

I

J

A

T

E

Volume 7

Issue 1

2020

International Journal of  
Assessment Tools in Education

International Journal of  
Assessment Tools in Education

International Journal of  
Assessment Tools in Education

<https://dergipark.org.tr/en/pub/ijate>

e-ISSN: 2148-7456



e-ISSN 2148-7456

<http://www.ijate.net/index.php/ijate/index>

**Volume 7**

**Issue 1**

**2020**

**Dr. İzzet KARA**

Publisher

International Journal of Assessment Tools in Education  
&  
Pamukkale University,

Education Faculty,

Department of Mathematic and Science Education,

20070, Denizli, Turkey

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : [ijate.editor@gmail.com](mailto:ijate.editor@gmail.com)

Frequency : 4 issues per year starting from June 2018 (March, June, September, December)

Online ISSN: 2148-7456

Website : <http://www.ijate.net/index.php/ijate>  
<http://dergipark.org.tr/en/pub/ijate>

Design & Graphic: IJATE

**Support Contact**

Dr. İzzet KARA

Journal Manager & Founding Editor

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : [ikara@pau.edu.tr](mailto:ikara@pau.edu.tr)

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).



## International Journal of Assessment Tools in Education

*International Journal of Assessment Tools in Education* (IJATE) is an international, peer-reviewed online journal. IJATE is aimed to receive manuscripts focusing on evaluation and assessment in education. It is expected that submitted manuscripts could direct national and international argumentations in the area. Both qualitative and quantitative studies can be accepted, however, it should be considered that all manuscripts need to focus on assessment and evaluation in education.

IJATE as an online journal is sponsored and hosted by **TUBITAK-ULAKBIM** (The Scientific and Technological Research Council of Turkey).

There is no submission or publication process charges for articles in IJATE.

### **IJATE is indexed in:**

- Emerging Sources Citation Index (ESCI) (Web of Science Core Collection)
- Education Resources Information Center (ERIC),
- TR Index (ULAKBIM),
- ERIH PLUS,
- DOAJ,
- Index Copernicus International
- SIS (Scientific Index Service) Database,
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,
- International Scientific Indexing

### **Editors**

*Dr. Eren Can Aybek, Pamukkale University, Turkey*

*Dr. Özen Yıldırım, Pamukkale University, Turkey*

### **Editorial Board**

*Dr. Beyza Aksu Dunya, Bartın University, Turkey*

*Dr. R. Şahin Arslan, Pamukkale University, Turkey*

*Dr. Murat Balkıs, Pamukkale University, Turkey*

*Dr. Gülşah Başol, Gaziosmanpaşa University, Turkey*

*Dr. Bengü Börkan, Boğaziçi University, Turkey*

*Dr. Kelly D. Bradley, University of Kentucky, United States*

*Dr. Okan Bulut, University of Alberta, Canada*

*Dr. Javier Fombona Cadavieco, University of Oviedo, Spain*

*Dr. William W. Cobern, Western Michigan University, United States*

*Dr. R. Nükhet Çıkrıkçı, İstanbul Aydın University, Turkey*

*Dr. Safiye Bilican Demir, Kocaeli University, Turkey*

*Dr. Nuri Doğan, Hacettepe University, Turkey*

*Dr. Selahattin Gelbal, Hacettepe University, Turkey*

*Dr. Anne Corinne Huggins-Manley, University of Florida, United States*

*Dr. Violeta Janusheva, "St. Kliment Ohridski" University, Republic of Macedonia*

*Dr. Francisco Andres Jimenez, Shadow Health, Inc., United States*

*Dr. Nicole Kaminski-Öztürk, The University of Illinois at Chicago, United States*

*Dr. Orhan Karamustafaoglu, Amasya University, Turkey*

*Dr. Yasemin Kaya, Atatürk University, Turkey*

*Dr. Hulya Kelecioğlu, Hacettepe University, Turkey*

*Dr. Hakan Koğar, Akdeniz University, Turkey*

*Dr. Sunbok Lee, University of Houston, United States*

*Dr. Froilan D. Mobo, Ama University, Philippines*

*Dr. Hamzeh Moradi, Sun Yat-sen University, China*

*Dr. Nesrin Ozturk, Ege University, Turkey*

*Dr. Turan Paker, Pamukkale University, Turkey*

*Dr. Abdurrahman Sahin, Pamukkale University, Turkey*

*Dr. Murat Dogan Sahin, Anadolu University, Turkey*

*Dr. Ragip Terzi, Harran University, Turkey*

*Dr. Hakan Türkmen, Ege University, Turkey*

*Dr. Hossein Salarian, University of Tehran, Iran*

*Dr. Kelly Feifei Ye, University of Pittsburgh, United States*

### **English Language Editors**

*Dr. Hatice Altun, Pamukkale University, Turkey*

*Dr. Çağla Atmaca, Pamukkale University, Turkey*

Dr. Sibel Kahraman, *Pamukkale University*, Turkey

Arzu Kanat Mutluođlu - *Pamukkale University*, Turkey

**Copy & Language Editor**

Anıl Kandemir, *Middle East Technical University*, Turkey

## Table of Contents

### *Research Article*

---

1. [Assessing Skills of Identifying Variables and Formulating Hypotheses Using Scenario-Based Multiple-Choice Questions](#) / Pages : 1-17 / [PDF](#)  
Burak TEMİZ
2. [Use of Item Response Theory to Validate Cyberbullying Sensibility Scale for University Students](#) / Pages : 18-29 / [PDF](#)  
Osman Tolga ARICAK, Akif AVCU, Feyza TOPÇU, Merve Gülçin TUTLU
3. [Investigation of PISA 2015 Reading Ability Achievement of Turkish Students in Terms of Student and School Level Variables](#) / Pages : 30-42 / [PDF](#)  
Kazım ÇELİK, Ahmet YURDAKUL
4. [The Young Adults Form of the Attitude toward Women's Working Scale: Development, Preliminary Validation and Measurement Invariance](#) / Pages : 43-61 / [PDF](#)  
Devrim ERDEM
5. [Physics Course Attitudes Scale for High School Students: A Validity and Reliability Study](#) / Pages : 62-72 / [PDF](#)  
Hülya ÇERMİK, İzzet KARA
6. [Comparison of Passing Scores Determined by The Angoff Method in Different Item Samples](#) / Pages : 80-97 / [PDF](#)  
Hakan KARA, Sevda ÇETİN
7. [What You might not be Assessing through a Multiple Choice Test Task](#) / Pages : 98-113 / [PDF](#)  
Burcu KAYARKAYA, Aylin ÜNALDI
8. [Scaling of Mood-State and Sample Cases Causing Anger in a Relationship with Rank-Order Judgment and Classifying Judgment](#) / Pages : 114-129 / [PDF](#)  
Merve YILDIRIM SEHERYELİ, Duygu ANIL
9. [A Study on Developing Scale for Teacher Perceptions towards Spelling Rules](#) / Pages : 130-144 / [PDF](#)  
Ali TURKEL

### *Rapid Communications*

---

1. [When interviewing: how many is enough?](#) / Pages : 73-79 / [PDF](#)  
William COBERN, Betty ADAMS

## Assessing Skills of Identifying Variables and Formulating Hypotheses Using Scenario-Based Multiple-Choice Questions

Burak Kağan Temiz<sup>1,\*</sup>

<sup>1</sup>Department of Science Education, Faculty of Education, Nigde Omer Halisdemir University, Nigde, Turkey

### ARTICLE HISTORY

Received: 08 May 2019

Revised: 24 December 2019

Accepted: 06 January 2020

### KEYWORDS

Science Process Skills,  
Multiple-Choice Items,  
Science Education

**Abstract:** The aim of this study was to investigate the effectiveness of scenario-based multiple-choice questions to assess students' science process skills. To achieve this objective, a test with 32 scenario-based multiple-choice questions evaluating students' skills in formulating hypotheses and identifying variables was prepared and administered to 370 high school freshmen. The questions were involved experiments with two different parts. Both parts of the experiments had the same dependent variable, and in each part the effect of a different manipulated variable on the dependent variable was examined. Therefore, the variables changed roles within the same experiment. In evaluating the test, questions about the first part of the experiments were coded A, and questions about the second part of the experiments were coded B. When the students' scores from the code A and code B items were compared, statistically significant differences were found. Analysis of the data revealed that some students were affected by the different roles played by the variables in the different parts of the experiment.

## 1. INTRODUCTION

As a result of their natural curiosity, human beings seek to understand the environment in which they live and to acquire new knowledge. The natural sciences that emerged as a result of these efforts embody two main components: the scientific knowledge itself, and the ways in which knowledge can be acquired. The skills that are used for acquiring knowledge in science are called science process skills (SPS). SPS are thus the activities that scientists engage in when they investigate a problem or phenomenon. SPS are mental and physical skills used in collecting, organizing and analyzing data through various methods. These skills are involved in identifying researchable questions, designing investigations, obtaining evidence, interpreting evidence in terms of the question addressed in the research, and communicating the findings of the investigative process. In addition, SPS are needed not only by scientists, but by all citizens in order for them to become scientifically literate people able to function in a society in which science plays a major role and has an impact on everyone's personal, social and global life. In fact, understanding scientific processes is a basic aspect of thinking, used both in science and in other fields to solve problems. For this reason, SPS are also life-long learning skills. In

**CONTACT:** Burak Kağan TEMİZ ✉ [bktemiz@ohu.edu.tr](mailto:bktemiz@ohu.edu.tr) 📍 Department of Science Education, Faculty of Education, Nigde Omer Halisdemir University, Nigde, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

elementary and middle school science education, the development of SPS is a major goal of science education.

In the literature different researchers have defined SPS differently (Gabel, 1992; Martin, 2002; Padilla, 1990). In general, SPS are the cognitive skills that we use to process information, to think through problems, and to formulate conclusions. These are the skills that scientists use when they work. By teaching students these important skills, we can enable them to understand their world and learn about it. These skills are fundamental to thinking and to research in science. In the Science-A Process Approach (SAPA), these skills are defined as a set of broadly transferable abilities, appropriate to many science disciplines and reflecting the behavior of scientists. The SAPA has grouped process skills into two categories, basic and integrated. The basic science process skills (BSPS) provide the intellectual ground work of scientific enquiry, such as the ability to order and describe natural objects and events. The BSPS are fundamental to the integrated science process skills (ISPS). The BSPS include observing, classifying, measuring and predicting, while the ISPS are essential skills for solving problems or conducting science experiments. The ISPS include identifying and defining variables, collecting and transferring data, constructing tables of data and graphs, describing relationships between variables, interpreting data, manipulating materials, formulating hypotheses, designing investigations, drawing conclusions and generalizing (Abruscato, 2000; Beaumont-Walters & Soyibo, 2001; Burns, Okey, & Wise, 1985; Carin, 1993; Carin & Bass, 2001; Esler & Esler, 2001; Harlen, 1993, 1999; Hughes & Wade, 1993; Ostlund, 1992; Rezba et al., 1995).

### **1.1. Formulating Hypotheses and Identifying Variables**

When we try to understand things in a scientific way, the complex subject at hand is divided into researchable and understandable elements. These elements of an event or a system are called variables. Variables are the factors, conditions or relations that change or that can be changed in an event or a system. In scientific research, there are three kinds of variables. These are manipulated, responding, and controlled variables (Bailer et al., 1995). A manipulated variable (independent variable) is a factor or a condition which is changed by the researcher on purpose in an experiment. A dependent variable (response variable) is a kind of variable that can be affected by the changes in the factor or the condition. Variables that remain constant through the experiment so as not to interfere with the results are called controlled variables. There can be more than one controlled variable in an experiment.

Formulating a hypothesis is the skill of developing a problem question which can be tested by an experiment about the effect of a manipulated variable on a dependent variable. To formulate a hypothesis means building testable statements based on ideas and experiences which are thought to be true. Hypothesizing means stating a testable solution to a problem. A hypothesis is usually proposed before any experiment or research and is a prediction about the relationships between variables. Being testable is the most important characteristic of a hypothesis.

According to Gabel (1993), a scientist must control all the variables that will affect the outcome of an experiment in order to be able to practice science, that is, to be able to test hypotheses or confirm assumptions. Before controlling variables, the scientist must identify the responding and manipulated variables. Later, a factor is changed on purpose and, as a result, a change occurs in the other variable. The strategy followed in manipulating and controlling variables is to change a variable (the manipulated variable) and examine changes occurring in the other variable (the response variable). At the same time, many other variables (controlled variables) must be defined and kept constant. This is because these variables have the potential to affect the results. If more than one variable is changed at the same time, the result of the experiment is not reliable (Carin & Bass, 2001). Bailer et al. (1995) associated the process of hypothesizing with the process of identifying and controlling variables. On this basis, a hypothesis is a kind of statement that predicts the effect of one variable on another.



## 1.2. Assessing Science Process Skills

With increased understanding of the importance and value of SPS in science education, the interest of researchers in the subject has also increased. Numerous models have been constructed for the teaching and acquisition of SPS. Additionally, several instruments have been developed to assess achievement in SPS for formative, summative and monitoring purposes. An examination of the literature reveals that numerous tests with various question formats have been developed in order to measure all or some SPS at different levels. [Table 1](#) shows some of these instruments.

As seen in [Table 1](#), most of the SPS assessment instruments were designed using a multiple-choice format, which is relatively easier and less time-consuming to administer. However, several researchers have emphasized the need to develop such instruments using alternative formats. Techniques suggested include systematic observations of students' laboratory work (Lunetta et al.1981), microcomputer simulations (Berger, 1982), technological applications (Kumar, 1996), and open-ended questions (Gabel, 1993). Moreover, Beaumont-Walters and Soyibo (2001) drew attention to the fact that although the commonly used multiple-choice test format has been criticized, only a few researchers have attempted to develop tests for SPS that also involve hands-on tasks. And although considerable attention has been given to assessing the performance of SPS, the development of standardized instruments for participants in a large sample has been difficult. In light of these difficulties, the multiple-choice format may be preferable for large samples (Aydınli et al., 2011).

**Table 1.** SPS Assessment Instruments Documented in the Research Literature.

Authors	Title	Year	Test Format
R. S. Tannenbaum	Test of Science Processes	1968	Multiple choice
J. W. Riley	The Test of Science Inquiry Skills	1972	Multiple choice
R. R. Ludeman	The Science Process Test	1974	Multiple choice
L. L. Molitor and K. D. George	The Science Process Test	1975	Multiple choice
F. G. Dillashaw and J. R. Okey	Test of Integrated Process Skills	1980	Multiple choice
K. G. Tobin and W. Capie	Test of Integrated Process Skills	1982	Multiple choice
J. C. Burns, J. R. Okey and K. C. Wise	Test of Integrated Process Skills II	1985	Multiple choice
K. A. Smith and P. W. Welliver	Science Process Assessments for Elementary School Students	1986	Multiple choice
K. A. Smith and P. W. Welliver	Science Process Assessments for Middle School Students	1994	Multiple choice
G. Solano-Flores	The "Bubbles" Task	2000	Hands-on Activity
Y. Beaumont-Walters and K. Soyibo	Test of Integrated Science Process Skills	2001	Multiple Format
Author, M. F. Taşar and M. Tan	Multiple Format Test of Science Process Skills	2006	Multiple Format
Author and M. Tan	Science Process Skills Test	2007	Multiple Format
Shahali E. H. M. and Halim L Feyzioglu, B., Demirdag, B., Ak yıldiz, M., & Altun, E.	Test of Integrated Science Process Science Process Skills Test	2010 2012	Multiple choice Multiple choice
Aydoğdu B., Tatar N., Yıldız E. and Buldur S.	Science Process Skills Scale	2012	Multiple choice
Aydoğdu, B. and Karakuş, F.	The Scale for Basic Process Skills of Pre-School Students	2017	Multiple choice
Tosun, C	Scientific Process Skills Test	2019	Multiple choice

### 1.3. Scenario-Based Learning and Assessment

Scenarios are narratives in the form of stories or speeches that emerge from real events or realistic situations. In scenario-based learning the real world is brought into the classroom. Thus, students are given opportunities to think about a problem, to use what they have learned in real or realistic situations, to become aware of their lack of knowledge and to do the necessary work to correct this. Furthermore, scenarios trigger students' higher-order thinking processes such as analysis, synthesis, evaluation and decision-making (Açıkgöz, 2003).

The increasing importance of scenario-based learning in recent years has brought new approaches to the teaching process, and scenarios are now included in many Science and Technology textbooks. With scenario-based learning, students are given the opportunity to discover different problems and situations through scenarios drawn from real life, to use their existing knowledge in these new situations, to offer creative ideas and to implement what they have learned (Erduran Avcı & Bayrak, 2013). Scenarios unique to a specific field can be used in activities involving measurement and evaluation in addition to normal learning activities. According to Thalheimer (2013), scenario-based questions present learners with one or more short paragraphs that describe a situation and include a question that asks learners to make their own decisions. There are many varieties to this basic design. We can use multiple scenes and multiple questions to form a scenario. We can add visual or auditory details to augment or even supplant the text-based scenario. We can also use different types of questions, including multiple-choice, open-ended, and yes-no questions, etc. Scenario-based multiple-choice test items have been used frequently in SPS assessments.

When the multiple-choice tests developed to assess skills in identifying variables and formulating hypotheses are examined, scenario-based questions are frequently encountered. Some of the tests used most frequently in science education research are the Test of Integrated Process Skills (TIPS) (Dillashaw & Okey, 1980; Tobin & Capie, 1982), the Test of Integrated Process Skills II (TIPSII) (Burns et al., 1985), and the Science Process Assessments for Middle School Students (Smith & Welliver, 1995). An examination of items in the multiple-choice format SPS measurement tests used to assess skills in formulating hypotheses and identifying variables shows that question developers generally provide one section from a single-stage experiment and ask the student to identify the hypothesis and the variables in the test. The example given in [Figure 1](#), which is a single-stage experiment, is from an SPS measurement test widely used in Turkey.

