STS test scores exhibit an expected trend for most pre-post-delayed post-test designs. The mean scores of the groups on STS_pre are not significantly different at .05 significance level, but the experimental group scored statistically higher than the comparison group on STS_post test. The difference still exists when math_gra and STRS_pre scores are controlled with ANCOVA tests (Table 5.11). It seems that the difference had not lasted for six months and it diminishes on the STS_del test (Table 5.6).

Repeated measures statistics about STS scores also give important results about individual group performances on STS tests. A statistically significant increase on STS_post scores and a statistically significant decrease on STS_del scores are observed in the case of the experimental group (Table 5.31). The increase on STS_post scores is an expected situation after an intensified systems instruction. However, the effect of the

instruction did not last long enough to be observed on the delayed post-test that was applied six months after the post test. One explanation could be the existence of so many factors that cannot be controlled such as contamination of learning (Yıldıran, 2006) on the skills and topics in different contexts, a long summer holiday, and attendance to some preparatory courses for the high stake exam by some of the subjects. The comparison group seems to increase their mean scores on STS tests throughout the study. However, these increases were not reported as statistically significant at .05 significant level ($F(2, 38) = 3.00$, $p > .05$).

STS tests were also examined on category and item-bases. No statistically significance difference was observed on category-based comparisons between the groups on STS tests. The only difference was observed when Chi-Square test was applied for examining category frequencies of the estimating delay (DEL) question across the groups (Table 5.17). There were more subjects who were not able to do the DEL question (Level 0) and who were able to complete the question (Level 2) in the comparison group than the experimental group. But, there were more subjects with partial credits (Level 1) on the DEL question in the experimental group. Utterly, more subjects (at Level 1 and 2) had a sense of "delay" in the experimental group (72 %) than the comparison group (55%).

STS categories and questions were also studied on within group basis. Both groups seem to increase their feedback thinking (FB) and delay (DEL) scores on STS_post test statistically at .05 significance level (Table 5.34 and Table 5.37). Besides, the increase of FB scores of the comparison group seem to last even on the STS_del test. The feedback question on the STS test is related to the topic; population dynamics. The question includes two stocks (chick population and chicken population stocks) that are connected with a flow "growth" from the chick to chicken population stock. Although, the comparison group was treated with a conventional method of instruction free from teaching systems structures, the subject includes dynamic content inherently. This might be an explanation for the increase of the FB scores of the comparison group.

DEL mean scores for the both groups show totally different trends than the other test distributions (Figure 5.5) with the comparison group having a higher mean score on STS_del test than the experimental group. But, it should be reminded that there was no statistical difference between DEL scores of the two groups at .05 significance level. When frequencies for the DEL categories were compared between the two groups with Chi-Square test, there were no statistical differences at .05 significance level on DEL_pre and DEL_post-test frequencies (Table 5.17). The only statistical difference ($\chi 2$ (1, 42) = 6.78, p = .03). lies on DEL_del test frequencies; it seems that more subjects in the experimental group were assigned to Level 1 than the subjects in the comparison group, when expected values were examined on Table 5.17.

Dynamic environmental scenarios (DES) test exhibit more enduring effects for the experimental group. The experimental group scored significantly higher than the comparison group both on DES and DES_del tests at .05 significance level. The statistical significance on DES test still exists when the covariates sci_gra, math_gra, and STRS_pre are controlled with ANCOVA tests (Table 5.18). The statistical difference also exists for DES_del when the covariate STRS_pre was under control with ANCOVA test (Table 5.19). The covariate sci_gra becomes important when comparing DES_del scores between the groups after six month- period. In other words, more successful subjects in science course in the experimental group constitute the reason for the difference on DES_del test between the groups. This finding could be explained with Information Processing Theory. According to this theory, new information is processed in mind and is associated with the existing and relevant knowledge in mind. It is assumed that an organized mind is the mind with more associations between each bit of knowledge (Schunk, 2012). Hence, more successful students are expected to have more organized mind at least about the academic subjects and tend to recall knowledge better in the long run.