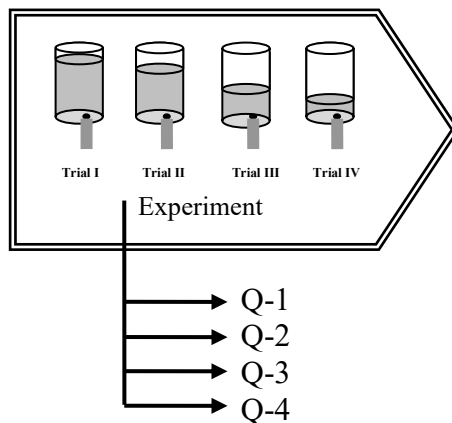
Most experiments conducted in science consist of more than one stage. At each stage the effects of a different manipulated variable on a dependent variable is examined. Therefore, the manipulated variable at one stage of the experiment can be a controlled variable at another stage. The idea that the same variable can play different roles in different parts of the experiment should be taken into consideration while developing questions to assess SPS. In the scenarios in the SPS measurement tests widely used in the literature, the idea that an experiment may be made up of more than one stage is not taken into consideration (see [Figure 2](#)). Does students' performance change if they are asked questions (see [Figure 3](#)) about situations where the effect of a different manipulated variable on a dependent variable at each different stage of the experiment is examined? The aim of this study was to find the answer to this question.

**Answer questions 29, 30, 31 and 32 by reading the paragraph given below.**

The effects on tomato production of leaves mixed in with the soil are being investigated. In the research an identical quantity and type of soil was placed in four large pots. However, 15 kg of mulched leaves were added to the first pot, 10 kg to the second and 5 kg to the third. No mulched leaves were added to the fourth pot. Tomatoes were then planted in these pots. All the pots were placed in sunlight and watered identically. Tomatoes obtained from each pot were weighed and recorded.

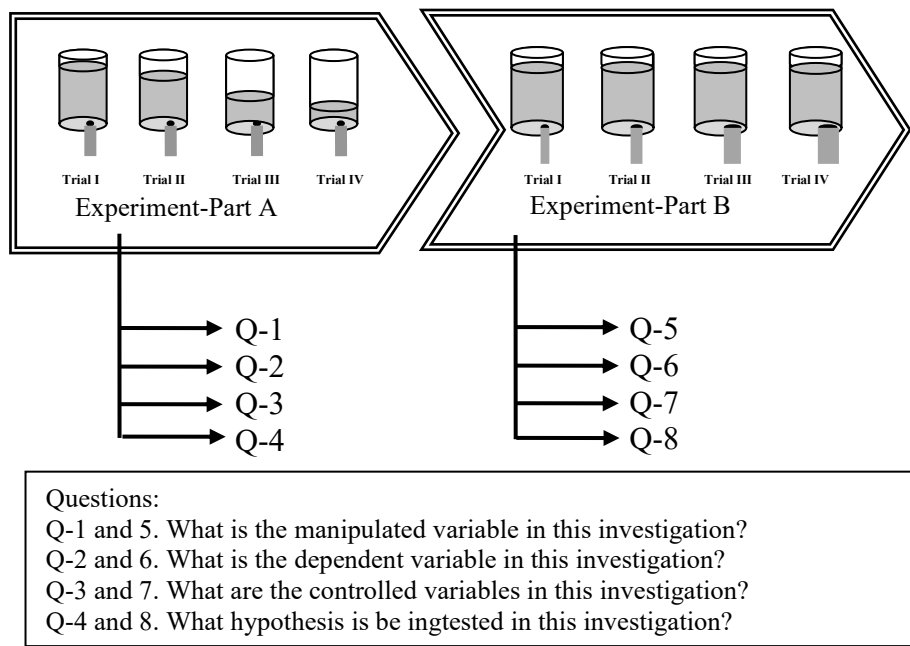
29. What is the hypothesis that was tested in this research?
  - a. Plants produce tomatoes in proportion to the sunlight they receive.
  - b. The larger the pots are the more mixed leaves are needed.
  - c. The more water in the pots, the faster the leaves rot.
  - d. The more mulched leaves are in the soil the more tomatoes are produced.
  
30. What is/are the controlled variable(s) in this research?
  - a. The amount of tomatoes obtained from each pot.
  - b. The amount of leaves mixed into the pots.
  - c. The amount of soil in the pots.
  - d. The number of pots with mulched leaves added.
  
31. What is the dependent variable in this research?
  - a. The amount of tomatoes obtained from each pot.
  - b. The amount of leaves mixed in the pots.
  - c. The amount of soil in the pots.
  - d. The number of pots with mulched leaves added.
  
32. What is the manipulated variable in research?
  - a. The amount of tomatoes obtained from each pot.
  - b. The amount of leaves mixed in the pots.
  - c. The amount of soil in the pots.
  - d. The number of pots with mulched leaves added.

**Figure 1.** Sample item in a scenario from a single-stage experiment



Questions:  
 Q-1. What is the manipulated variable in this investigation?  
 Q-2. What is the dependent variable in this investigation?  
 Q-3. What are the controlled variables in this investigation?  
 Q-4. What hypothesis is being tested in this investigation?

**Figure 2.** Traditional scenario-based SPS questions about single stage experiments



**Figure 3.** Scenario-based SPS questions about two stage experiments

This study was conducted with the aim of examining whether the students' ability to use SPS (formulating hypotheses and identifying variables) would change with questions about two-part experiments where a different hypothesis was tested in each part and where variables played different roles in different parts. As scenario-based questions are frequently used in the literature to assess the skills of identifying variables and formulating hypotheses, these were the skills that this study examined.

## 2. METHOD

This study uses a type of descriptive research model with a survey method. Descriptive models that are used commonly aim to describe the situation and find out the factors that are the subjects of the study. The survey type methods contain collecting, classifying, describing, analyzing and inferring results from the data which aim to determine any presence and/or degree of together-change amongst two or more variables (Büyüköztürk et al., 2009; Karasar, 2011).

### 2.1. Study Group

370 (191 females, 179 males) high school freshmen selected by stratified sampling from five different high schools participated in this study. The majority of students were 15 years old. The participants had just completed their elementary education and had not yet chosen any future field of study.

### 2.2. Data Collection Process and Assessment Tool

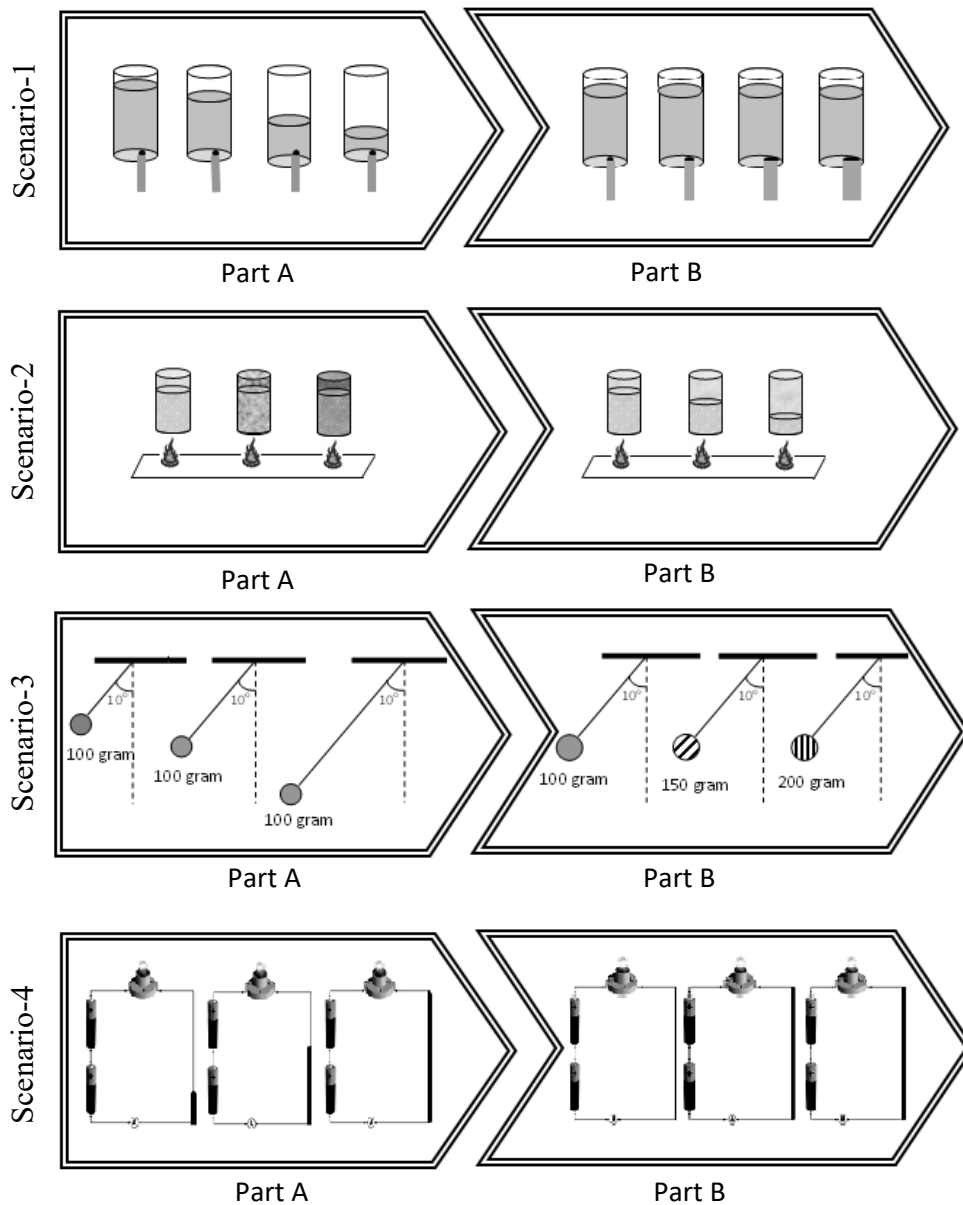
In order to measure students' skills in identifying variables and formulating hypotheses, a test with scenario-based multiple-choice items was used. The 40 items in this test were compiled from The Science Process Skills Test (SPST) question pool developed by the author (Temiz, 2007). The SPST was developed for the purpose of assessing skills in identifying variables, formulating hypotheses, controlling variables, recording data (constructing the data table), constructing graphs and interpreting graphs. The SPST is composed of three multiple-choice and three open-ended modules, with a total of six modules. Module 1 assesses the skills of defining variables and formulating hypotheses and has 60 multiple-choice questions; Module 2

assesses the skill of controlling variables (designing experiments) and has five open-ended and 25 multiple-choice questions; Module 3 assesses the skill of constructing a data table and has eight open-ended questions; Module 4 assesses the skill of drawing graphs and has eight open-ended questions; Module 5 assesses the skills of interpreting graphs and has 60 multiple choice questions; and Module 6 assesses the skills of defining the variables and formulating hypotheses and has 10 open-ended questions. The SPST was developed after pilot tests conducted on 1584 Grade 9 students. To collect evidence for the test's validity, content-related, criterion-related and construct-related validity analyses were conducted, and internal-consistency, test-retest and inter-rater agreement analyses were carried out to determine the SPST's reliability. Detailed statistics about test development process can be found in the work "Evaluating students' science process skills in physics teaching" (Temiz, 2007).

The data in this work was collected using 40 multiple-choice questions from among the questions in Module 1 of the SPST. These are related to five experimental scenarios which are individually made of two parts. Each experiment is presented with a paragraph of text and pictures supporting that text, followed by multiple-choice questions come based on what is given. This test was named the Formulating Hypotheses and Identifying Variables Skills (FHIVS) Test.

To examine the test reliability and item indices, the FHIVS test was administered to high school students. A total of 87 students were involved in this pilot test. Students' answers were processed with the Excel software package, and test reliability was investigated by internal consistency analyses. The total scores of the five experimental scenarios test ranged from 4 to 40 (mean=23.9, S.D.=10.6) for the students overall. The total test reliability (KR 20 coefficient) was 0.944. Item difficulty indices ranged from 0.25 to 0.81 with an average of 0.61. Item discrimination indices obtained by using the upper 27% and lower 27% of the sample group showed that 32 of 40 items were above 0.50 with an average of 0.63. Each of these indices fell well within the acceptable range for a reliable test. After the item analysis conducted with the data obtained from the pilot application, one scenario (and eight questions related to this scenario) was taken out of the test.

The revised version of the FHIVS test includes four experimental scenarios; each of which consists of two parts. Each experimental scenario features a single paragraph describing an experiment accompanied by supporting diagrams, and four scenario-based multiple-choice questions about the experiment described. These experimental scenarios are given in [Figure 4](#). The first question related to the experiment was about the manipulated variable, the second was about the dependent variable, the third involved the control variables, and the fourth was about the hypothesis tested in the experiment. The same dependent variable was involved in the first and second parts of all the experimental scenarios given but in each part the effect of a different manipulated variable on the dependent variable was involved while all other variables were kept constant. Consequently, different hypotheses were tested in the first and second parts of the experiments. Additionally, the distractors in the answers to the questions in the first and second parts of the experiments were also identical. One example scenario and eight questions related to this scenario are given in the appendix. In the analysis of students' answers, responses to questions in the first and second parts of experiments were coded as A and B respectively. Items coded A and B were then compared to determine differences in the students' skills in identifying variables and formulating hypotheses in the two parts of the experiment. The contexts of experimental scenarios given in [Figure 4](#) can be summarized as follows:



**Figure 4.** Experimental scenarios used in the FHIVS test

In Scenario 1, two stages of an experiment about the discharge of water from a glass with a hole under it were described. In the first stage of the experiment (Part A), while variables like the size of the hole, the type of liquid and the shape of the container were fixed, the amount of liquid amount was changed and the discharge time was measured. In the second stage of the experiment (part B), while variables like the type of liquid, the amount of liquid and the stage of the container remained the same, the size of the hole changed and the discharge time was measured.

In Scenario 2, two stages of an experiment about boiling water in metal containers were described. In the first stage of the experiment (Part A), while variables like the amount water amount, the amount of heat given to the container and the size of the container were fixed, the metal which the container was made of changed and the boiling times were measured. In the second stage of the experiment (Part B), while variables like the amount of heat given to the container, the size of the container and the metal which it was made of remained the same, the amount of water changed and the boiling times were recorded.

In Scenario 3, two stages of an experiment about a simple pendulum were described. In the first stage of the experiment (Part A), while variables like angle of amplitude, mass of the oscillated object and volume of the object remained the same, the length of the rope and length of oscillation time were measured. In the second stage of the experiment (Part B), while variables like angle of amplitude, length of the rope and volume of the oscillated object remained the same, the mass was changed and the oscillation times were measured.

In Scenario 4, two stages of an experiment about a simple electric circuit were described. In the first stage of the experiment (Part A), while variables like the number of batteries in the circuit, the type of the material the conductive wire is made from and the width of the wire remained the same, the length of the wire changed and the intensity of the current going through the circuit was measured. In the second stage of the experiment (Part B), while variables like the number of batteries, type of the material the conductive wire was made from and length of the wire remained the same, the width of the wire changed and again the intensity of the current going through the circuit was measured.

As shown above, two different stages of an experiment were described in these four scenarios. The number of stages can be increased. In fact, at each stage the effects of a different independent variable on the same controlled variable are investigated and a different hypothesis is tested.

### **3. RESULTS/FINDINGS**







#### **3.1. Consistency of SPS**

To examine the consistency of the students' SPS performance, the responses of each student to the questions about the first and second stages, coded as A and B, were compared for accuracy. For this purpose, as shown in [Table 2](#), students' answers were categorized into four groups with different levels of performance consistency.

To describe each group given in [Table 2](#), students' answers to code A and B questions were compared separately for each skill. This comparison was done for all four groups, and the number of students in the groups and percentages in each group were found. The average number of students grouped in terms of skills is given in [Table 3](#). It was found that students falling into Groups 1 and 2 exhibited consistent performances whereas students in Groups 3 and 4 exhibited inconsistent performances.

According to the results presented in [Table 3](#), only about half the students were able to answer both code A and code B items correctly. In questions assessing the skill of identifying controlled variables, this number even dropped to 35%. The percentages of students who answered both code A and code B items incorrectly ranged between 15% and 35%. The percentages of students exhibiting an inconsistent performance by incorrectly answering any one of the code A or B items ranged between 20% and 25%. The skill with the highest level of inconsistent performance was formulating hypotheses (25%). The percentage of students exhibiting consistent performance (Group 1 + Group 2) was in the range 65% - 75%. The skill with the highest level of consistent performance was identifying the dependent variable (75%). All these descriptive statistics demonstrate that some (nearly a fifth) of the students exhibited different performances in the FHIVS test with regard to the two different parts of an experiment.