Familiar and unfamiliar sections of DES were also examined on between groups and repeated measure bases. It is found that the experimental group scores were significantly better on DES_f part, but the difference does not exist for DES_f_del scores at .05 significance level (Table 5.20). When items on DES tests were examined individually, the experimental group significantly scored higher on three DES questions (two familiar and one unfamiliar part questions). To examine in depth, Chi-Square test was

also applied to compare frequencies of the respondents in both groups. Two intersecting questions (bioaccumulation question as the familiar one and third bridge question as the unfamiliar one) were identified as the questions in which the experimental group performed better. No statistical significant difference was reported for DES_del questions when Independent t-test and Chi-Square tests were applied. One of the possible reasons of decreasing difference on DES_f_del is that bioaccumulation and modeling question, which the experimental group performed significantly better, demands high cognitive load. Modeling is a high level systems thinking skill (Stave & Hopper, 2007) that needs to be practiced for a long period of time for retention. It should be also noted that although there was no statistical difference of performance of groups on individual DES_del questions, it was reported that the experimental group performed significantly better on both DES_unf and DES_unf_del (Table 5.21).

SAT test is the only test that was not designed by the researcher. The test aims to control the environmental content taught throughout the study. The SAT test includes all the content knowledge that the Seventh Grade Science and Technology Curriculum suggests for the "Human and Environment" unit. It was hypothesized that the groups should have performed similarly on the SAT tests. No significant differences were observed on SAT and SAT_del scores between the groups at .05 significance level (Table 5.6).

Besides the quantitative measures, interviews were conducted within two week-period after the completion of the treatments. Hence, it can be assumed that the interviews enabled to measure some delayed effects of the treatments. In most of the cases, the interview responses were parallel with the responses on the STS and DES tests, but examining individual responses result in more insight and enable to make discussions in more depth. Besides, some differences became more apparent after interviewing with the randomly selected respondents from both groups.

The first question of the interviews; Causal Loop (CL) Thinking question constitutes the question called as "feedback question" on STS test in the Quantitative Results section. Although the feedback question scores and frequencies were not reported as statistically significant at .05 significance level, the Chi-Square test results on the

interview responses demonstrate that there was a significant difference on frequencies of the categories of the CL question between the groups. Moreover, the quality of Level 2 responses differs among the groups. Four responses from the comparison group were assigned to Level 2 for this particular question. All these responses concentrate on the "increase/decrease" words when examining the interrelationships between the given cases. On the other hand, among the eight Level 2 responses from the experimental group, two of them explained the interrelations with the completely correct systems terminology and two of them explained the interrelationships the mathematical terms as inverse and direct proportions.

The DEL question on the STS tests and the interviews show rather a different trend than the other questions. The experimental group performed better than the comparison group on solely the STS_post test, but none of the differences were reported as statistically significant at .05 significance level. During the interviews, it was found that the responses of the eight respondents from the experimental group and the six respondents from the comparison group included a sense of delay (Level 1 & Level 2 responses).

The third question on the interviews is related to stock-flow thinking (SF). The experimental group had a higher percentage of correct responses on the interviews than the STS_post test. Probing about the graphs seems to be more helpful for the respondents from the experimental groups. Another important issue about DEL and SF questions is that these questions include mostly mathematical skills (like constructing and interpreting graphs) and the questions are convergent rather than being divergent or open-ended. These features might explain the distinguished patterns between science and mathematics grades and levels of the responses.

Suggestions about the bluefish population and traffic problem in Istanbul were asked both on the DES tests and during the interviews. No significance differences were found in scores between the groups on DES and DES_del tests at .05 significance level. Besides, the frequencies with respect to levels for the interview questions show a similar trend between groups. One of the reasons for the similar results for these suggestions questions is that people are likely to comment on solutions of some environmental problems, even though they could not understand the hidden complex structure of the

problem. For instance, most respondents suggested that mass transportation should have been in progress for the 3rd Bridge question, although they could not explain the phrase; "Every bridge creates its own traffic." In other words, the cognitive demand for this question is rather smaller than the other highly conceptual DES questions, so a significant difference for these suggestion questions could not be found between the groups.

The bluefish population question includes feedback thinking, stock-flow thinking, and predicting behavior of a system skills. The systems-based intervention clearly results in progress of this systems thinking skill, because the statistically significant differences in favor of the experimental group are reported on both the quantitative (DES and DES_del) and qualitative (the interviews) measures. Moreover, as an open-ended and a cognitively demanding question, no regular pattern between science and mathematics grades and assigned levels for the interview question could be identified for the responses from the experimental group. In other words, the significance of pre-defined achievement criteria (course grades) were diminished after participating the systems-based intervention.

The 3rd Bridge question is related to feedback thinking skill. This is one of the questions that results in significant difference in favor of the experimental group on Chi-Square test based on category frequencies of the interview questions (χ^2 (2,20) = 6.23, p = .036) and on the DES test (χ^2 (1,42) = 6.64, p = .010). It should be noted that this question was assessed as a dichotomous question on the DES tests. In other words, the subjects, who were able to complete and close the bridge-traffic loop, were given full credit; while the other subjects were given no points at all. On the other hand, the interview responses were assessed over three levels. So, evaluation of the interview responses includes more details for this question. By taking into account all the measures, it could be concluded that the systems-based intervention results in progress of feedback thinking skill, because the statistically significant differences in favor of the experimental group are reported at .05 significance level on both the quantitative (DES) and qualitative (the interviews) measures.

To sum up, the system based intervention results in statistically significant differences on STS and DES tests and the difference is reported as permanent when DES_del scores are taken into account. But, the difference on STS scores diminished in a six-month period. There are two possible reasons for this situation. One reason is related to

the content of the interventions and the instruments. STS test includes mostly mathematics-related questions, while DES has mathematics-free content. And, both interventions include some activities with mathematical applications, but foci of the both interventions are on environmental issues. Hence, the mathematical applications seemed to be forgotten in the long run, but understanding and conceptualizing dynamic environmental issues had been still enduring for the experimental group even after six months. Another possible reason is that six months is a very long time that might result in uncontrollable contamination with new or intersecting knowledge. During this six-month period, the subjects were exposed to some mathematics instruction at school and some subjects were also exposed to extra mathematics instruction in courses that prepare students for high stake exams. These instructions might explain the increase of STS_del scores of the comparison group.

7.2. Limitations and Future Research

This study was designed as a quasi-experimental research. The school and the subjects were not selected randomly. The school was selected, since it is a public primary school close to the university. All the seventh grade students were included in the study. In addition to non-random selection of the subjects, some subjects could not attend all the classes and all the tests due to extra-curricular activities like drama rehearsals and sport activities. There were 52 seventh grade students in the school, but 45 subjects attended most of the classes due to extra-curricular activities. There were also three more missing subjects who missed to attend one or more than one test. So, all the quantitative analyses were done over 42 subjects. To increase generalizability of the study, the study should be replicated with higher number of students from different schools and from different locations.

The current study includes a pretest-posttest-delayed test design. Including comparison and experimental groups and administering tests at different time periods with respect to the interventions are the strengths of this study. Systems Thinking Required Skills Tests (STRS) and Systems Thinking Skills Tests (STS) were applied three times in the study as pre, post, and delayed tests. The duration between conducting pre and post-tests was around one month. So, to decrease learning effect of the tests themselves,

equivalence forms of STRS and STS were developed. Duration between conducting post and delayed tests was around six months, so the same alternate forms were selected for all post and delayed tests to measure the delayed effect of the interventions. However, SAT and DES tests were administered as post and delayed tests. There were two reasons for this research design. Firstly, the selection of the population as seventh grade students was intentional. "Human and the Environment" unit is the first science unit devoted to solely ecological themes. So, there was an assumption that the subjects attended to this study had limited ecological knowledge. Besides, the difficulty of developing two alternate forms (Gay, Mills, and Airasian, 2006) of multiple choice tests with 20 questions and learning effect (in case of administering two identical tests) are the two risks of application of SAT pre-test. There are some additional reasons for the absence of DES_pre test in terms of the placements of DES tests within the research design. The research design includes a sequence in terms of pre-requisites. There are pre-requisite skills (STRS) for improvement of systems thinking skills (STS) and STS are set as prerequisites for solving dynamic environmental scenarios test. So, DES test was shown up as a post-test at the first time. In addition to sequential order of the research design in terms of the instruments, DES test includes questions to assess near and far transfer (via familiar and unfamiliar questions on DES, respectively) of the STS taught during the systems-based instruction. The existence of familiar and unfamiliar questions on the test enables to differentiate the performances of the subjects in different groups. And, the results have confirmed this assumption: The subjects in the experimental group were able to transfer what they had learnt during the systems-based instruction to even unfamiliar conditions and their far transfer seemed to be enduring as indicated in DES_unf_del scores. On the other hand, the following researches should include SAT_pre and DES_pre in their designs because conducting all the tests as pre, post, and delayed tests would enable to collect tremendous data and to conduct several more analyses in the form of between groups and repeated measures.