**Table 2.** Identification of groups

Groups		Group description	Performance consistency
Group 1		Students correctly answered both questions.	Consistent
Group 2		Students incorrectly answered questions.	Consistent
Group 3		Students answered code A questions correctly but code B questions incorrectly.	Inconsistent
Group 4		Students answered code A questions incorrectly but code B questions correctly.	Inconsistent
Other		Students left at least one question unanswered in the same experiment.	Undetermined

**Table 3.** Average Numbers of Students in Groups According to Skills (N = 370)

Groups	Identifying Manipulated Variable		Identifying Responding Variable		Identifying Controlled Variables		Formulating Hypotheses	
	N	%	N	%	N	%	N	%
Group 1	191	51.62	201	54.19	130	35	186	50.14
Group 2	73	19.73	76	20.41	130	35.2	56	15
Group 3	36	9.8	35	9.32	40	10.88	56	15.2
Group 4	41	11.15	31	8.38	31	8.45	36	9.73
Consistent Performance	264	71.35	276	74.6	260	70.2	241	65.14
Inconsistent Performance	78	20.95	66	17.7	72	19.33	92	24.93
Other	29	7.7	29	7.7	39	10.47	37	9.93
Total	370	100	370	100	370	100	370	100

### 3.2. Comparison of SPS achievement in different parts of the same experimental scenario

Would the test scores of students be affected when the variables in two different parts of an experiment testing different hypotheses changed roles? To address this question, the test scores for both code A and code B items were compared. For this purpose, a paired samples t-test was conducted for each skill. The results of the paired samples t-test are given in [Table 4](#).

**Table 4.** Paired Samples t-test Results

Skills	$\bar{X}_A$	$\bar{X}_B$	$S_A$	$S_B$	$t$	$p$
Identifying Manipulated Variable	2.49	2.58	1.47	1.38	-1,99	0.046
Identifying Dependent variable	2.60	2.51	1.46	1.42	1,99	0.046
Identifying Controlled Variables	1.91	1.76	1.63	1.50	2.96	0.003
Formulating Hypotheses	2.72	2.41	1.31	1.24	6.65	0.000



According to the data in Table 4, there were statistically significant differences in the total scores for the code A and code B items of the test for all the specific skills. These data demonstrate that some students were affected by the variables having different roles in different parts of the experiment. Most differences between the code A and code B questions were found with regard to the skill of formulating hypotheses. When the eta-squared values ( $\eta^2$  of the manipulated variable=0.01,  $\eta^2$  of the dependent variable=0.01,  $\eta^2$  of the controlled variables=0.02,  $\eta^2$  of formulating hypotheses=0.11) were computed separately for the skills taken into consideration, it could be stated that the two-stage nature of the experiments had a small effect on students' performance scores in terms of identifying variables and a moderate effect on their performance scores for formulating hypotheses.

### 3.3. Stability of answers in different parts of the same experimental scenario

In a new situation where a different hypothesis is tested, did the students understand the changing role of the variable? To address this question, same responses from each student in both parts of the experiments were compared with one another. The number of students choosing the same response for both code A and code B items for all the experiments and skills were identified. The data obtained are presented in Table 5. The data in Table 5 show that nearly 64% of the students marked the same response in both parts of the experiment while identifying the dependent variable. This can be interpreted as positive since the same dependent variable had been worked on in both parts of all experiments. However, on the other hand, in identifying the manipulated variable 18% of the students marked the same response for the two parts; in identifying the controlled variable 28% of the students marked the same response for the two parts; in formulating hypotheses 14% of the students marked the same response for the two parts. These results are interesting since they demonstrate that some students did not take into consideration the different parts of the experiment while identifying the variables and testing the hypotheses.

**Table 5.** Percentage of the students who gave the same response for both parts of the experimental scenarios

Skills	Identifying Manipulated Variable		Identifying Dependent variable		Identifying Controlled Variables		Formulating Hypotheses	
	N	%	N	%	N	%	N	%
Scenario1	58	15.68	233	62.97	53	14.32	33	8.92
Scenario2	59	15.95	241	65.14	126	34.05	37	10.00
Scenario3	60	16.22	226	61.08	103	27.84	55	14.86
Scenario4	89	24.05	250	67.57	132	35.68	85	22.97
Overall	66.50	17.97	237.50	64.19	103.50	27.97	52.50	14.19

The data collected in the research show that the scores of nearly 22% of the students for formulating hypotheses and identifying variables changed depending on the part of the experiment. In other words, some students' performance changed depending on different parts of the same test. Furthermore, it has been established that a significant portion of the students ignored different parts of the experiment while identifying the variables or hypotheses tested in the experiment. In the second part of the experiment where the hypothesis was tested, these students did not mind putting the same answer they had done in the first part. For example, in the questions given in the Appendix, the effect of the "height of liquid in a glass" variable on the "emptying time" variable was examined in the first part of the experiment. In the second part, the effect of "hole size" variable on the "emptying time" variable was examined. Some students mistakenly selected the "height of liquid in glass" variable as the manipulated variable

in the first and second parts of the experiment. If these questions assessing the skill of identifying variables had only been developed for single stage experiments, this confusion would not have been revealed.

#### 4. DISCUSSION and CONCLUSION

SPS are intellectual and physical skills we use to acquire information, think about problems and formulate conclusions. These skills are an inseparable component of inquiry-based science education. Learning with understanding in science involves using SPS. Thus, the development of SPS is a major goal of science education. Several science education curricula have been developed with the intention of teaching the acquisition of SPS, and measuring and assessing these skills is an important aspect of science education. Over recent years many tools have been developed in various forms with the objective of measuring these important skills (Harlen, 1999; Aydınli et al., 2011).

The measurement of SPS comes with various difficulties. These difficulties may be discussed from two aspects. The first concerns how SPS should be measured; in other words, it is about the types of question to be used in SPS measurement. Some researchers think that the best way to measure the SPS of students is by using laboratory reports, oral presentations and observations (Lavinghousez, 1973; Gabel, 1992; Ostlund, 1992; Haury & Rillero, 1994; Kazeni, 2005). A more appropriate way of measuring SPS is the use of hands-on activities, but due to their ease of application, simplicity of evaluation, and because they do not require expensive resources, paper and pencil tests are still often currently preferred. According to Rezba et al. (1995), a transition from multiple-choice measurement methods to multi-formatted measurement methods is taking place. However, multiple-choice tests are still frequently preferred because they can be easily applied to large groups of students. According to Burns et al. (1985), assessing students' skills through observation in laboratory situations can be difficult and time-consuming. While an instructor may obtain an intuitive feel for a student's competence in process skills via observation, high-quality tests are needed to achieve accurate measures of students' performance.

The second aspects concern the difficulties in selecting content and contexts when measuring and assessing SPS. Harlen (1999) asserts that SPS have to be used in concert with specific content. Therein lies the difficulty in assessing these skills. Students' performance in any task involving these skills will be influenced by the nature of the content as well as by the students' ability. In the literature is examined various studies have demonstrated that the content of the tasks utilized in SPS measurement tools have an influence on students' performance. Zimmerman and Glaser (2001) conducted a study on this. They investigated whether sixth-grade students were affected by variations in the scenario given while designing an experiment about plants. It was found that student performances were affected when the scenarios were chosen from among topics in the curriculum. These studies also demonstrate that the performance of SPS is affected by whether the content of tests relates to everyday life or to scientific issues. While a question referring school or a laboratory context can point toward a specific idea, a subject from everyday life might not produce a similar association. According to these studies, students demonstrated better SPS when the content was drawn from everyday life, while their application skills were better in scientific contexts (Song & Black, 1991, 1992; Temiz, 2010). In this study, these effects were also taken into consideration when the scenarios were created. Some of the scenarios were created using content from everyday life (scenarios 1 and 2) and some were formulated using scientific contexts (scenarios 3 and 4). The findings obtained in this study add a new dimension to the discussion on content and context selection in SPS measurement. This dimension is the development of multi-stage scenarios.

In this study, two different stages of an experiment used in experimental scenarios were

explained. At each stage, the effects of a different independent variable on the same controlled variable were investigated. In other words, at each stage a different hypothesis was tested. The method of testing a variable's effect on another effect is called "fair testing". According to Hughes and Wade (1993) children have difficulty in controlling variables and see no problem in simultaneously exchanging two or more variables even up to the ages of 13-15. For this reason, the development of the concept of fair testing should commence early in schools. According to Carin and Bass (2001), in controlled variable studies conducted among primary and middle schools, students better understand the experiment when they learn about the fair testing technique. In addition to this, teaching the students that "variables can exchange roles in various parts of an experiment" is a finding which this study contributes to the literature.

Test writers have focused on content validity, reliability, difficulty level and discrimination indices, all of which are important for the development of high-quality tests. Many of the SPS tests widely used in the literature have been developed to meet these requirements. However, due to the nature of SPS, if multiple-choice questions are to be used, the scenarios must be carefully formulated in the question stem. For example, when writing a question, the multi-stage experimental scenario needs to be considered. This study researched the effectiveness of the scenario-based multiple-choice tests widely used in SPS measurement. In multiple-choice SPS tests, item writers generally require the student to determine what hypothesis is being tested in an experiment and to identify the variables in a single-stage experiment. But in science, experiments can have several stages, and a different hypothesis can be tested in each part. Therefore, a manipulated variable in the first part of an experiment can become the controlled variable in the second part. The data collected in this study have demonstrated that multi-stage experiments are effective in ascertaining students' SPS competence. The findings of this research show that students exhibited differing performance in FHIVS questions with regard to differing parts of the same experiment. This variation originates from students' miscomprehension of the reality that variables may play different roles in different parts of an experiment. Nearly one fifth of the students failed to notice that a manipulated variable in the first part of an experiment was the controlled variable in the next part of the experiment. This situation affected their scores for identifying variables in addition to formulating hypotheses.

The results obtained in this study should be considered when assessing the skills of identifying variables and formulating hypotheses, skills which are among the most important SPS. If the students' performance in these areas is to be measured using multiple-choice test items, multi-stage experimental situations where a different hypothesis is tested at each stage should be used instead of single-stage experiments. The ways in which variables can change should be taken into consideration while selecting content to measure SPS. If a student chooses the right answer in a multiple-choice test, this is still not enough to conclude that student's knowledge of the subject is complete and accurate. In addition, a student may choose a distractor as the correct answer due to lack of information and mistakes made during the test. In addition to these factors, not being able to comprehend the changing role of the variables may cause the emergence of Groups 3 and 4 above. If the two-stage scenarios had not been used, this situation would not have been observed. This could have misled the researcher and the researcher may have believed that the student's SPS were more developed (or not as developed) as they were. Some researchers suggest using multiple stages in multiple-choice tests in order to determine misconceptions (Bahar, 2001; Karataş et al., 2003; Aykutlu & Şen, 2012). A similar approach should be followed for measuring SPS. For a student to be considered successful at a skill, she or he should be able to correctly answer the two parts of a related scenario, like the students in Group 1 above.


The advantages of using dual-stage questions while measuring the SPS can be summarized as follows: In reality, scientific experiments consist of multiple stages. Therefore, to use multi-

stage experimental scenarios to measure SPS is more realistic. While a variable can be an “independent variable” at a certain stage of the experiment, the same variable can also be a “controlled variable” at another stage of the experiment. The idea that a variable can play a different role at different stages of the experiment is a part of the “fair testing” strategy. For this reason, while measuring the skills of manipulating variables and formulating hypotheses, using multi-staged scenarios will give more sound results. Data collected from single-stage multiple-choice tests can be misleading. To make more consistent assessments, it is thus better to use multi-stage items.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### ORCID

Burak Kağan TEMİZ  <https://orcid.org/0000-0001-8636-8743>

## 5. REFERENCES

- Abruscato, J. (2000). *Teaching children science*, Needham Heights, M.A: Allyn and Bacon.
- Açıköz ÜN, K. (2003). *Aktif öğrenme*, İzmir: Eğitim Dünyası Yayınları.
- Temiz, B. K., Taşar, M. F., & Tan, M. (2006) Development and validation of a multiple format test of science process skills. *International Education Journal*, 7(7), 1007-1027.
- Temiz, B. K. (2007). *Fizik öğretiminde öğrencilerin bilimsel süreç becerilerinin ölçülmesi [Evaluating students' science process skills at physics teaching]* (Unpublished dissertation). Gazi University, Ankara, Turkey.
- Temiz, B. K. (2010). Bilimsel Süreç Becerilerini Ölçmede İçerik Seçiminin Önemi. *e-Journal of New World Sciences Academy*, 5(2), 614-628.
- Aykutlu, I., & Şen, A. İ. (2012). Determination of secondary school students' misconceptions about the electric current using a three tier test, concept maps and analogies. *Education and Science*, 37(166), 275-288.
- Aydınlı, E., Dökme, I., Ünlü, Z. K., Öztürk, N., Demir, R., & Benli, E. (2011). Turkish elementary school students' performance on integrated science process skills. *Procedia-Social and Behavioral Sciences*, 15, 3469-3475.
- Aydoğdu B., Tatar N., Yıldız E., & Buldur S. (2012). The science process skills scale development for elementary school students. *Journal of Theoretical Educational Science*, 5(3), 292-311.
- Aydoğdu, B., & Karakuş, F. (2017). Basic Process Skills Scale of towards Pre-School Students: A Scale Development Study, *Journal of Theoretical Educational Science*, 10(1), 49-72.
- Bahar, M. (2001). Çoktan Seçmeli Testlere Eleştirel Bir Yaklaşım ve Alternatif Metotlar. *Kuram ve Uygulamada Eğitim Bilimleri*, 1(1), 23-28.
- Beaumont-Walters, Y., & Soyibo, K. (2001). An analysis of high school students' performance on five integrated science process skills. *Research in Science and Technological Education*, 19(2). 133-145.
- Berger, C. F. (1982). Attainment of skill in using science processes. I. Instrumentation. methodology and analysis. *Journal of Research in Science Teaching*, 19(3), 249-260.
- Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching*, 22(2), 169-177.
- Büyüköztürk Ş, Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2009). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.
- Carin, A. A. (1993). *Teaching science through discovery*. Toronto: Macmillan Publishing Company.

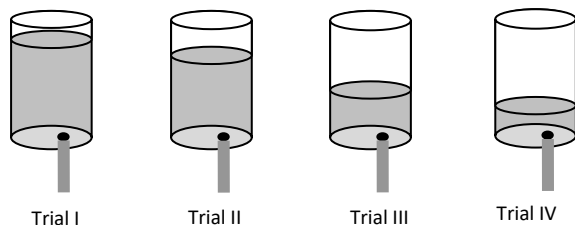
- Carin, A. A., & Bass, J. E. (2001). *Teaching science as inquiry*. Upper Saddle River, New Jersey: Merrill Prentice Hall.
- Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary students. *Science Education*, 64, 601-608.
- Erduran Avcı, D., & Bayrak, E. B. (2013). Investigating Teacher Candidates' Opinions Related to Scenario-Based Learning: An Action Research. *Ilkogretim Online*, 12(2), 528-549.
- Esler, M. K., & Esler, W. K. (2001). *Teaching elementary science. A full-spectrum science instruction approach*. Belmont CA.: Wadsworth Publishing.
- Feyzioglu, B., Demirdag, B., Akyildiz, M., & Altun, E. (2012). Developing a science process skills test for secondary students: validity and reliability study. *Educational sciences: Theory and Practice*, 12(3), 1899-1906.
- Gabel, D. L. (1993). *Introductory science skills*, Illinois: Waveland Press, Inc.
- Harlen, W. (1993). *Teaching and learning primary science*. London: Corwin Press.
- Harlen, W. (1999). Purposes and Procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice*, 6(1), 129-146.
- Haury, D. L., & Rillero, P. (1994). *Perspectives of hands-on science teaching*. Retrieved from <http://www.ncrel.org/sdrs/areas/issues/content/contareas/science/eric/eric-toc.htm>
- Hughes, C., & Wade, W. (1993). *Inspirations for investigations in science*. Warwickshire: Scholastic Publication.
- Karasar, N. (2000). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayın Dağıtım.
- Karataş, F.Ö., Köse, S., & Coştu, B. (2003). Öğrencilerin Yanılgılarını ve Anlama Düzeylerini Belirlemede Kullanılan İki Aşamalı Testler. *Pamukkale Üniversitesi, Eğitim Fakültesi Dergisi*, 13(1), 54-69.
- Kazeni, M. M. M. (2005). *Development and validation of a test of integrated science process skills for the further education and training learners* (Unpublished master's dissertation). University of Pretoria, South Africa.
- Kumar, D. (1999). *Computers and Assessment in Science Education*. Retrieved from ERIC database. (ED395770).
- Lavinghousez, W. E. Jr. (1973, February). *The analysis of the biology readiness scale (BRS), as a measure of inquiry skills required*. Paper presented at BSCS Biology, College of Education, University of Central Florida.
- Ludeman, R. R. (1975). *Development of the Science Processes Test (TSPT)* (Unpublished doctoral dissertation). Michigan State University. MI.
- Lunetta, V. N., Hofstein, A., & Giddings, G. (1981). Evaluating science laboratory skills. *The Science Teacher*, 48, 22-25.
- Martin, D. J. (2002). *Elementary Science Methods a Constructivist Approach*. New York: Delmar Publishers.
- Molitor, L. L., & George, K. D. (1976). Development of a test of science process skills. *Journal of Research in Science Teaching*, 13(5), 405-412.
- Ostlund, K. L. (1992). *Science process skills: assessing hands on student performance*. California: Addison Wesley.
- Padilla, M. (1990). The science process skills. *Research Matters-to the Science Teacher*. No. 9004. Retrieved from <http://www.educ.sfu.ca/narstsite/publications/research/skill.htm>
- Rezba, R. J., Sprague, C. S., Fiel, R. L., Funk, H. J., Okey, J. R., & Jaus, H. H. (1995) *Learning and assessing science process skills*. Iowa: Kendall/Hunt Publishing Company.
- Riley, J. W. (1972). *The Development and Use of a Group Process Test for Selected Processes of the science Curriculum Improvement Study* (Unpublished Doctoral dissertation). Michigan State University. MI.
- Shahali E. H. M., & Halim L., (2010), Development and validation of a test of integrated science process skills, *Procedia Social and Behavioral Sciences*, 9, 142-146.