The nature of the systems approach is holistic itself. As Brown (1992) argues about discipline-oriented and fragmented structure of the education system, the following researches should be interdisciplinary with a focus on more than one subject matter at a time and should include collaboration among teachers from different disciplines and researchers. An exemplary teacher project was conducted by Zaraza and Fisher (1999). CC-STADUS Project aims to generate a population of teachers who are able to interpret, develop and implement models and curriculum materials based on systems approach. 36

teachers from various disciplines trained every year and the trained teachers became instructors of the upcoming teachers throughout the project. The project can be identified as an authentic project in the sense that it supports forming and sustaining networks and collaboration between teachers by enabling master teachers to train novice teachers. In addition to collaboration, training by experienced teachers increase credibility and effectiveness of the training programs by focusing on real classroom systems approach practices. It was found that after implication of dynamic models in their classes, project teachers realized that using and building models allowed their students to explore problems in more depth, ask and answer better questions, and develop an understanding that their world is full of systems. When dynamic models are introduced in classrooms, students tend to be more involved and active during both individual and group works. Moreover, an independent system modeling course was opened in one of the project schools. High number of students took the course as an elective. The students expressed that they found the content of the course very interesting. Teachers expressed that the course was a particularly attractive for creative students who are not excited by traditional mathematics and science classes.

To spread systems-based education and to study effects of different system-based education designs, the first step should be teacher training. Training teachers is a more sustainable way rather than conducting single studies depended on researchers acting also as practitioners. A possible next study would be about training teachers and examining effects on both teachers and their students for a longer period of time. Training science and biology teachers both about systems approach and its applications on the ecological subjects would be a priority for the researcher. But for a more holistic point of view, teachers with other backgrounds should be included to ensure prevalence within a school environment. It would be interesting to collect data and study in Turkish schools with trained teachers and students who are exposed to systems thinking in their various lessons. Analyzing long term effects of continuous system-based education at different grades would be a distinguishable study for both national and international levels.

7.3. Conclusion

This research is the first Turkish academic study on application of system dynamics in environmental education of K-12 students. The study includes development of an educational program with authentic learning activities and instruments which contribute to both environmental education and systems dynamics literature. This research has some strength like including two groups to enable group comparisons, conducting pre and posttests to enable both group comparisons and repeated measures, and conducting delayed tests to analyze long term effects of the interventions on both groups. Another strength of the research is that the study does not depend only on quantitative analyses. Semi-structured interviews were conducted with equal number of participants from both groups. Collecting and analyzing two different types of data empower the validity of the results and mixed method of data collection enables deeper analysis of the available data. Moreover, the researcher taught to both groups to eliminate any effects related to teacher differences.

A Turkish study also supports inclusion of system thinking related practices in the current curriculum. Nuhoğlu and Nuhoğlu (2007) made a research on seventh grade students and modified "spring mass systems" topic with aspects of system dynamics for the experimental group. They concluded that students developed some systems thinking skills like identifying cause-effect relationships, drawing graphics and arguing about structure of the system. Moreover, students in the experimental group constructed their own dynamic model on STELLA that represents relatively higher level of systems thinking skill.

Nuhoğlu and Nuhoğlu's study (2007) is the first Turkish academic study that incorporates systems dynamics perspective and its tools in science courses. They developed their own instruments, because there have been no Turkish instruments to measure systems thinking skills for children. The study included development of Drawing Graphs, Problem Solving, Cause and Effect, and Conceptual System Dynamic Tests. These tests are useful to further related researches with different age groups. Conceptual System Dynamic Test includes identification of systems terminology. This test is more appropriate with non-experimental designs. On the other hand, STS and DES tests are more appropriate for experimental studies due to systems-free terminology.

The Science and Technology Curriculum allocates 14 hours for the "Human and Environment" unit. This unit includes basic terminology like ecosystem, ecology, population, and habitat; interrelations within ecosystems, food chains and food webs; biodiversity, and environmental problems. "Human and Environment" unit includes mostly factual knowledge. The curriculum proposes 14 objectives for the unit. Among these 14 objectives, nine of them are objectives at cognitive domain (objectives related to mental skills) and the remaining five objectives are affective (related to feelings). According to Bloom's famous taxonomy (1956), only two objectives at cognitive domain in the unit serve for high order thinking. Bloom classifies the higher level objectives at analysis, synthesis, and evaluation levels, while the lower level of objectives at knowledge and comprehension levels. This is a good justification that the curriculum itself presents the only environmental unit in a superficial manner. The system-based intervention focuses more on higher-level of thinking by identifying variables within ecosystems, constructing causal loop diagrams and building stock-flow structures related to population dynamics, and modeling a change in an ecosystem. With a multi-disciplinary and holistic perspective, the intervention includes some classroom activities that are also based on mathematical skills. The final product of the intervention is a dynamic model about bioaccumulation problem in a coastal settlement.

Another justification about superficial presentation of the unit is related to the science process skills addressed in this unit. Developing science process skills is another endeavor of the Science and Technology Curriculum (Ministry of Education, 2005). Science process skills are defined as "the thinking skills that scientists use during formation of knowledge, focusing on problems, formulating the results" (Ministry of Education, 2005; p.66). Different from other science units on the curriculum, only two objectives are devoted to development of science process skills. These objectives are only related to prediction and comparison as science process skills. On the other hand, the system-based intervention includes the skills; comparison, prediction, identifying variables, hypothesizing, collecting data, discussing results, modeling, and presenting data. These justifications are not limited to Turkish Science Curriculum. Grotzer and Basca (2003) also mentioned about inadequate ecological content in curricula of other countries. Besides, they also mentioned about teachers' tendency for oversimplification of ecological

content within curricula. This tendency might be result of superficial and low-level content of the ecology-related subjects.

The systems based intervention designed for this research includes authentic learning activities and instruments. The literature supports the notion of developing specific instruments for the intended objectives of an educational program rather than using standardized tests (Anderson and Burns, 1987). STRS tests were designed to measure pre-requisite mathematical skills for developing systems thinking skills as addressed in Sweeney and Sterman's study (2000). The STS tests were designed according to the skills presented in the same study. It should be noted that this is the first Turkish STS test designed for children together with its alternate form. The STS questions include application of some mathematical skills and they are free from environmental content. In contrast, DES test is composed of current, dynamic, and local environmental issues. DES does not include any explicit applications of mathematics skills, but the questions could only be answered in a systems-based perspective. SAT is the only instrument that is not developed by the researcher, but the questions of SAT were taken from three text books. The test includes mostly factual questions as the Science and Technology Curriculum demands.

From this point on, the research questions are answered in the light of the findings:

1) Do the subjects already have systems thinking pre-requisite skills?

Most of the subjects (88 %) were able to draw behavior of time graphs in the presence of a story and 48 % of them were able to interpret data from the given graphs.

The problematical pre-requisite skills were related to unit identification. None of the subjects were able to answer four sub-questions related to units. Only 14% of the subjects were able to answer the question with a familiar unit (km/h).

Another challenging pre-requisite skill was related to writing equations. Only 14% of the subjects were able to write the correct equation about the given problem.

2) Which systems thinking skills do the subjects already have?

On STS_pre test, 21 % of the subjects were able to answer the questions related to stock-flow thinking and feedback thinking skills, while 17% of the subjects were able to answer the question related to realizing and expressing delay.

3) Are there any significant differences between systems thinking skills of the subjects in the experimental and the comparison group right after the interventions and after six-month period?

The experimental group scored significantly higher on STS_post test than the comparison group at .05 significance level (t (40) = 2.74, p =.009, d = .84). However, the difference was not reported as enduring (t (40) = .548, p =.587, d = .17) at .05 significance level.

4) Are there any significant differences between conceptualizations of dynamic environmental problems of the subjects in the experimental and the comparison group after the interventions and six-month period?

The experimental group scored significantly higher on DES test than the comparison group at .05 significance level (t (40) = 3.04, p = .004, d = .94) and there was still a statistically significance difference between the groups (t (40) = 2.10, p = .042, d = .64).at .05 significance level after six-month period.

5) Are there any significant differences between science achievement level on environmental questions of the subjects in the experimental and the comparison group right after the intervention and after six-month-period?

There was no significant difference between the groups on SAT (U = 246, p = .52, d = .19) and SAT_del tests (t (40) = .71, p = .486, d = .22) at .05 significance level.

6) How do the participants in both groups understand and verbalize the structures of some dynamic environmental questions?

Based on the interviews, it was found that more participants from the experimental group were able to explain the bluefish population problem on a stock-flow basis and the construction of the 3rd Bridge problem on a feedback loop basis than the participants from the comparison group.

7) Are the subjects in the experimental group able to transfer knowledge and skills between two environmental tasks with similar dynamic structure after the intervention and after six-month-period?

The experimental group scored significantly higher on DES_f questions than the comparison group at .05 significance level ($t(40) = 2.62$, $p = .012$, $d = .81$). However, the difference is statistically significant ($t(40) = 1.74$, $p = .89$, $d = .54$) between groups at .10 significance level after six-month period.