- 
- Smith, K. A., & Welliver, P. W. (1995). *Science process assessments for elementary and middle school students*. Smith and Welliver Educational Services. Retrieved from <http://www.scienceprocesstests.com>
- Solano-Flores, G. (2000). Teaching and assessing science process skills in physics: The "bubbles" task. *Science Activities*, 37(1), 31-37.
- Song, J., & Black, P. J. (1991) The effects of task contexts on pupil's performance in science process skills. *International Journal in Science Education*, 13, 49-58.
- Song, J., & Black, P. J. (1992) The effects of concept requirements of task contexts on pupil's performance in control variables. *International Journal in Science Education*, 14, 83-93.
- Tannenbaum, R. S. (1971). Development of the test of science processes. *Journal of Research in Science Teaching*, 8(2), 123-136.
- Thalheimer, W. (2013). The Power of Scenario-Based Questions. Retrieved from <http://www.immersivelearninguniversity.com/articlethalheimersep13>
- Tobin, K. G., & Capie, W. (1982). Development and validation of a group test of integrated science processes. *Journal of Research in Science Teaching*, 19(2), 133-141.
- Tosun, C. (2019). Scientific process skills test development within the topic "Matter and its Nature" and the predictive effect of different variables on 7th and 8th grade students' scientific process skill levels. *Chemical Education Research Practice*, 20, 160-174.
- Zimmerman, C., & Glaser, R. (2001). *Testing positive versus negative claims: A preliminary investigation of the role of cover story in the assessment of experimental design skills* (Tech. Rep. No. 554). Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

## Appendix: Examples of the Items from FHIVS Test

### Scenario-A

Susan has conducted an experiment which is shown below with a glass with a hole under it. Answer the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> questions that follow.



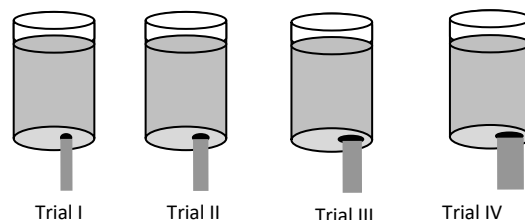
Susan, in her first attempt put liquid into the glass to a height of 15 cm and measured the time as 15 seconds for the glass to be completely emptied. In her second attempt, she put the same liquid into the same glass but this time to a height of 10 cm and measured the time for to empty the glass as 10 seconds. In her third attempt she put same liquid into the same glass to a height of 6 cm and measured the time to empty the glass as 7 seconds. In her fourth and last attempt she put the same liquid into the same glass to a height of 4 cm and measured the time to empty the glass as 5 seconds.

- What is the manipulated variable in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.
- What is the dependent variable in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.
- What is/are the controlled variable(s) in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.

a. i    b. i and ii    c. ii, iv and v    d. iii, iv and v  
e. ii and iii
- What is the hypothesis that was tested in this research?
  - If the size of the hole in the bottom of the glass decreases, then the intensity of the liquid will decrease.
  - If the height of the liquid in the glass increases, then the emptying time of the liquid will increase.
  - If the number of the holes' increases, then the emptying time of the liquid will decrease.
  - If the intensity of the liquid in the glass increases, then the emptying time of the liquid will increase too.
  - If the size of the hole in the bottom of the glass increases, then the emptying time of the liquid will increase too.

### Scenario-B

Susan has conducted the new experiment below, with four similar size glasses with different size holes in the bottom. Answer the 5<sup>th</sup>, 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> questions that follow.



Susan, in her first try put liquid into the glass with 15 cm height and 2 mm hole scale and measured the time as 15 seconds for glass's getting emptied completely. In her second try, she put the same liquid into the same glass but this time with 15 cm height and 3 mm hole scale and measured the emptying time as 10 seconds. In her third try she put same liquid into the same glass with 15 cm height and 4 mm hole scale and measured emptying time as 7 seconds and in her fourth and last try she put same liquid into the same glass with 15 cm height and 5 mm hole scale and measured the emptying time as 7 seconds.

- What is the manipulated variable in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.
- What is the dependent variable in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.
- What is/are the controlled variable(s) in this research?
  - Height of the liquid in the glass.
  - Liquid's emptying time.
  - Number of holes in the bottom of the glass.
  - The size of the hole in the bottom of the glass.
  - The type of the liquid in the glass.

a. i    b. i and ii    c. ii, iv and v    d. iii, iv and v  
e. ii and iii
- What is the hypothesis that was tested in this research?
  - If the size of the hole in the bottom of the glass decreases, then the intensity of the liquid will decrease.
  - If the height of the liquid in the glass increases, then the emptying time of the liquid will increase.
  - If the number of the holes' increases, then the emptying time of the liquid will decrease.
  - If the intensity of the liquid in the glass increases, then the emptying time of the liquid will increase too.
  - If the size of the hole in the bottom of the glass increases, then the emptying time of the liquid will increase too.

## Use of Item Response Theory to Validate Cyberbullying Sensibility Scale for University Students

Osman Tolga Arıçak <sup>1</sup>, Akif Avcu <sup>2,\*</sup>, Feyza Topçu <sup>1</sup>, Merve Gülçin Tutlu <sup>1</sup>

<sup>1</sup>Department of Psychology, Hasan Kalyoncu University, Gaziantep

<sup>2</sup>Department of Educational Sciences, Marmara University, Istanbul

### ARTICLE HISTORY

Received: 04 October 2019

Revised: 03 December 2019

Accepted: 06 February 2020

### KEYWORDS

Cyberbullying sensibility,  
Test validation,  
Item response theory,  
Graded response model,  
Item selection

**Abstract:** A thirteen-item cyberbullying sensibility scale (CSS), developed by Tanrikulu, Kinay, and Arıçak (2013) and extensively used by researchers, was used to measure the cyberbullying sensibility levels of high school students. Unlike other similar concepts, such as cyberbullying and cyber victimization, there are no scales developed to measure the cyberbullying sensibility among university students. In this study, the data obtained from 727 university students were analyzed based on item response theory (IRT) techniques, and psychometric evidences were obtained to evaluate whether it is appropriate to use the scale on the university students. Accordingly, a parameterization of CSS items was performed by using the graded response model. Using the discrimination parameters and item fit statistics, some items were removed from the original scale and a seven-item CSS version was developed since preliminary exploratory and confirmatory factor analyses provide inadequate evidence for the validity of a one-dimensional structure of cyberbullying sensibility. However, an IRT-based item removal process yielded an acceptable improvement. In this way, despite the six items being removed from the original CSS form, the scale retained 64% of the information it provided. The reliability values computed based on the classical approach and IRT were above .8 after the item elimination process with only a minor drop. With the validation process, the CSS will be a valuable measurement tool to determine the level of cyberbullying sensibility among university students and allow academicians to conduct research with this population.

## 1. INTRODUCTION

Today, young people who use information technologies are under extreme risk of cyberbullying in cyberspace's unknown and virtual social relations (Willoughby, 2018). They may be engaging in bullying, be exposed to bullying, or be bystanders (Gahagan, Vaterlaus, & Frost, 2016). A noteworthy concept in prevention of bullying in cyberspace is the cyberbullying sensibility. Studies show that young people not only intentionally hurt others, but also can bully others just for fun (Arıçak, 2015; Tolia, 2016). This is an important finding that emphasizes the lack of knowledge and awareness of the consequences of such actions. The cyber-bullying sensibility is defined as the awareness of young people about cyberbullying behaviors while

CONTACT: Akif Avcu ✉ [avcuakif@gmail.com](mailto:avcuakif@gmail.com) 📧 Marmara University, Atatürk Education Faculty, İstanbul, Turkey

ISSN-e: 2148-7456 /© IJATE 2020



using electronic media and how sensitive they are toward these kinds of behaviors (Tanrikulu, Kinay, & Arıcak, 2015). Studies aimed at preventing cyberbullying (Gaffney, Farrington, Espelage, & Ttofi, 2018) and increasing the sensibility to cyberbullying (Nedim Bal & Kahraman, 2015; Tanrikulu, Kinay & Arıcak, 2015) have begun to take their place in the literature. This trend in the literature requires the use of instruments that measure the sensibility to cyberbullying behaviors. For this aim “Cyberbullying Sensibility Scale” (CSS), developed by Tanrikulu, Kinay and Arıcak in 2013 to measure the cyberbullying sensibility of adolescents, has been used in various studies in the last six years (i.e., Aktan & Çakmak, 2015; Baştaç & Altınova, 2015; Doğan, Cansu, & Şahin, 2016).

IRT (item response theory) is an important psychometric approach used in the processes to obtain valid measurement instruments. Its use has become widespread among test developers since this method can solve many measurement difficulties encountered during test development process and provide richer output (Samejima, 1968; Embretson, 1996). For this reason, the validation of the CSS for university students was carried out by taking advantage of IRT.

The most salient difference between the Classical Test Theory (CTT) and IRT is that the CTT assumes equal measurement accuracy across all test takers, regardless of their ability levels. However, in IRT, the measurement accuracy depends on the level of the latent trait being measured. This leads to a differentiation between results obtained from CTT and IRT. When model-data fit is achieved, IRT provides the test information functions (TIF) (the amount of the information the test provides to the users) and the amount of error for different ability levels (Hambleton et al., 2000).

As stated by Hambleton et al. (1991), IRT models have two basic assumptions provided that the measured property is one-dimensional: unidimensionality and local dependence. The first assumption of unidimensionality means that only one trait is being measured by a set of items composing the test. It requires the presence of a dominant factor explaining most of the variability on test scores. In other words, the covariance between items can be explained by the single dimension. Hattie (1985) recommended that the unidimensionality could be tested with investigating eigenvalues and variability explained by the first factor based on exploratory factor analysis (EFA) testing one dimensional factor analysis via confirmatory factor analysis (CFA). Another IRT assumption is local independence. According to this assumption, responses to an item should not be statistically related to each other, even after the latent trait being measured is kept statistically constant.

Another IRT assumption is local independence. According to this assumption, responses to an item should not be statistically related to each other, after the latent trait being measured is kept statistically constant. It implies that, an observed responses must not be affected by any unrelated factors other than the ability levels of participants. Different statistics developed so far to investigate whether or not local independence assumption holds. Most commonly preferred statistics are  $\chi^2$  statistics, G2 statistics (Chen & Thissen, 1997) and Q3 statistics (Yen, 1984). For the current study, only Q3 statistics were taken into account in order to determine whether Local dependence (LD) was present or not. Even though there is no consensus on the cut off value of Q3 statistics, the value of 0.3 were generally considered as an evidence for the existence of LD.

Many different models have been developed to analyze Likert-type items with more than two response options also known as polytomous items. Although these polytomous response models differ among themselves in terms of parametrization, they all include the specification of a location and slope parameter (and the characteristic curve accordingly) for each response category (Thissen & Steinberg, 1986). The Graded Response Model (GRM) (Samejima, 1968) is a polytomous IRT model developed for item responses characterized by graded categories.

The GRM and is considered to be the generalization of two parameter logistics models (Keller, 2005). The model is particularly suitable for use with Likert type items and has been preferred in different studies (i.e. Rubio et al., 2007; Mielenz et al., 2010). Even though there are some other alternative polytomous models available in the literature (see Hambleton & Swaminathan, 1985) GRM was preferred for the current study.

The item level fit statistics could also be obtained by utilizing IRT based approach and was evaluated with polytomous extension of S-X<sup>2</sup> item-fit index (Orlando & Thissen, 2000). This index is Chi-square based and uses a significance test. Generally, *p* values lower than .05 were considered a poor item fit.

Another important advantage of the IRT is the provision of item and test information. In IRT terminology, the term information implies the amount of accuracy of the measurement and closely related to reliability. For a two-parameter model, the item information is determined as the function of the item discrimination and the item location parameters in each value of the ability parameter. The item information function shows the contribution of each item to the measurement of the latent trait being measured. Items with more discrimination power contribute more to the accuracy of measurement (Hambleton et al., 2000).

Given these advantages cited above, the use of IRT in test development/revision processes provides advantages that cannot be achieved with classical test theory. Accordingly, the number of studies using IRT for test development, test revision and obtaining shorter versions of available ones increased in the last decade (i.e. Zanon, Hutz, Yoo & Hambleton, 2016; Istiyono et al., 2019; Bilker et al., 2012). In the light of this fact, the revision of CSS using IRT based approach is the main purpose of this study.

## **2. METHOD**

### **2.1. Study Group**

The participants in this study were 727 university students. The average age of students was 22.03 (SD=22.03, ranged between 18-26 years), and the majority, 462, were female (63.4%) and the rest, 266, were male (36.6%). The participants were selected with convenient sampling among the students studying in various faculties of a private university. Since IRT studies are model-based, large sample size is of great importance for the accuracy of the measurement. For models with more parameters estimated, Tsutakawa and Johnson (1990) stated that a sample size of 500 would be sufficient. For this reason, participation in the research was on a voluntary basis, and the sample size was intentionally kept high, above the size recommended by Tsutakawa and Johnson (1990).

### **2.2. Measurement Instrument**

The Cyberbullying Sensibility Scale (CSS) was developed by Tanrıku, Kınay & Arıcak in 2013. The scale development process was conducted in Istanbul metropolitan area with 663 high school students. For construct and construct validity, both the EFA and the CFA were carried out. The results of the EFA showed that the scale has a one-dimensional structure, explaining 47% of the total variance with factor loadings varying between .61 and .76. The CFA results further confirmed the one-dimensional structure of the scale ( $\chi^2/df=3.22$ , RMSEA=.082). The loadings coefficients varied between .31 and .65. The Cronbach alpha coefficient was estimated as .90 and the test-retest reliability as .63.

### **2.3. Procedures**

The participants were asked to answer the items of CSS and a demographic information form in their own classes. The students were informed that participation in the research was voluntary and that any information they provided would be kept confidential. The students were asked to read the questions carefully and fill in the scale to reflect their views. Data collection was

performed in a single session for each class.

## 2.4. Analysis

To test the unidimensionality of the data, a confirmatory factor analysis was performed by utilizing MPLUS 6 (Muthén & Muthén, 1998-2012). The fit of the model was evaluated  $\chi^2$  statistic, Root Mean Square Approximation (RMSEA), standardized root mean squared residual (SRMR) and confirmatory fit index (CFI) and Tucker-Lewis index (TLI) indexes were used. In addition, unidimensionality was also evaluated by applying an exploratory factor analysis. For this analysis, SPSS 21 was utilized. In addition, local independence assumption was evaluated by Q3 statistics (Yen, 1984). After checking the assumptions. The GRM (Samejima, 1968) was used for IRT based item calibration. After obtaining item parameters, item fit statistics were computed with  $S-X^2$  index (Orlando & Thissen, 2000). IRT based factor analysis was performed with a *mirt* package (Chalmers, 2012) in R program (R core team, 2017). In addition, the item level fit statistic and information functions for the test and items were computed with the same package.

## 3. RESULT / FINDINGS

### 3.1. Testing IRT assumptions

The confirmatory factor analysis (CFA) results showed that a one-dimensional model was not confirmed at acceptable level:  $\chi^2$  (N=727, df=65) = 492.62,  $p < 0.001$ , CFI = 0.860 TLI = 0.832, RMSEA = .95, %95C.I. = [0.087-0.102], and SRMR = 0.053. Factor loadings varied between 0.44 and 0.70 at  $p < 0.001$  significance level. In addition, an exploratory factor analysis (EFA) was conducted as a supplemental. The results showed that two factors were extracted with eigenvalues greater than one. The first factor explained 39.95% of total variance while the second factor explained 8.91%. Even though the CFA results did not confirm the one factor structure of the CSS, the EFA results provided evidence that the original unidimensional structure of the CSS was present because, as stated earlier, Hattie (1985) stated that 20% or more variability on the first factor is a proof for the presence of a dominant factor explaining most of the instrument scores. At this point, IRT based analyses were performed to obtain more acceptable results to keep original one-dimensional structure of the CSS for university students. We achieved this by investigating the information contribution of each item and item level fit statistics.

The LD assumption was tested by computing the Q3 statistics. The results showed that only one item pair had a Q3 value greater than 0.3 (item 1 - item 2). At this point, one of the items had to be eliminated from the scale. We eliminated item 1 because it was not only locally dependent with item 2 and it didn't satisfy the criterion related to the amount of information it had (see below). In addition, the LD was tested with a final 7-item version of the scale, and no item pairs were found with LD.

### 3.2. Fitting the Graded Response Model

As previously mentioned, the CSS has three graded response options. Hence, the GRM was preferred for model fitting. As a result of the GRM estimations, one discrimination and two threshold parameters were obtained. The analysis showed that the fit values of the single-factor GRM model were at an acceptable level: CFI = 0.942 TLI = 0.954, RMSEA = .074, and SRMR = 0.067.

Table 1 lists item parameters, information results, and  $S-X^2$  statistics to evaluate item fit. The factor loadings for the CSS items ranged from .52 to .72. In addition, the communality values ranged from .33 to .63. The results provided evidence that the items are fairly different in terms of the amount of variation with a common factor. The discrimination parameters ranged from 1.02 to 2.44. The contribution of items to total information varied between 1.49 and 4.19, and

the percentage of the contribution to the total test information varied between 4.94% and 11.11%. This finding shows that the contribution of items to the model shows a significant variance. The difficulty parameters of the CSS vary between -0.62 and 4.02. This finding shows that the CSS is not useful enough for identifying individuals with low levels of cyberbullying sensibility but was a scale more suitable for identifying average-to-high-level individuals. The item fit was also evaluated by using  $S-X^2$  statistics. The results showed that three items (item 7, item 12, and item 13) had  $p$  values less than .05, which suggest that the items did not fit the model well.

### 3.3. Item Selection for the CSS for University Students

The items were selected from the 13 item CSS to increase the model fit index, contributing to the validation process of the scale for university students. Two different criteria were used in this process. First, the amount total item information was obtained by adding of the information values of each item. Later, total item value was divided by the number of items in CSS. In this way, average information value was obtained. The items with above-average contribution to total information were determined, and these items were kept in the new CSS form. This operation was carried out only once. Secondly, the items with a poor level of item fit statistics (items with  $p$  values corresponding to  $S-X^2$  statistics below .05) were also eliminated from the new form of the CSS. This process continued until no poor fit items remained. As seen in Table 1, five items below the average were removed from the scale. Three of these items also had a poor fit ( $p < .05$ ). Thus, the second criterion was not used to eliminate items at this stage. After removing these five items from the scale, the GRM model was repeated with the remaining eight items, the item fit statistics were examined, and each item with a poor fit level was determined and eliminated (item 1). When the GRM model was repeated with the remaining seven items, it was found that there were no items to be eliminated according to this criterion.

**Table 1.** Initial item loadings, communalities, four parameters, fit statistics, and information

	Factor analysis		Item Parameter			Item Fit			Test info=37.74 (M=2.90)	
	F	h2	a1	d1	d2	S-X <sup>2</sup>	df (S-X <sup>2</sup> )	p (S-X <sup>2</sup> )	Item info.	%
item 1	0.77	0.59	2.04	0.77	3.94	29.84	25	0.230	3.68	9.74
item2	0.79	0.62	2.19	0.95	4.07	17.45	25	0.865	3.91	10.35
item3	0.52	0.27	1.02	0.23	1.95	31.11	25	0.185	<b>1.49</b>	3.96
item4	0.73	0.53	1.82	0.32	2.87	39.14	29	0.099	3.02	8.00
item5	0.75	0.56	1.91	0.24	2.65	24.95	29	0.681	3.14	8.32
item6	0.73	0.53	1.80	0.47	3.07	32.79	29	0.286	3.05	8.08
item7	0.61	0.37	1.31	0.99	3.11	58.90	33	<b>0.004**</b>	<b>2.11</b>	5.60
item8	0.58	0.33	1.20	1.10	3.00	39.46	34	0.239	<b>1.87</b>	4.94
item9	0.82	0.67	2.44	0.07	2.95	30.58	26	0.244	4.19	11.11
item10	0.72	0.52	1.77	-0.49	2.21	33.77	28	0.209	2.96	7.84
item11	0.79	0.63	2.20	0.18	2.89	24.35	27	0.611	3.62	9.60
item12	0.65	0.43	1.46	-0.62	2.01	48.03	31	<b>0.026*</b>	<b>2.38</b>	6.29
item13	0.70	0.48	1.65	1.03	2.56	39.53	22	<b>0.012*</b>	<b>2.33</b>	6.17

Note: \* $p < 0.05$ ; \*\* $p < 0.01$ ; h2=communality; bolded results are pointing the items with below average information contribution and poorly fitting items at first stage investigation.

### 3.4. Fitting the Graded Response Model for Seven-item CSS Form

The results for the GRM fitted to the seven-item CSS form are provided at Table 2. Factor loadings ranged from .71 to .85, and communalities ranged from .50 to .72. When compared with the 13-item version, it was seen that the variance of items with common factor showed less variability. Item discrimination values of seven items varied between 1.71 to 2.74. Again,

items in the seven-item CSS form are more homogenous in terms of their discrimination powers. In addition, the results showed that the threshold parameters ranged from .52 to 3.71. In addition, for seven item version, S-X<sup>2</sup> statistics and their corresponding p values suggested that all items were fitted to the GRM model at an acceptable level ( $p > .05$ ). The seven-item version retained 64% of the total information provided by the original CSS, and the average contribution of each item to total test information increased to 18%.

**Table 2.** Final item loadings, communalities, parameters, fit statistics and information for 7-item CSS.

	Factor analysis		Item Parameter			Item Fit			Test info=24.00 (M=3.43)	
	F	h2	a1	d1	d2	S-X <sup>2</sup>	df(S-X <sup>2</sup> )	p(S-X <sup>2</sup> )	Value	%
item2	0.73	0.54	1.84	3.71	0.83	18.56	16	0.292	3.23	13.47
item4	0.71	0.50	1.71	2.80	0.29	18.34	17	0.367	2.89	12.04
item5	0.74	0.55	1.88	2.65	0.23	18.85	16	0.277	3.15	13.13
item6	0.72	0.52	1.76	3.06	0.45	18.72	16	0.284	3.01	12.54
item9	0.85	0.72	2.74	3.20	0.06	17.82	14	0.215	4.92	20.48
item10	0.75	0.56	1.91	2.30	-0.52	23.20	15	0.080	3.33	13.89
item11	0.76	0.59	2.02	2.76	1.50	10.06	16	0.863	3.47	14.44

Not: h2=communality

### 3.5. Comparison of the 13-item CSS Form and the Seven-item CSS Form

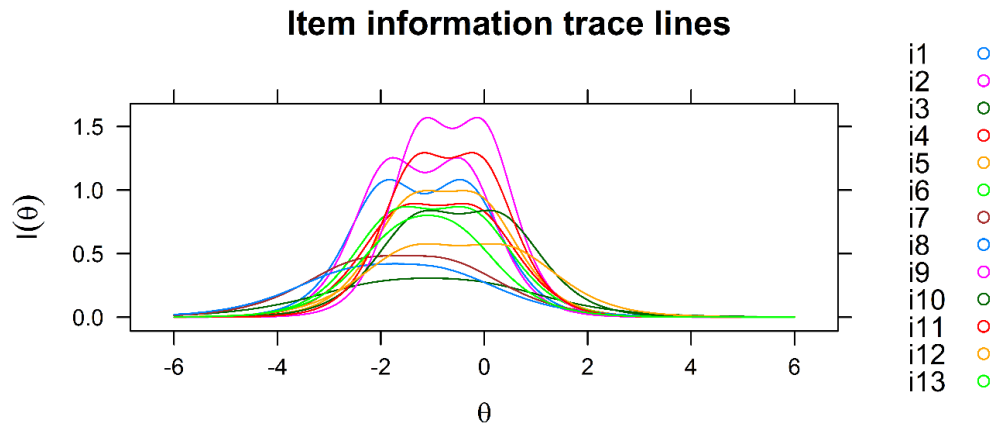
To compare both forms, the CFA analysis was repeated with a seven-item version of the CSS. The results showed that a one factor solution was confirmed with acceptable fit indices:  $\chi^2(2) = 101.05$ ,  $df = 14$ ,  $\chi^2(2) / df = 7.21$  CFI = 0.943, TLI = 0.914. RMSEA = 0.092(95% CI = 0.076-0.110), SRMR = 0.038. As presented above, the fit indices were not at an acceptable level for the 13-item version. It was clearly seen that removing items contributes to model data fit. The IRT based factor analysis was also computed by fitting the GRM model. The analysis showed that the fit values of the single-factor GRM model were at an acceptable level: CFI = 0.976. TLI = 0.952. RMSEA = 0.072. and SRMR = 0.055. As compared to the GRM model fitted with 13-items, the seven-item version provides better IRT based fit indices for a one factor solution.

Cronbach alpha coefficients were also computed for both versions in order to see how the internal consistency of the scale was affected by reducing the number of the scale. While the alpha value was .87 for the 13-item version, it was found that the value for the seven-item version was .83, showing a minimal drop of internal consistency after reducing to six items. The IRT-based empirical reliability values were also compared between both forms. The results were similar for both forms (.87 for the 13-item version vs .81 for the seven-item form). The correlation between both forms using were also computed using their total scores. It is .95 ( $p < 0.001$ ), indicating that the seven-item version could be used for same purposes as the 13-item version of the CSS.

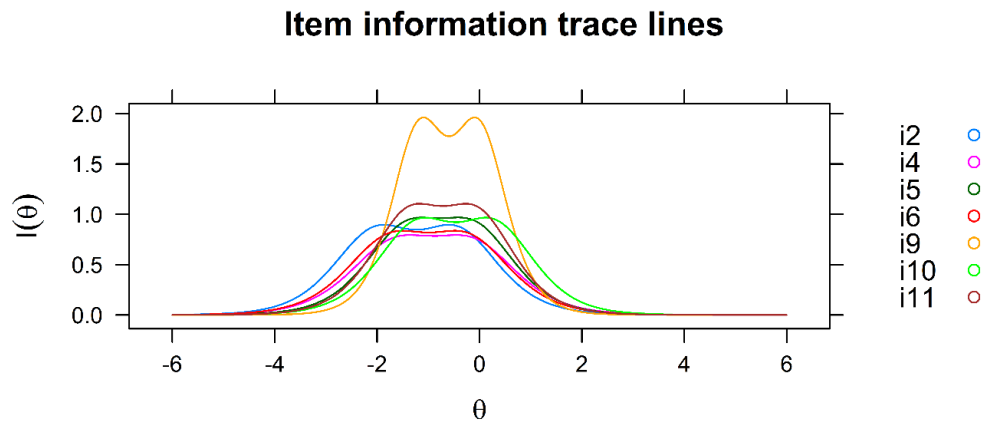
The two different versions of the scale were compared based on the item and test information functions. The item information functions of both versions are presented in Figures 1a and b respectively. As depicted in Figures 1a and b, both versions of the CSS scale contain items with varying information levels, which contributes to total test information, while most of the less contributing items were eliminated in the seven-item version. In addition, while the information they contribute show variation, the items are slightly homogenous in terms of the location where they provide their highest information. The range of the ability both versions provide is narrow to some extent but similar to each other while it could be inferred that both forms provided accurate measurement within the same range of cyberbullying sensibility levels in terms of the range of the information they provide (see Figure 2).

In the seven-item version, the two items with relatively less information were retained in the test because they could possibly provide information at the level of the ability range. All in all, we can infer that the item removing process mostly eliminates item with lower information except for two items that provided information at different locations.

a



b

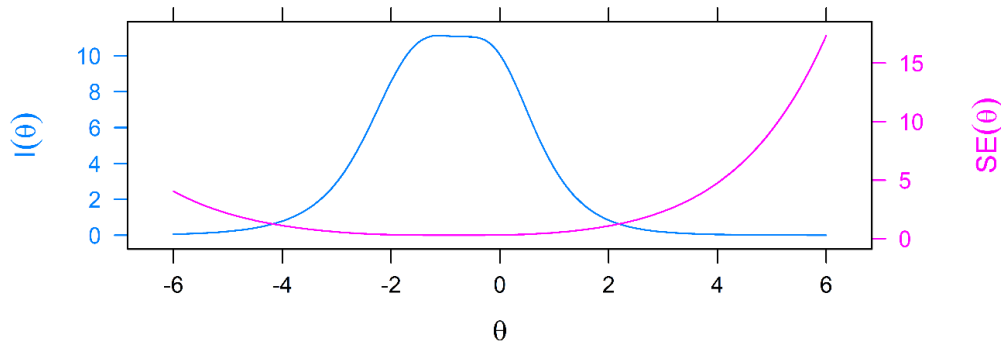


**Figure 1.** (a) Item Information Curves for thirteen item version of CSS. (b) Item Information Curves for seven items version of CSS. Note: Numbers in shorter version indicate the item number after items removed from original version of CSS

Figures 2a and b represent test information functions and a standard error of measurement of the 13-item and the seven-item versions of the CSS. As shown in Figure 2, both forms are similar in terms of the maximum information they provide across the ability spectrum. As expected, in terms of the information and precision of items, the 13-item version is better because it contains more items while the drop in information (and increase in the standard error) is not comparable with the proportion of the reduced items. The percentage of the remaining items were 54% while 65% of the information still retained.

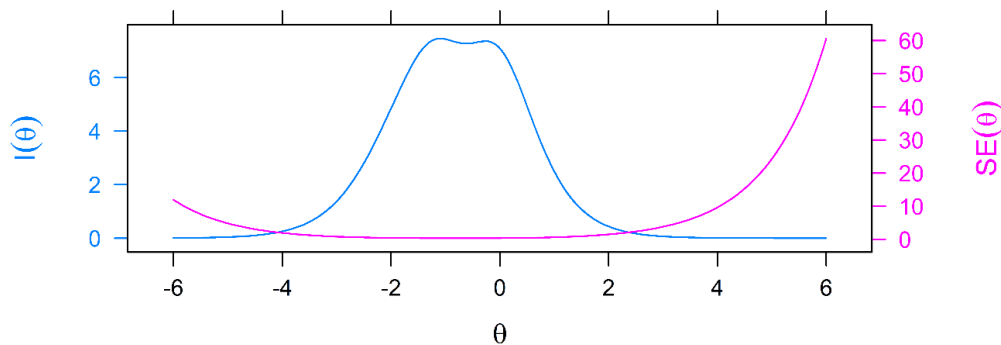
a

### Test Information and Standard Errors



b

### Test Information and Standard Errors



**Figure 2.** (a) Total test information for the initial 13-item CSS. (b) Total test information function for the seven-item CSS scale.

#### 4. DISCUSSION and CONCLUSION

The primary aim of this research was to validate the 13-item Cyberbullying Sensibility Scale developed by Tanrikulu, Kinay and Arıcak (2013) for university students. In this respect, item response theory (IRT) was used in order to obtain findings related to the psychometric properties of the scale. IRT based techniques were used to remove some items that interfered with the one-dimensional structure and provided below-average information. We expect that the IRT based item removal process would not adversely affect the reliability of psychometric properties of the scale and the amount of information it provided. Moreover, we hoped that its usefulness would increase thanks to the shortening of the scale. Shorter scales, in addition to increasing the usefulness in practice by enabling the addition of more variables to the research, expand the nomological network of cyberbullying to other psychological constructs. Finally, this validation study made the measuring of the cyberbullying sensibility among university students possible for interested researchers. Although there are scales of cyberbullying and cyber victimization that could be used with university students, there is a lack of a standardized scale of cyberbullying sensibility suitable for university students. This study will contribute to filling this gap in the literature.

The discrimination parameters obtained from the IRT analyzes showed that the seven-item cyberbullying sensibility scale validated for the university students was effective in distinguishing students with low and high levels of cyberbullying sensibility. On the other hand,

when the IRT-based difficulty parameters were examined, it was seen that the scale provided more accurate measurements for average-to-high-level individuals.

The item removal process was conducted based on two criteria: (a) removing items with below-average contribution to total test information, (b) removing items with poor item fit statistics. As a result of the IRT analysis performed with 13 items in the original form, the fit statistics of two items were found to be less than  $p < 0.05$  significance level (these two items were also the items that provide information below the average).

The IRT analyzes were repeated with the remaining eight items after dropping five items that had provided below average information. When the item fit statistics were examined again, it was found that the fit values of the first item were not at an acceptable level of significance ( $p < 0.05$ ), and this item was also removed from the scale. As a result of the IRT analysis carried out with the remaining seven items, the item elimination process was terminated by acknowledging that all of the remaining items fit well to the model.

As a result of the primary EFA and CFA analyzes performed, no evidence was obtained regarding the fact that the one-dimensional factor structure of the cyberbullying scale developed by Tanrikulu, Kınay, and Arıcak (2013) was maintained. After the item removal process, it was confirmed by CFA analysis that the seven-item version showed a one-dimensional structure. In addition, the reliability analysis conducted based on both the classical approach and the IRT showed that the reliability of the scale was similar to the 13-item version. In other words, as a result of the IRT-based item elimination, the construct validity of the scale approached the desired structure while the reliability level of the scale stayed almost the same.

This study contributed to the literature and experts working in the field of cyber psychology in many different ways. Firstly, as emphasized by different experts, understanding the importance of the concept of cyberbullying sensibility and its relationship to other psychological structures is vital. Even though there are many studies forming the nomological network of cyber victimization and cyberbullying with other psychological constructs (Ang & Goh, 2010; Ojedokun & Idemudia, 2013; Kokkinos, Antoniadou, & Markos, 2014;), there is an absence of literature on the construct of cyber sensibility. In addition, there are appropriate instruments for measuring the cyberbullying sensibility among high school students (Tanrikulu, Kınay, & Arıcak, 2013) while there is no instrument for such studies to be carried out with university students. This validated instrument will contribute to a better understanding of the cyberbullying sensibility of university students, and our understanding of the relationship between cyberbullying sensibility and other structures will expand for this population.

Secondly, reducing the number of items increased the usefulness of the scale while the amount of information it provides was mostly retained. In addition, the reliability level of the scale was kept at almost similar levels despite item removal. We expect that a shorter but still-reliable scale will be preferred by the researchers and practitioners. On the other hand, the research has some limitations that should be taken into consideration by the readers and scientists who will use this instrument in their research and practices. The data obtained in this study were collected only from the students who were studying at a private university. For this reason, there is a question about the generalizability of the findings. The university where the data collection process took place is located in a medium-sized metropolis. Inclusion of university students in larger metropolitan cities such as Istanbul into the data collection process and the inclusion of students in state owned universities will increase the generalizability of the findings. On the other hand, as Baker (2001) states, IRT parameters are independent of the sample data collected. Therefore, the parameters obtained are independent of the sample group. Hence, we think that generalizability of the results may not pose a serious problem based on the fact that the IRT was employed.



Secondly, evidence was obtained for the validation of the cyberbullying sensibility scale for university students. On the other hand, no data collection process was carried out to gather evidence of criterion validity. In addition, no test re-test process was carried out to determine whether the scale gives stable results. We recommend that future studies should focus on collecting evidence for criterion related validity and the stability of scores across time for the 7-item cyberbullying sensibility scale.