8) Are the subjects in the experimental group able to analyze more general environmental problems that are dynamic and complex after the intervention and after six-month-period?

The experimental group scored significantly higher on DES_unf questions than the comparison group at .05 significance level ($U = 116.50$, $p = .008$, $d = .85$). The difference is still statistically significant ($U = 61.5$, $p < .001$, $d = 1.63$). at .05 significance level with a very big effect size.

After answering to all the research questions, it is time to answer the question about the main motivation of the study:

Whether systems approach provides efficient means to teach dynamic environmental issues to seventh grade students?

The answer is yes. The systems based intervention enables students to

- Realize changes in time within an ecosystem
- Represent changes in an ecosystem and its components in time with a variety of tools (behavior of time graphs, causal loop diagram, stock-flow diagrams, and dynamic computer models)

It was found that the subjects in the experimental group were able to conceptualize dynamic and complex environmental issues. Another important contribution of this study to the field environmental education is that teaching children food chains in a loop fashion makes more sense for children rather than emphasizing a linear, one-way relationship between organisms on a food chain. To teach food chains in a loop fashion is important to teach basic population dynamics by taking into account changes in populations and natural resources over time and their reverse effects on one another.

Recent news about revised curricula for 1st to 12th grades for all subject matters was disseminated in February, 2013. According to the recent Science and Technology Curriculum, the time spent for "Human and Environment" unit has been decreased from 14 to 10 lesson hours and the number of objectives has been decreased from 13 to four objectives. The revised unit does not contain the section devoted to environmental problems (TTKB, 2013). Besides, the revised unit contains solely factual knowledge about environmental terminology and does not contain any dynamic issues in contrast to the nature of environmental issues. These revisions seem to be very contradictory with today's needs. This is the most evident proof that environmental issues are underestimated in Turkey.

To conclude, even one-month instruction with system-based approach on “Human and Environment” unit helped 12-14 years old students to develop systems thinking skills on environmental issues. The next step would be to train teachers from different backgrounds on systems thinking and system dynamics and on its application on education.

REFERENCES

Ackoff, R. L., 1994. Systems thinking and thinking systems. System Dynamics Review, 10, 2-3, 175-188.

Anderson, L. W., & Burns, R. B. (1987). Values, evidence, and mastery learning. Review of Educational Research, 57, 215–223.

Assaraf, B. O. , Orion, N., 2005. Development of systems thinking skills in the context of Earth system education. Journal of Research in Science Teaching, 42, 5, 518-560.

Barlas, Y., 2002. System Dynamics: Systemic Feedback Modeling for Policy Analysis" in Knowledge for Sustainable Development - An Insight into the Encyclopedia of Life Support Systems, UNESCO-Eolss Publishers, Paris, France, Oxford, UK., pp.1131-1175.

Bernard, H.R., Ryan, G.W., 2010. Analyzing Qualitative Data: Systematic Approaches. Sage Publications, California, USA.

Black, T. R. , 1999. Doing quantitative research in the social sciences : An integrated approach to research design, measurement and statistics. Sage Publications, London, UK.

Brown, G. S.,1992. Improving education in public schools: innovative teachers to the rescue. System Dynamics Review, 8 (1), 83-89.

Capra, F., 1997. Web of Life- A New Scientific Understanding of Living Systems. Anchor Books, NY, USA.

Capra, F. 1998. Ecology, Systems Thinking and Project-Based Learning. Paper presented at the Sixth Annual Conference on Project-Based Learning, 14th March 1998, San Francisco.
http://www.coopecology.com/Coop_Ecology/Download_Documents_files/Ecology,%20Systems%20Thinking,%20%26%20Project-Based%20Learning.pdf (accessed June 2010).

Capra, F., 2005. Speaking Nature's Language: Principles for Sustainability. In Stone, M. K. & Barlow, Z. (Eds.). Ecological Literacy- Educating Our Children for a Sustainable World. Sierra Club Books, San Francisco, USA.

Chesire, G., 2007. Mountains and Deserts [*Turkish: Dağlar ve Çöller*]. NTV Yayınları, İstanbul, Turkey.

Creswell, C. W., 2003. Research design : qualitative, quantitative, and mixed methods approaches. Sage Publications, California, USA.