Third, the resulting difficulty parameters show that the scale can perform more accurate measurements for the average-to-high level of sensibility. This finding was observed for both the 13-item version and the seven-item version (see Table 1 and 2). When it is considered that the individuals who need preventive interventions by experts have low levels of cyberbullying sensibility, it is seen that it is necessary to add more questions to the scale that can give information for the lower level of cyberbullying sensibility. In the further revisions of the scale, we recommend that authors add items suitable for providing information on individuals with lower cyberbullying sensibility level for risky populations.

Differential Item Functioning (DIF) was not investigated in the current study. DIF occurs when different subgroups of participants (e.g., male and female) with the same latent trait level yield different response patterns. If DIF is detected, it poses a risk to the validity of the scale. Because CSS is a relatively new construct, there is little knowledge about whether some subgroups present more cyberbullying sensibility. Hence, we recommend the investigation of DIF across different subgroups of cyberbullying sensibility. Such an investigation might reveal possible differences in subgroups.

#### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

#### **ORCID**

Osman Tolca Arıcak  <https://orcid.org/0000-0001-8598-5539>

Akif Avcu  <https://orcid.org/0000-0003-1977-7592>

Feyza Topçu  <https://orcid.org/0000-0002-5853-2670>

Merve Gülçin Tutlu  <https://orcid.org/0000-0003-4225-7982>

#### **5. REFERENCES**

- Álvarez-García, D., Núñez, J.C., González-Castro, P., Rodríguez, C., and Cerezo, R. (2019) The Effect of Parental Control on Cyber-Victimization in Adolescence: The Mediating Role of Impulsivity and High-Risk Behaviors. *Front. Psychol.*, 10, 1159.
- Ang, R.P., & Goh, D.H. (2010). Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry & Human Development*, 41(4), 387-397.
- Bilker, W.B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, 19(3), 354-369.
- Baker, F.B. (2001). *The basics of item response theory (2nd ed.)*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland. Retrieved February, 3 2019 from <http://files.eric.ed.gov/fulltext/ED458219.pdf>
- Baştak, G., & Altınova, H.H. (2015). Lise Öğrencilerinde Yaratıcı Drama Yöntemiyle Siber Zorbalık Hakkında Duyarlılık Oluşturma. *Yaratıcı Drama Dergisi*, 10(1), 91-102.
- Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29.

- Doğan, E., Cansu, Ç., & Şahin, Y.L. (2016). A Study on Online Social Network Games Players' Cyberbullying Sensibility and Aims of Facebook Usage/Çevrimiçi Sosyal Ağ Oyunu Oynayan Bireylerin Siber Zorbalığa Duyarlılık Düzeyleri ile Facebook Kullanım Amaçları Üzerine Bir Çalışma. *Eğitimde Kuram ve Uygulama*, 12(3), 501-520.
- Embretson, S., Reise, S.P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates. Inc. Mahwah.
- Gaffney, H., Farrington, D. P., Espelage, D. L., and Ttofi, M. M. (2019). Are cyberbullying intervention and prevention programs effective? a systematic and meta-analytical review. *Aggress. Violent Behav.* 45, 134–153. doi: 10.1016/j.avb.2018.07.002
- Gahagan, K., Vaterlaus, J.M., & Frost, L.R. (2016). College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility. *Computers in human behavior*, 55, 1097-1105.
- Hambleton, R.K, Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Thousand Oaks: Sage Publications.
- Hambleton, R.K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In: Tinsley HEA, Brown SD, editors. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. San Diego: Academic. p. 553–85.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139–64.
- Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Istiyono, E., Dwandaru, W.S.B., Ledo, Y.A., Rahayu, F., & Nadapdap, A. (2019). Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge. *International Journal of Instruction*, 12(4), 267-280.
- Kokkinos, C.M., Antoniadou, N., & Markos, A. (2014). Cyber-bullying: An investigation of the psychological profile of university student participants. *Journal of Applied Developmental Psychology*, 35(3), 204-214.
- Lee, J., Abell, N., & Holmes, J.L. (2015). Validation of measures of cyberbullying perpetration and victimization in emerging adulthood. *Research on Social Work Practice*. <http://dx.doi.org/10.1177/10497315155578535>
- Mielenz, T.J., Edwards, M.C. & Callahan, L.F. (2010). Item response theory analysis of two questionnaire measures of arthritis-related self efficacy beliefs from community based US samples. *Hindawi Publishing Corporation Arthritis*.
- Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus User's Guide: Statistical Analysis with Latent Variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nedim-Bal, P., & Kahraman, S. (2015). The Effect of Cyber Bullying Sensibility Improvement Group Training Program on Gifted Students. *Journal of Gifted Education Research*, 3(2). 48-57.
- Ojedokun, O., & Idemudia, E. S. (2013). The moderating role of emotional intelligence between PEN personality factors and cyberbullying in a student population. *Life Science Journal*, 10(3), 1924-1930.
- Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL <https://www.R-project.org/>

- Rubio, V.J., Aguado, D., Hontangas, P.M., & Hernandez, J.M. (2007). Psychometric properties of an emotional adjustment measure. *European Journal of Psychological Assessment*, 23(1), 39-46.
- Samejima, F. (1969) Estimation of Latent Ability Using a Response Pattern of Graded Scores. (Psychometrika Monograph, No. 17). Psychometric Society, Richmond. <http://www.psychometrika.org/journal/online/MN17.pdf>
- Tanrikulu, I. (2018). Cyberbullying prevention and intervention programs in schools: A systematic review. *School psychology international*, 39(1), 74-91.
- Tanrikulu, T., Kınay, H. & Arıcak, O.T. (2013). Cyberbullying sensibility scale: validity and reliability study. *Trakya University Journal of Education*, 3(1), 38-47.
- Tanrikulu, T., Kınay, H., & Arıcak, O. T. (2015). Sensibility development program against cyberbullying. *New Media & Society*, 17(5), 708-719.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567–577.
- Tolia, A. (2016). Cyberbullying: Psychological effect on children. *The International Journal of Indian Psychology*, 3(2), No. 1. 48-51.
- Tsutakawa, R.K., & Johnson, J.C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, 55, 371–390.
- Willoughby, M. (2018). A review of the risks associated with children and young people's social media use and the implications for social work practice. *Journal of Social Work Practice*. 1-14. doi:10.1080/02650533.2018.1460587
- Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Zanon, C., Hutz, C.S., Yoo, H., & Hambleton, R.K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29, 1-10.

## Investigation of PISA 2015 Reading Ability Achievement of Turkish Students in Terms of Student and School Level Variables

Kazim Celik<sup>1,\*</sup>, Ahmet Yurdakul<sup>2</sup>

<sup>1</sup>Department of Education, Faculty of Education, Pamukkale University, Denizli, Turkey

<sup>2</sup>Directorate of National Education, Usak, Turkey

### ARTICLE HISTORY

Received: 09 July 2019

Revised: 25 November 2019

Accepted: 26 January 2020

### KEYWORDS

PISA,  
Reading ability,  
Hierarchical linear model,  
Student level factors,  
School level factors,

**Abstract:** The aim of this study is to determine the students-level and school-level factors that are related to reading ability achievement of students who participated PISA 2015 (Programme for International Student Assessment) from Turkey. The effects of the student and school level factors on reading achievement of students were tested by 2 level hierarchical linear model. According to the findings, there are differences between schools in terms of students' reading ability scores in Turkey. When the findings of the effects of the student level variables on the reading ability scores are examined; mother's socio-economic status, parental emotional support, and unfair teacher behavior variables seem to affect students' reading ability achievement. When the findings of the effects of the school level variables on the reading ability scores are examined; school size, teacher education level, and student behavior that hinders learning variables have a significant effect on the average reading ability scores of schools. When the student and school level variables mentioned above were modeled together, the significant effect of the school size variable was lost while the teacher education level and the student behavior that hinders learning variables continued to have a significant effect on the schools average reading ability scores.

## 1. INTRODUCTION

The evaluation of the quality of education is crucial. The most commons of these evaluation methods in recent years are the PISA and TIMMS exams which are also applied in Turkey. Thanks to these exams, countries can see their level of education in the world compared to other countries. The PISA (Programme for International Student Assessment) is a screening survey conducted by the Organization for Economic Co-operation and Development (OECD) every 3 years and assesses the knowledge and skills gained by 15-year-old students (OECD, 2000; Schleicher, 2007; Breakspear, 2012). PISA focuses on the ability of young people using their skills and knowledge to cope with real life challenges (Reinikainen, 2012). The assessment, which focuses on reading, mathematics, science and problem-solving, not only recognizes that students can repeat what they have learned, but also examines how well they are able to benefit what they have learned and how they can apply this knowledge in and out of the school

CONTACT: Kazim Celik ✉ [kcelik@pau.edu.tr](mailto:kcelik@pau.edu.tr) ☒ Department of Education, Faculty of Education, Pamukkale University, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

environment. This approach reflects the fact that modern societies reward what the students can do instead of what they know (OECD, 2014).

In a short time, PISA has gone a long way and reinforced the role of OECD and its Education Directorate as a leading global organization to develop and analyze comparative international education performance data. PISA results now have a very high profile in the national media and are in the awareness of high-level politicians (Fairclough, 2000; Lingard & Rawolle, 2004, Grek, 2009). As Gür, Çelik and Özoğlu (2012) stated countries shape their educational policies according to PISA results. PISA results enable policy makers from around the world to measure the knowledge and skills of their country's students compared to other countries, to set policy targets, and to learn from applied policies and practices elsewhere (Ringarp & Rothland, 2010).

PISA assesses the application of knowledge in mathematics, reading and science literacy to problems in the context of real life (OECD, 1999). The PISA uses the term "literacy" in each area to show focus on the application of knowledge and skills. For example, when reading is assessed, PISA assesses how well students in the 15<sup>th</sup> year understand, use and reflect the written text for various purposes and environments. In science, PISA assesses how students can apply scientific knowledge and skills to different situations they may encounter in their lives. Similarly, in mathematics, PISA evaluates how students analyze, reason, and interpret mathematical problems in various situations. The scores on the PISA scales represent skill levels throughout the continuity of literacy skills. PISA provides a range of proficiency levels associated with points that define what a student can typically do at each level (OECD, 2006).

Mathematics and science education constitute large and dynamic elements of schooling that are generally viewed as important to individual students in enhancing their understanding of the world and improving their chances of lifetime achievement and also important at the larger societal level in today's knowledge-based economy where the capacities of the citizenry are directly linked to the well-being of the nation. However, the importance of mathematics and science education is a distant second compared to the importance ascribed to language and literacy education, especially reading (Yore, Anderson & Chiu, 2010). According to Wellington and Osborne (2001); to be successful in math and science, students should understand what they read. Therefore, in this study the factors that affect the reading literacy of pupils are investigated.

Apart from tests that assess students' knowledge in mathematics reading and science literacy, in PISA, some questionnaires are also applied. These questionnaires are designed to get information about students and their families' background including their economic, social and cultural capital, students' attitudes towards learning, the life aspects of students such as their habits in and out of the school and, their lives and families, the quality of human and material resources of schools, aspects of the school's such as teaching and learning processes, staffing practices and emphasis on curriculum and extracurricular activities, organizational structures and genres, class size, classroom and school environment and instructional content, including science activities in the classroom (Rindermann, 2007). In the current study the factors that affect the reading literacy achievement of students are determined according to the results of these questionnaires.

When the literature is examined, some research studies on the subject are available. Yildirim (2012) investigated the student and school variables that influence PISA 2009 reading comprehension skills, Willms (2001) investigated the differences in the level of reading comprehension in the Canadian provinces, and the factors that make this difference, Lietz and Kotte (2004) compared the factors that affect the achievement of Finland's, which is the most successful country in reading skills according to PISA 2000 results, and Germany's, which is under average, Linnakyla, Malin and Taube (2004) compared PISA 2000 reading ability scores of Finnish and Swedish students and tried to explain the reason of the difference, Kotte, Lietz

and Lopez (2005) used PISA 2000 data to investigate the factors that encourage and impede students' reading achievement in Germany and Spain, Nonoyama (2006) investigated the factors that affect the school-based and family-based factors in achievement of students in reading skills. Besides these Thomson, Bortoli, Nicholas, Hillman and Buckley (2010) investigated PISA results for Australia, Wilms (2001) and Catwright and Allen (2002) for Canada, Rindermann (2007) for Denmark, Brozo, Shiel and Topping (2007) for Ireland, Scotland and the U.S.A, Grek (2009) for England, Finland and Germany.

When the literature was investigated, no studies were conducted on PISA 2015 data. When we look at the work done in the past years, we have not encountered a study which deals specifically with the variables of our study. PISA results reveal what is possible in education by demonstrating what the highest performing and fastest growing educational system can do. While applying cognitive tests to students, at the same time PISA also applies student, parents and school questionnaires in order to evaluate the factors that affect students' achievement. In the student questionnaire, the student is asked questions about his / her home (family, computer use, technology use, etc.); In the school questionnaires, the school administrator or an authorized person is asked questions about the structure of the school, the resources of the school, the situation of students and teachers, educational policies and school climate. While the PISA project demonstrates the academic achievement of students on an international scale through cognitive tests, it also measures the relationship between school resources at national and international levels with the data obtained from school surveys; it also reveals similarities and differences between different schools. Findings allow policy makers around the world to measure the knowledge and skills of students in their countries compared to other countries, to set policy targets against measurable targets in other education systems, and to learn from applied policies and practices elsewhere (MEB, 2015). But according to Özdemir (2017), in Turkey the studies on PISA are mostly not original, these studies are repetition or the interpretation of the results published by MEB (Turkish Ministry of National Education) and OECD. Therefore, original studies in this field are needed. The hierarchical linear modeling approach is a two-level strategy (Hoffman, Griffin and Gavin, 2000) that investigates the variables involved in two-step analysis. As the PISA data are hierarchical, this approach is very convenient. Students who are the sample group of PISA are in classrooms, classes in schools, schools in cities, and cities in countries. In this context the aim of this study is to determine the personal and school-based factors that affect Turkish students' reading achievement in PISA 2015.

## **2. METHOD**

### **2.1. Study Group**

In PISA 2015, the 15 year-old student population was determined to be 925.366 students. In the PISA study, school sampling was determined by stratified random sampling method. At the first stage for PISA 2015 application, schools were selected by stratified random sampling method in the Classification of Statistical Region Units (NUTS) Level 1, type of education, type of school, place of schools and administrative forms of schools, and in the second stage students who were to participate in these schools were determined by random method . 5895 students from 187 schools in 61 cities that represent 12 regions participated in the exam (MEB, 2015). 20.7 % of students participating in PISA 2015 application are 9th grade students and 72.9% are 10th grade students. 75% of the students attend vocational high schools and Anatolian high schools. 50% of the students are male and 50% of the students are female.

### **2.2. Data Collection Tools**

The data on students' achievements in reading comprehension, mathematics and science, and demographic, socio-economic and educational variables that may be related to achievement in

these areas were collected by PISA 2015 performance tests and questionnaires. In the research, these tests and questionnaires were used. These data for PISA 2015 was obtained from the official website of the OECD (<http://www.oecd.org/pisa/data/2015database>) which carried out PISA applications. The student and school level variables determined in the study are given in [Table 1](#).

**Table 1.** Student and School Level Explanatory Variables

Student Level (1st Level)	School level (2nd Level)
Mother's Education Level	School Size
Father's Education Level	Educational Leadership of School Director
Mother's Socio Economic Status	Instructional Leadership of School Director
Father's Socio Economic Status	Lack of Educational Equipment
Class Repetition	Lack of Staff
Study Time Out of School	Teacher Behavior That Hinders Learning
Math Study Time	Student behavior That Hinders Learning
Turkish Study Time	
Science Study Time	
Belonging to School	
Co-operation Skills	
Parental Emotional Support	
Unfair Teacher Behavior	

### 2.3. Data Analysis Procedure

After the normality assumptions are tested by Kolmogorov-Smirnov test and met, a two-level HLM was conducted to determine the relation between PISA 2015 students' reading ability achievements and student and school characteristics. Data statistical programs and; a hierarchical linear model program was used. A minimum of .05 was taken as the basis for the statistical significance test.

In social sciences, data are generally nested hierarchically in structure. The best example of this situation is seen in educational sciences. The students are in the classes, the classes are in the schools, the schools are in the regions and the regions are in the countries. Therefore, when we do analyses, we can not consider the students separate from the classes or the schools they are in. Hierarchical linear models allow us to analyze these variables together (Raudenbush & Bryk, 2002). In this study, a two-level hierarchical model was established by taking student variables as Level 1 and school variables as Level 2.

Four HLM models were used to reach the objectives of the study. These; The One-Way ANOVA with random effects model, the random coefficient regression model, the regression model in which the intercepts are outcomes, and the model in which intercepts and slopes are outcomes. A One-Way ANOVA with random effects model; is the simplest form of hierarchical linear models. It is also called an empty model (Hox, 2002). First, a One-Way ANOVA with random effects model is established and hierarchical linear models are started. The objective is to distinguish the dependent variable according to the different levels of the hierarchy. This model includes only random groups and variances within these groups. The One-Way ANOVA with random effects model is used to generate the point estimate and confidence interval for the

large intercept. It also provides information on output variability in each of the two levels (Acar, 2013). In the random coefficient regression model, all of the submodels are treated with the assumption that the fixed parameter is a randomly changing model. There are no Level 2 independent variables in the model that explain the intercept and slope parameters. In the regression model in which the intercepts are outcomes, the predictions are made using the Level 2 variables. Regression model consists of group intercepts which are predicted by Level 2 variables. Within the scope of the research, Level 1 of this model was constructed as the first step of the random-effects ANOVA model. In Level 2, school-level variables, the effects of which are sought on students' reading achievement, are added to the model. The last model is the model in which intercepts and slopes are outcomes. It is also called as full model as it contains all the 1st and 2nd Level variables together (Raudenbush & Byrk, 2002). In this model Level 1 variables are added and the change on the effect of Level 2 variables on the dependent variable is observed.