DeBruine, L. ,2011. Reporting statistics in psychology- Lecture notes.
Retrieved from :
http://facelab.org/debruine/Teaching/Meth_A/files/Reporting_Statistics.pdf (accessed on March 2013)

Doğança, Z., 2007. Developing environmental education program for primary school students and assessing its effects on prospective science teachers. M.S. Thesis, Bogazici University.

Fadem, T. J., 2009. The Art of Asking- Ask Better Questions, Get Better Answers. Pearson Education, Inc. New Jersey, USA.

Feurzeig, W., Roberts, N., 1999. Modeling and Simulation in Science and Mathematics Education, Springer, NY, USA.

Forrester, J. W., 1994. System dynamics, systems thinking and soft OR. System Dynamics Review, 10, 2-3, 245-256.

Forrester, J. W., 1996. System Dynamics and K-12 Teachers --*A lecture at the University of Virginia School of Education*
http://sysdyn.clexchange.org/sdep/Roadmaps/RM1/D-4337.pdf (accessed on June 2010).

Gay, L. R., Airasian, P. ,2006. Educational Research- Competencies for Analysis and Applications. (8th Edition). Merrill Prentice Hall, Ohio, USA.

George and Mallery, 2005. SPSS for Windows Step by Step- A Simple Guide and Reference. (5th Edition). Allyn and Bacon, Boston, USA.

Giusti, G. A., 2009. Human influences to Clear Lake, California. Retrieved from: http://www.lakecountywinegrape.org/growers/growersfiles/Human%20Influences%20on%20Clear%20Lake%202009.pdf (accessed on February 2011).

Grotzer, T. A. , Basca, B. B., 2003. How does grasping the underlying causal structures of ecosystems impact students' understanding? Journal of Biological Education, 38, 1, 16-29.

Gribbons, B., Herman, J., 1997. True and quasi-experimental designs. Practical Assessment, Research & Evaluation, 5(14). http://pareonline.net/getvn.asp?v=5&n=14 (accessed on September 2010)

Güvender Yayıncılık, 2009. 7th Grade 100% SBS Science and Technology Test Book *[Turkish: 7. Sınıf %100 SBS Fen ve Teknoloji Soru Bankası]* Istanbul, Turkey.

Hopper, M., Stave, K. A., 2008. Assessing the effectiveness of systems thinking interventions in the classroom. Proceedings of the 26th International Conference of the System Dynamics Society, Athens, Greece, July 20-24, 2008. http://www.systemdynamics.org/conferences/2008/proceed/index.htm (accessed on June 2010).

Huck. S. W. 2012. Reading Statistics and Research. (6th Edition). Allyn & Bacon, Boston, USA.

Laerd Statistics https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide.php (accessed on January 2013)

Lyneis, D. A., Fox-Melanson, D., 2001. The challenges of infusing system dynamics into a K-8 curriculum, 19th International Conference of the Systems Dynamics Society, Atlanta, Georgia, 23-27 July, 2001. www.systemdynamics.org/conferences/2001/papers/Lyneis_1.pdf_ (accessed on June 2010)

Mandinach, E. B. & Cline, H. F., 1994. Classroom Dynamics: Implementing a Technology-Based Learning Environment. Lawrence Erlbaum Assoc., New Jersey, USA.

McGraw, K. O. & Wong, S.P.,1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1; 30–46.

Ministry of National Education, 2005. The Guide for Primary School Science and Technology Curriculum-6th, 7th, and 8th Grades. Board of Governmental Books, Ankara, Turkey.

Ministry of Education Teacher Portal- *"Talim Terbiye Kurulu Başkanlığı (TTKB) Öğretmen Portalı*"- http://ttkb.meb.gov.tr/www/guncellenen-ogretim-programlari-ve-kurul-kararlari/icerik/150 (accessed on March 2013)

MIT, 1997. Roadmaps for System Dynamics in Education Project. Retrieved from Creative Learning Exchange- System Dynamics and Systems Thinking in K-12 Education Website: http://www.clexchange.org/curriculum/roadmaps.asp (accessed on February 2010).

Moxnes, E. & Saysel, A. K., 2009. Misperceptions of global climate change: information policies. Climate Change, 93; 15-37.

Nuhoğlu, H., 2008. Studying Effects of Systems Approach on Attitude, Achievement, and Different Skills in Science and Technology Lesson, *(Turkish: İlköğretim Fen ve Teknoloji Dersinde Sistem Dinamiği Yaklaşımının Tutum, Başarıya ve Farklı Becerilere Etkisinin Araştırılması),* PhD Dissertation, Gazi University.