### 3. RESULT / FINDINGS

We presented the descriptive statistics for the first sub-problem of the study in Table 2. We established a One-Way ANOVA with random effects model in order to answer the research question: “Are there differences between schools in terms of the students' reading achievement?”

**Table 2.** Fixed Coefficients for One-Way ANOVA with Random Effects Model

Fixed Effects	Coefficients	Standard Error	t	df	p
Reading Skills	458.861	5.728	80.105	110	0.000

As depicted in Table 2, there is a significant difference in reading skills achievement among schools ( $p < 0.01$ ). The average reading scores of schools are 458,861. We presented the random effect for one-way ANOVA with random effects model in table 3.

**Table 3.** Random Effects for One-Way ANOVA with Random Effects Model

Random Effect	Variance	df	$\chi^2$	p
INTERCEPT	2861.809	110	887.301	.000
Level1 Effect	2213.043			

As seen in Table 3, the random effects are significant in school level ( $\chi^2 = 887.301$ ,  $df = 110$ ,  $p < 0.01$ ). This indicates that the difference between the schools in terms of the average reading comprehension scores is random. In addition, it was determined that the change in the intercept score of reading between schools was caused 56% by school variables and 44% by student variables ( $2861 / (2861 + 2213)$ ). Reliability is calculated as 0.78 when the reliability of the Level 1 coefficients, which give information about whether the average obtained from the sample is a sign of the actual school average. This suggests that the average obtained from the sample is a reliable indicator of the true school average. In this respect, the model is established as follows.

Level1 Model

$$READING_{ij} = \beta_{0j} + r_{ij}$$

Level2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Mixed Model



$$\text{READING}_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

READING = It is the literacy intercept of the students attended PISA 2015 application from Turkey.

$\beta_{0j}$  = It is the literacy intercept of the students for school.

$r_{ij}$  = It is the error of Level 1 equation.

$\gamma_{00}$  = It is the intercept of schools number  $j$ .

$u_{0j}$  = It is the random effect.

We established the random coefficient regression model in order to answer the second sub-problem of the study “What are the student-level variables that have significant effects on students' reading comprehension achievements?”. We initially included 13 variables to the model. These are; the educational status of the mother, the educational status of the father, the socio-economic level of the mother, the socio-economic level of the father, the grade repetition, the study time outside the school, the student behavior that prevents learning, mathematics study time, Turkish study time, science study time, feelings of belonging, enjoyment of cooperation, emotional support of the family, unfair teacher behavior. The significant ones among these variables are presented in [Table 4](#).

**Table 4.** Fixed Coefficients for Random Coefficient Regression Model

Fixed Effects	Coefficients	Standard Error	t	df	p
Intercept, $\beta_{00}$	456.077067	5.531659	82.449	118	0.000
Mother's Socio-Economic Status, $\gamma_{10}$	0.504272	0.109915	4.588	374	0.000
Parents Emotional Support, $\gamma_{20}$	5.715517	2.388353	2.393	374	0.017
Unfair Teacher Behavior, $\gamma_{30}$	-2.107095	0.614764	-3.427	374	0.000

When [Table 4](#) is investigated, it is seen that the variables that affect students' reading ability achievement are mother's socio-economic status, parents' emotional support and unfair teacher behavior. According to [Table 4](#), there is a significant positive correlation between the student's reading ability score and the mother's socio-economic status ( $\beta_{10} = 0.504$ , SE= 0.10,  $p < 0.05$ ). When other variables are held constant, a unit increase in mother's socio-economic status increases reading skills by 0.504 units. There is a significant positive correlation between students' reading ability scores and family emotional support ( $\beta_{20} = 5.715$ , SE= 2.388,  $p < 0.05$ ). Students who perceive the emotional support of the family are more successful in the field of reading skills than other students. There is a significant negative correlation between students' reading ability scores and unfair teacher behavior perception variables ( $\beta_{30} = -2.107$ , SE= 0.61,  $p < 0.05$ ). It is seen that the more unfair teacher behaviors are, the less students' reading ability scores are. According to these data, we presented the model as;

Level 1 Model

$$\text{READING}_{ij} = \beta_{0j} + \beta_{1j}*(\text{BSMJ}_{1ij}) + \beta_{2j}*(\text{PARSUP}_{ij}) + \beta_{3j}*(\text{UNFAIRTE}_{ij}) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Mixed Model

$$READING_{ij} = \gamma_{00} + \gamma_{10} * BSMJ1_{ij} + \gamma_{20} * PARSUP_{ij} + \gamma_{30} * UNFAIRTE_{ij} + u_{0j} + r_{ij}$$

READING = It is the literacy intercept of the students attended PISA 2015 application from Turkey.

BSMJ1 = It is mother’s socio-economic status.

PARSUP = It is parental emotional support.

UNFAIRTE = It is unfair teacher behavior.

$\beta_{0j}$  = It is the reading intercept of j school.

$\beta_{1j} \dots \beta_{4j}$  = Intercept differences between schools.

$\gamma_{00}$  = It is the expected value of constant parameters on Level 2 units.

$\gamma_{10} \dots \gamma_{40}$  = They are the expected value of slope parameters on Level 2 units.

$u_{0j}$  = It is Level 2 j unit’s change in constant parameter.

$r_{ij}$  = It is the error of Level 1.

The regression model in which the intercepts are outcomes was established in order to answer the 3<sup>rd</sup> sub-question of the study “What are the school level variables that that have significant effects on students' reading comprehension achievements?”. Table 5 shows the fixed and random effects on the intercept reading score of the school according to the regression model in which the intercepts are outcomes of school variables.

**Table 5.** Fixed and Random Effects of the Regression Model in Which The Intercepts Are Outcomes of School Variables

Fixed Effects		Coefficients	Standard Error	t	df	p
INTERCEPT, $\gamma_{00}$		458.707530	4.880027	93.997	107	0.000
SCHOOLSIZE, $\gamma_{01}$		-0.024048	0.007177	-3.351	107	0.001
TEACHER EDUCATON LEVEL, $\gamma_{02}$		477.0057	179.673	2.655	107	0.009
STUDENT BEHAVIOR, $\gamma_{03}$		-22.774649	4.843012	-4.703	115	0.000
Random Effect	Variance	df	$\chi^2$	p		
INTERCEPT, $r_0$	2008.94783	107	613.75481	0.000		
Level 1 Effect, e	2208.15542					

As depicted in Table 5, the school level variables that affect students’ reading ability achievement are school size, teachers’ educational level, and student behavior that hinders learning. There is a negative correlation between students' reading ability scores and school size ( $\beta_{01} = -0.024$ , Standard Error (SE) = 0.007,  $p < 0.05$ ). As the number of students increases, the scores of reading skills of the students decrease. There is a positive significant relationship between the reading ability scores of the students and the education levels of the teachers ( $\beta_{02} = 477.005$ , SE = 179.673,  $p < 0.05$ ). As the level of education of teachers increases, the scores of students' reading skills also increase. There is a negative relationship between students' reading ability scores and student behaviors that prevent learning ( $\beta_{03} = -22,774$  SE = 4.84,  $p < 0.05$ ). The more students’ behaviors that prevent students from learning is, the lower their reading ability scores are. According to these data the regression model in which the intercepts are outcomes is established as;

Level-1 Model

$$READING_{ij} = \beta_{0j} + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(SCHSIZE_j) + \gamma_{02}*(TEACEDLEV_j) + \gamma_{03}*(TEACBEH_j) + u_{0j}$$

Mixed Model

$$READING_{ij} = \gamma_{00} + \gamma_{01}*(SCHSIZE_j) + \gamma_{02}*(TEACEDLEV_j) + \gamma_{03}*(TEACBEH_j) + u_{0j} + r_{ij}$$

SCHSIZE = It is the school size.

TEACEDLEV = It is the teachers' education level.

TEACBEH = It is the teacher behavior that hinders learning.

$\beta_{0j}$  = It is the reading achievement of school j.

$\beta_{1j}$ .....  $\beta_{4j}$  = Intercept differences between schools.

$\gamma_{00}$  = It is the expected value of constant parameters on Level 2 units.

$\gamma_{01}$ .....  $\gamma_{03}$  = They are the differentiating effects of school level variables on school average achievement.

$\gamma_{10}$ ..... $\gamma_{40}$  = They are the expected value of slope parameters on Level 2 units.

$u_{0j}$  = It is Level 2 j unit's change in constant parameter.

$r_{ij}$  = It is the error of Level 1.

We established the model in which intercepts and slopes are outcomes to answer the 4th sub-question of the study “When the student-level variables that affect reading ability achievement of the students significantly are added in the model, how do school level variables affect reading ability achievement of the students?”. In [Table 5](#) we presented the results of the model in which intercepts and slopes are outcomes.

**Table 6.** Fixed and Random Effects of the Model in Which Intercepts and Slopes Are Outcomes

Fixed Effects	Coefficients	Standard Error	t	df	p
Intercept, $\gamma_{00}$	443.436647	6.627748	66.906	107	<0.001
SCHOOLSIZE, $\gamma_{01}$	-0.002900	0.011157	-0.260	107	0.795
TEACHER EDUCATION LEVEL, $\gamma_{02}$	432.939496	170.412240	2.541	107	0.013
STUDENT BEHAVIOR, $\gamma_{03}$	-19.945674	4.590593	-4.345	107	<0.001
MOTHER'S SOCIO-ECONOMIC STATUS, $\beta_1$					
Constant, $\gamma_{10}$	0.358402	0.102493	3.497	371	<0.001
School Size, $\gamma_{11}$	-0.000528	0.000186	-2.840	371	0.005
Constant, $\gamma_{20}$	5.428292	2.251499	2.411	371	0.016
UNFAIR TEACHER BEHAVIOR, $\beta_3$					
Constant, $\gamma_{30}$	-1.579701	0.520183	-3.037	371	0.003
School Size, $\gamma_{31}$	0.002506	0.001194	2.099	371	0.037
Random Effect	Variance	df	$\chi^2$	p	
Intercept, $r_0$	2034.02824	115	710.76624	0.000	
Level1 Effect, e	1980.39148				

When [Table 6](#) is investigated it can be seen that the student level variables are added in the model the significant effect of school level variables on students' reading ability achievement that were presented in the regression model in which the intercepts are outcomes, except for

school size, continues ( $p < 0,05$ ). However, the significant effect of the school size variable was lost when the student-level variables included in the model. However, when the school size variable is combined with the unfair teacher behavior variable, it has a significant effect. When the school size variable model is combined, the negative effect of students' perception of unfair teacher behavior decreases to some extent. Again, the school size becomes significant when it is combined with the socio-economic status of mother variable. In large schools, the effect of socio-economic status of the mother is less. In this respect, the model in which intercepts and slopes are outcomes is;

Level-1 Model

$$\text{READING}_{ij} = \beta_{0j} + \beta_{1j} * (\text{BSMJ1}_{ij}) + \beta_{2j} * (\text{PARSUP}_{ij}) + \beta_{3j} * (\text{UNFAIRTE}_{ij}) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * (\text{SCHSIZE}_j) + \gamma_{02} * (\text{TEACEDLEV}_j) + \gamma_{03} * (\text{TEACBEH}_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * (\text{SCHSIZE}_j)$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31} * (\text{SCHSIZE}_j)$$

Mixed Model

$$\text{READING}_{ij} = \gamma_{00} + \gamma_{01} * \text{SCHSIZE}_j + \gamma_{02} * \text{TEACEDLEV}_j + \gamma_{03} * \text{TEACBEH}_j + \gamma_{10} * \text{BSMJ1}_{ij} + \gamma_{11} * \text{SCHSIZE}_j * \text{BSMJ1}_{ij} + \gamma_{20} * \text{PARSUP}_{ij} + \gamma_{30} * \text{UNFAIRTE}_{ij} + \gamma_{31} * \text{SCHSIZE}_j * \text{UNFAIRTE}_{ij} + u_{0j} + r_{ij}$$

SCHSIZE = It is the school size.

TEACEDLEV = It is the teachers' education level.

TEACBEH = It is the teacher behavior that hinders learning.

BSMJ1 = It is mother's socio-economic status.

PARSUP = It is parents emotional support.

UNFAIRTE = It is unfair teacher behavior.

$\beta_{0j}$  = It is the reading intercept of j school.

$\beta_{1j}$ .....  $\beta_{4j}$  = Intercept differences between schools.

$\gamma_{00}$  = It is the expected value of constant parameters on Level 2 units.

$\gamma_{01}$ .....  $\gamma_{03}$  = They are the differentiating effects of school level variables on school average achievement.

$\gamma_{10}$ ..... $\gamma_{40}$  = They are the expected value of slope parameters on Level 2 units.

$u_{0j}$  = It is Level 2 j unit's change in constant parameter.

$r_{ij}$  = It is the error of Level 1.

#### 4. DISCUSSION and CONCLUSION

Our study states that there is a significant difference between schools in terms of reading ability achievement of students in Turkey. This is consistent with many studies. Thomson et.al. (2010) examined PISA 2009 reading skills outcomes for Australia and found that there were significant differences between Australian schools in reading skills. Similar studies were conducted by, Rindermann (2007) for Denmark, Brozo, Shiel and Topping (2007) for Ireland, Scotland and the USA, Grek (2009) for England, Finland and Germany and Yıldırım (2012) for Turkey. In all of these studies, the PISA reading ability scores showed significant differences between schools. As the studies were conducted in different years and in different countries, it can be said that there are differences in reading skills across the world in all years.

In our study, we investigated the factors that cause the differences between schools and we found out that the student level variables are; socio-economic status of the mother, emotional

support of the family, and perception of the teacher's unfair teacher behavior. Acar (2013) investigated the results of 2005 and 2008 Turkey Student Achievement Test (SAT) and found the student-level factors that affect reading ability achievement of students as; gender of the student, father's education level, the number of books the student has, the time for reading, the belief of success in Turkish class and the state of taking private lessons from Turkish. In her study, in which she investigated the PISA 2009 reading ability achievements of students, Yıldırım (2012) revealed that the factors that affect students' success are; gender, enjoyment of reading, parents' socio-economic status and number of the books the student has.

As a result of the research, it was revealed that the socio-economic status of the mother affects the reading ability score positively. This result is consistent with the studies made by Anılan (1998), Ates (2008), Bölükbaşı (2010), Kaldan (2007), Öztürk (2010) and Kahraman and Çelik (2017). As the socio-economic status and educational level of the parents increase, students' reading achievement increases. Moreover, this is not only the case for reading achievement, but also for other courses. For example, Erberber (2010) found that parents' socio-economic status was influential on the mathematical achievement of students in TIMSS. According to our study another variable that affects reading achievement of students is parents' emotional support. Christenson, Rounds and Gorney (1992), Desimone (1999), Swap (1993), Gümüseli (2004) and Çelenk (2003) put forward that family support is a prerequisite for success. From this point of view, we suggest that organizing seminars for parents to increase their support for their children may be helpful to increase students' success.

Another result of our study is that there is not a significant correlation between study time out of school and having private Turkish lessons and students' reading achievement. In the study that they investigated the effect of homework on students' achievement, Kapıkıran and Kıran (1999) concluded that students who get less homework are more successful than students who get more homework. However; Grodner and Rupp (2013) concluded that homework had positive effects on the learning of students. Considering this data, it can be said that the quality of the homework given to the students is important and sufficient homework should be given so that students can do reinforcement instead of giving too much homework.

We found out that unfair teacher behaviour perceptions of students affect students' reading achievement in a negative way. Fryer (2013), Allen, Gregory, Mikami, Lun, J., Hamre and Pianta (2013) and Jones and Jones (2015) found significant relationships between teacher behavior and student achievement. As the positive teacher behaviors increase, the success and attitudes of the students are found to be positive. For this reason, it can be said that the positive attitude and behavior of the teachers towards the students will increase the success.

As a result of the research, school variables that affect students' reading achievement were school size, teachers' education level and student behavior that hinders learning. There is a negative relationship between school size and the success of reading skills. Fredriksson, Öckert and Oosterbeek (2013) and Leithwood and Jantzi (2009) found negative relationships between school size and success. Therefore, smaller schools instead of larger ones may increase the success.

According to the results of the research, there is a positive relation between the education level of the teachers in the schools and the reading ability scores of the students. As the education levels of teacher's increase, the success of students' reading skills also increases. In this data, teachers should be encouraged to constantly improve themselves and to continue their master education.

In this study the data from Turkey were used. Other countries can be included in the study and comparative statistics can be made. Results related to reading skills in Turkey was investigated. The results of science and mathematics achievement may also be investigated. Two levels of HLM was done in the study by taking the students and school levels into consideration. Levels

such as class and district levels can be added and three or four level models can be established. The research was conducted for PISA exam results. Similar examinations can be made in examinations like TIMMS and the results of the research can be compared.