Nuhoğlu, H., Nuhoğlu M., 2007. System dynamics approach in science and technology education. Journal of Turkish Science Education (TUSED), 4, 2, 91-108.

Oran Yayıncılık, 2008. 7th Grade SBS Science and Technology Test Book *[Turkish:7. Sınıf SBS Fen ve Teknoloji Soru Bankası]*. İzmir, Türkiye.

Ossimitz, G. Stock-flow thinking and reading stock-flow related graphs: An empirical investigation in dynamic thinking abilities. 20th International Conference of the Systems

Dynamics Society, Palermo, Italy, July 28- August 1, 2002. http://wwwu.uni-klu.ac.at/gossimit/pap/sfthink.pdf (accessed on June, 2010).

Patton, M. Q. 1983. Qualitative Evaluation Methods. (4th Edition). Sage Publications, California, USA.

Perkins, D. N., & Salomon, G., 1992. Transfer of learning.*International Encyclopedia of Education* (2nd Edition). Oxford, UK: Pergamon Press.

Quaden, R., Ticotsky, A. & Lyneis, D., 2008. The Shape of Change: Stocks and Flows. (Revised Edition). Creative Learning Exchange; Massachusetts, USA.

Rainforest Alliance: Facts about Tropical Rainforests http://www.rainforest-alliance.org/ (accessed on May, 2013).

Razali, N. M. & Wah, Y. B. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

Riess, W. & Mischo, C., 2010. Promoting systems thinking through biology lessons. International Journal of Science Education, 32, 6, 705-725.

Senge, P. M., 1990. The Fifth Discipline-The Arts & Practice of the Learning Organization, Doubleday/Currency, New York, USA.

Schunk, D. H., 2012. Learning theories : an educational perspective. Pearson Inc., Boston, USA.

Stave, K.A., Hopper, M., 2007. What constitutes Systems Thinking? A proposed taxonomy. 25th International Conference of the System Dynamics Society, July 29-August 2, 2007, Boston. Retrieved from: http://www.systemdynamics.org/conferences/2007/proceed/papers/STAVE210.pdf (accessed on January 2010)

Sterman, J. D., 1994. Learning in and about complex systems. System Dynamics Review, 10, 2-3, 291-330.

Sterman, J. D., 2000. Business Dynamics- Systems Thinking and Modeling for a Complex World, McGraw Hill, Boston, USA.

Sterman, J. D., Sweeney, L. B., 2002. Cloudy skies: assessing public understanding of global warming. System Dynamics Review, 18, 2, 207- 240.

Sweeney, L. B., Sterman, J. D., 2000. Bathtub dynamics: initial results of a systems thinking inventory. System Dynamics Review, 16, 4, 249-286.

Sweeney, L. B., Sterman, J. D., 2007. Thinking about systems: student and teacher conceptions of natural and social systems. System Dynamics Review, 23, 2-3, 285-312.

Stone, M. K. , Barlow, Z., 2005. (Eds.). Ecological Literacy- Educating Our Children for a Sustainable World. Sierra Club Books, San Francisco, USA.

Torchim, W. M. K., n.d. Types of Reliability. Retrieved from: http://www.socialresearchmethods.net/kb/reltypes.php (accessed on January, 2011).

Tunç, T., Bağcı, N., Yörük, N., Gürsoy Köroğlu, N., Çeltikli Altunoğlu, Ü., Başdağ, G., Keleş, Ö., İpek, İ., Bakar, E., 2011. Science and Technology 7th Grade Course Book [*Fen ve Teknoloji 7.Sınıf Ders Kitabı*]. MEB Devlet Kitapları, Ankara, Turkey.

Waters Foundation, n.d. Systems Thinking in Schools Project Website http://watersfoundation.org/resources/be-nice-to-spiders/ (accessed on February 2010).

Yıldıran, G., 2006. Multicultural aplications of mastery learning : our thoughts our deeds and our hopes for education. Bogazici University Press, Istanbul, Turkey.

Zaraza, R. & Fisher, D. M., 1999. Training System Modelers: the NSF CC-STADUS and CC-SUSTAIN Projects. In Feurzeig, W., Roberts, N. (Eds.), Modeling and Simulation in Science and Mathematics Education, Springer, NY, USA.

Zientek, L. R., Yetkiner-Ozel, E.Z., Ozel, S., Allen, J.,2012. Reporting Confidence Intervals and Effect Sizes: Collecting the Evidence. Career and Technical Education Research, 37, 3, 277-295.