### Acknowledgements

This study was derived from the PhD thesis supported by Pamukkale University Scientific Research Projects Coordination Unit with project number 2018EĞBE009.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### ORCID

Kazım ÇELİK  <https://orcid.org/0000-0001-7319-6567>

Ahmet YURDAKUL  <https://orcid.org/0000-0002-9995-7157>

## 5. REFERENCES

- Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review, 42*(1), 76.
- Anılan, H. (1998). *Beşinci sınıf öğrencilerinin Türkçe dersinde okuduğunu anlama becerisiyle ilgili hedef davranışların gerçekleşme düzeyleri [Level of achievement of the general objectives related to the ability of understanding reading in Turkish lessons by fifth grade students]*. Unpublished Master Dissertation, Pamukkale Üniversitesi, Denizli.
- Akyol, H., Ateş, S., & Yıldırım, K. (2008). *Sınıf öğretmenlerinin kullandıkları okuma stratejileri ve bu stratejileri tercih nedenleri [Reading strategies used by primary school teachers and the reasons for choosing these strategies]*. VII. National Symposium on Teacher Education (2-4 May 2008). Çanakkale. Çanakkale Onsekiz Mart Üniversitesi. Ankara: Nobel Yayın Dağıtım, 723-728.
- Bölükbaşı, C., & Sarıbaş, M. (2011). İlköğretim Birinci Kademe (1, 2, 3. Sınıf) Türkçe Öğretimi Sorunları. [Primary Education (1st, 2nd, 3rd Grade) Turkish Teaching Problems]. *Academic student journal of Turkish researches, 1*(1), 20-26.
- Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers, 71*, 1.
- Brozo, W. G., Shiel, G., & Topping, K. (2007). Engagement in reading: Lessons learned from three PISA countries. *Journal of Adolescent & Adult Literacy, 51*(4), 304-315.
- Cartwright, F., & Allen, M. K. (2002). *Understanding the rural-urban reading gap*. Human Resources Development Canada. Service Canada, Ottawa, ON K1A 0J9, Canada.
- Christenson, S. L., Rounds, T., & Gorney, D. (1992). Family factors and student achievement: An avenue to increase students' success. *School Psychology Quarterly, 7*(3), 178.
- Çelenk, S. (2003). Okul başarısının ön koşulu: Okul aile dayanışması [The Prerequisite for School Success: Home-School Cooperation]. *İlköğretim online, 2*(2).
- Desimone, L. (1999). Linking parent involvement with student achievement: Do race and income matter? *The journal of educational research, 93*(1), 11-30.
- Erberber, E. (2010). Analyzing Turkey's data from TIMSS 2007 to investigate regional disparities in eighth grade science achievement. In *the Impact of International Achievement Studies on National Education Policymaking* (pp. 119-142). Emerald Group Publishing Limited.

- Fairclough, N. (2000). *New Labour, New Language?* (London, Routledge).
- Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *Quarterly Journal of Economics*, 128(1), 249-285.
- Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, 31(2), 373-407.
- Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of education policy*, 24(1), 23-37.
- Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, 44(2), 93-109.
- Gümüseli, A. İ. (2004). Ailenin katılım ve desteğinin öğrenci başarısına etkisi. [The Family Factor on the Student's Success of School]. *Özel okullar birliği bülteni*, 2(6), 14-17.
- Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, 27(1), 1-21.
- Güvendir, M. A. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarı ile ilişkisi [Student and School Characteristics' Relation to Turkish Achievement in Student Achievement Determination Exam]. *Eğitim ve Bilim, Education and Science*, 39(172), 20-26.
- Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (2000). *The application of hierarchical linear modeling to organizational research*. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 467-511).
- Hox, J. (2002). Quantitative methodology series. *Multilevel analysis techniques and applications*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Kahraman, Ü., & Çelik, K. (2017). Analysis of PISA 2012 results in terms of some variables. *Journal of Human Sciences*, 14(4), 4797-4808.
- Kaldan, E. S. (2007). *İlköğretim 3. Sınıf Öğrencilerinin Türkçe Dersinde Okuduğunu Anlama Becerilerini Etkileyen Ekonomik ve Demografik Faktörler* [Economic and demographic factors that affect the 3rd grade primary school students' reading comprehension skill in Turkish class]. Unpublished Master thesis, Gaziantep Üniversitesi.
- Kapıkıran, Ş., & Kıran, H. (1999). Ev ödevinin öğrencinin akademik başarısına etkisi. [The effect of homework on student's academic achievement]. *Pamukkale University Faculty of Education Journal*, 5(5), 54-60.
- Kotte, D., Lietz, P., & Lopez, M. M. (2005). Factors Influencing Reading Achievement in Germany and Spain: Evidence from PISA 2000. *International Education Journal*, 6(1), 113-124.
- Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. *Review of educational research*, 79(1), 464-490.
- Lietz, P., & Kotte, D. (2004). Factors influencing reading achievement in Germany and Finland: Evidence from PISA 2000. *The Seeker*, 215-228.
- Lingard, B. & Rawolle, S. (2004) Mediatizing educational policy: the journalistic field, science policy, and cross-field effects, *Journal of Education Policy*, 19(3), 361-380.
- Linnakyla, P., Malin, A., & Taube, K. (2004). Factors behind low reading literacy achievement. *Scandinavian Journal of Educational Research*, 48(3), 231-249.
- Nonoyama, Y. (2006). *A cross-national, multi-level study of family background and school resource effects on student achievement*. Unpublished doctoral dissertation. Columbia University. Retrieved 6 July, 2007 from UMI.
- Organization for Economic Cooperation and Development (OECD). (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: Author.
- OECD. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematics and science literacy*. Paris: OECD

- OECD, P. (2014). *Results: What Students Know and Can Do Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014).
- Organization for Economic Cooperation and Development (OECD). (2006). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*. Paris:
- Özdemir, C. (2017). OECD PISA Türkiye Verisi Kullanılarak Yapılan Araştırmaların Metodolojik Taraması. [A Methodological Review of Research Using OECD PISA Turkey Data] *Eğitim Bilim Toplum*, 14(56), 10-27.
- Öztürk, P. (2010). *İlköğretim II. kademe Türkçe dersi performans görevi başarı puanları ile akademik başarı ve derse yönelik tutum arasındaki ilişkinin değerlendirilmesi [The evaluation of relation between success grades of Turkish lesson performance task and academic success and attitude to lesson in primary education school level]*. Unpublished Master thesis, Karadeniz Teknik Üniversitesi, Trabzon.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage Publishing.
- Reinikainen, P. (2012). Amazing PISA results in Finnish comprehensive schools. In *Miracle of education* (pp. 3-18). Sense Publishers.
- Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, 21(5), 667-706.
- Ringarp, J., & Rothland, M. (2010). Is the grass always greener? The effect of the PISA results on education debates in Sweden and Germany. *European Educational Research Journal*, 9(3), 422-430.
- Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change*, 8(4), 349-357.
- Swap, S. M. (1993). *Developing Home-School Partnerships: From Concepts to Practice*. Teachers' College Press, Columbia University, 1234 Amsterdam Avenue, New York, NY 10027 (cloth--ISBN-0-8077-3231-1).
- Taş, U. E., Arıcı, Ö., Ozarkan, H. B., & Özgürlük, B. (2016). PISA 2015 Ulusal Raporu. [PISA 2015 National Report]. *Ankara: Milli Eğitim Bakanlığı*.
- Thomson, S., De Bortoli, L., Nicholas, M., Hillman, K., & Buckley, S. (2010). *PISA in brief: highlights from the full Australian report: challenges for Australian education: results from PISA 2009: the PISA 2009 assessment of students' reading, mathematical and scientific literacy*.
- Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. McGraw-Hill Education (UK).
- Yıldırım, K. (2012). The main factors determining the quality of education in Turkey according to the PISA 2006 data. *The Journal of Turkish Educational Sciences*, 10(2), 229-255.
- Yore, L. D., Anderson, J. O., & Chiu, M. H. (2010). Moving PISA results into the policy arena: Perspectives on knowledge transfer for future considerations and preparations. *International Journal of Science and Mathematics Education*, 8(3), 593-609.



## The Young Adults Form of the Attitude toward Women's Working Scale: Development, Preliminary Validation and Measurement Invariance

Devrim Erdem <sup>1,\*</sup>

<sup>1</sup>Department of Educational Sciences, Nigde Omer Halisdemir University, Turkey

### ARTICLE HISTORY

Received: 08 May 2019

Revised: 24 December 2019

Accepted: 06 January 2020

### KEYWORDS

Attitude,  
Women's working,  
Scale development,  
Configural invariance,  
Metric invariance

**Abstract:** The purpose of this study was to develop a scale measuring attitudes toward women's working. In line with this main purpose, two studies were conducted to develop the tool and investigate its psychometric properties in two different samples. The study 1 started with generating item pool, conducting exploratory factor analysis to identify underlying factor structure of the latent variable. In study 1 after testing the structure of measure, a brief 9-item, tri-factor scale for the assessment of attitudes toward women's working was emerged. The study 2 utilized a different sample. In study 2, it was aimed to examine model fit, test measurement invariance across gender and investigate reliability. Validity and reliability of the scale indicated that the attitude toward women's working scale (ATWWS) had satisfactory psychometric properties. In study 2, configural and metric invariances of the ATWWS were supported for females and males.

## 1. INTRODUCTION

One of the important criteria for a society's advancement is that individuals can participate in a fair labor market and have equal opportunity for acquisition of welfare. Given that women are nearly half of any given human society, examination of a society's human capital and the efficiency which it is being used cannot be done without taking women into account. Women's participation into public life as well as into work place is of paramount importance in today's societies. On the other hand, the existing gender inequalities pose extra challenges for those women who do participate in the labor market (Forsythe, Korzeniewicz, & Durrant, 2000; Himmelweit, 2002).

One could easily claim that women have always partaken in production throughout all human history. However, when compared to men, women's attendance in the public life and in paying jobs has not been to a satisfactory degree all along (Kakıcı, Emeç, & Üçdoğruk, 2007). In the case of Turkey, while women's employment in both industrial and service sectors was 3.86% in 1955, it was 40.9% in 2000 (Turkish Statistical Institute, [TUIK], 1990, 2000). Women's employment rate in the total labor market was 23.3% in 2004, it reached 30.8% in 2013 (Turkish Ministry of Labor, Social Services and Family, [MLSSF], 2014). Looking at women's (age 15 and up) employment according to their educational level shows intriguing results: Their

---

CONTACT: Devrim Erdem ✉ [erdem\\_devrim@yahoo.com](mailto:erdem_devrim@yahoo.com) 📧 Department of Educational Sciences, Nigde Omer Halisdemir University, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

employment rates increase with their levels of education. Women with illiteracy had a rate of 16% while the rate of those with university degrees was 71.3% (TUIK, 2016). These statistics show that women's participation in the work place has been increasing but it is still far from being equal to those of men. Indeed, employment rate among men is 65.1% while women's rate is 28% (TUIK, 2016). In other words, women's employment is not even at the half level of that of men. A closer look at the nature of males' and females' employment shows even more striking discrepancies. Women work in part time job three times more than men (6.5% men; 19.1% women). Women's employment in mid or high-level administrative positions is only 16.7% (TUIK, 2016). In other words, males occupy incomparably more decision-making positions than females.

When women's employment rates in Turkey are compared to those of women in Europe or in the OECD countries, the results are not quite promising. These rates are over 60% in Europe and in member countries of the OECD while they are under 30% in Turkey (OECD, 2016). In its 2018 report on the Global Gender Gap Report (GGGR, 2018), the World Economic Forum indicated that Turkey ranked 130<sup>th</sup> among 149 countries with respect to women's employment. The same report illustrates that Turkey ranked 131<sup>st</sup> in terms of women's participation in the economy and of equal opportunities. Further, it ranked 106<sup>th</sup> in terms of women's education and 113<sup>rd</sup> regarding women's political participation. The same report shows that a significant increase occurred in women's employment rate, during the current decade while their earning was 51% of that of men. Statistics show a global decrease in women's employment and attribute this decrease to increased utilization of technology. The report finally indicates that women's participation in areas of science, technology and mathematics is still far from being equal to that of men.

Considering cultural values and norms in Turkey, these statistics are not surprising. A host of aspects of the culture in Turkey reflect a highly sexist and gendered view (Kuzgun & Sevim, 2004). Thus, from how parents raise female and male children, to how motherhood is idealized to a lack of legal protection of women, there are numerous aspects of the culture, its legal system, family functioning that limit women's roles merely within a traditional patriarchal domain. Gender refers to meanings and expectations a society or culture attributes to its males and female members (Lips, 2001). In the Turkish culture, while females are expected to be patient, sensitive, caring, passive and dependent, males are expected to be assertive, dominant, strong, independent, competitive and determined (Dökmen, 2004; Heilman, 2001; Özkaplan, 2013). Both families and the general society provide messages defining gender roles by both overt and covert means (Ersoy, 2009; Tan, 2000).

Women embracing the roles of wives and mothers (Özcatal, 2011), women who have economic resources preferring not to work (Koray, 1992), women's work depending upon the permission of husbands (Özcatal, 2011) are all indicative of the fact that the culture still heavily promotes traditional-patriarchal gender roles. The home and household works seem to be essential in defining women's identity (Bora, 2011). In other words, the cultural gender stereotypes attributing roles of wives and mothers as women's primary duties seem to still highly prevalent (Aktaş, 2013, Bingöl, 2014; Nergiz & Yemen, 2011). Such traditional roles and expectations are likely to impede with women's likelihood to partake in the work force and limiting their aspirations (Aktaş, 2013; Dökmen, 2004).

Gender roles have significant impact on individuals' careers as well as institutions of any given society. A male-oriented workforce and organizational culture also constitutes challenges for women and impact their career choices (Britton, 2000; Özar, 2005). Indeed, teaching seems to be number one preferred occupation that women choose both for themselves and for their daughters (Özcatal, 2011). Likewise, they prefer occupational areas such as nursing or counseling that extend their caring roles as wives and mothers to their careers.

The relevant literature on women's participation into the workforce shows that there are numerous factors such as motherhood and familial responsibilities, caring for the elderly (Dayıoğlu & Kırdar, 2010; Palaz, 2015; Yamak, Abdioğlu & Mert, 2012) age, marital status, educational level, place of residence, number of children, spouse's level of education seem to significantly impact their involvement in paying jobs (Akın, 2002; Bölüköğlu, 2018; Gürler & Üçdoruk, 2007; Kılıç & Öztürk, 2014; Kırıl & Karlılar, 2017; Şengül & Kırıl, 2006; Yıldırım & Doğrul, 2008). Moreover, high unemployment rates of women are major discouraging factors affecting women's participation in the labor market (Tansel 2002; Kızılırmak, 2005). Thus, many developed and developing countries have implemented state policies encouraging women's employment (Kakıcı, Emeç & Üçdoğruk, 2007). Turkey too, in its highest document for state policies, 10<sup>th</sup> Development Program (2014-2018) specified concrete measures to take to improve women's participation in workforce. Indeed, part of the development program has been the Priority Transformation Program that identifies obstacles to women's employment and specific measures by which they can be reduced (MLSSF, 2014). Gender equality and empowerment of women are essential among the most important priorities declared by the United Nations Millennium Development Goals (United Nations, [UN], 2012). This requires significant improvements in women's rates of participation in the workforce.

Even in the development plans of those countries that have relatively advanced in development, there are interventions geared toward improvements in women's employment. Thus, the need for advancement in women's participation in the work force is far from being specific to Turkey and is rather a global issue. Traditionally, modern society allocated different social roles to men and women in the division of working life (Alwin, Braun & Scott, 1992). Women, in particular those with young children, work part-time, and they still provide free care (Scott & Clery, 2013). The fact that women have more responsibility in childcare leads to a decrease in women's working hours, while it on the contrary leads to increase in men's working hours (Kaufman & Uhlenberg, 2000). Time spent on domestic labor adversely affects women's earnings and career performance (Hersch & Stratton, 2002). Regardless of national boundaries men are not high likely to approve of women working when there are preschool or young children at home, and clearly, it is considered more reasonable for women to take care of children (Alwin, Braun, & Scott, 1992). Despite progress made around the world, women are still concentrated in gender-segregated jobs such as teaching, nursing, clerical, sales and service occupations (Ferraro, 2010). In Canada women still occupy the majority of part-time low-income jobs; of all part-time workers in 2009, nearly seventy percent were women and they were taking upon the added burden of childcare (Ferraro, 2010). In the UK, even if women are beginning to represent a growing proportion of the working population, this does not indicate that they are beginning to breakdown gender segregation within certain professions (i.e., women are overrepresented in office and secretarial positions) (Agapiou, 2002). Again, in the UK, 41% of women in employment were working part-time compared to 13% of men; and because the per hour earning of part-time workers is less than full-timer workers, the gender pay gap was greater for all employees (Powell, 2019). In the US, although women participation in labor pool has increased, women still undertake 65% to 80% of chores in home (Bianchi, Milkie, Sayer, & Robinson, 2000; Coltrane, 2000). Furthermore, 43.2% of women in employment were working in gender-segregated jobs such as health care, non-governmental education, leisure, janitor, secretary, accountant and other services (U.S. Bureau of Labor Statistics, 2019).

Despite changes in policies, improved legislation and efforts toward improving women's participation in the labor market, patriarchal values seem to significantly influence women's decision making in choosing the kinds of workplaces, occupations and jobs (Köseoğlu, 2017). Likewise, discrimination and inequality pose extra challenges in women's education as well as work lives. Norms, perceptions and prejudice on gender stem from past history, economic and

societal circumstances, political regimes, religion and cultural values (Kırkpınar, 2001). In this way, individuals are exposed to these norms and values from the onset of their lives. Corrigan and Konrad (2007) pointed out that early gender role attitudes of women affect women's later career and earnings. Since both married women and men support traditional gender roles (Gubernskaya, 2010), the thoughts and behavior patterns of children regarding gender roles have been shaped from an early age. As a result, individuals' schemas, perceptions and attitudes regarding males' and females' work are shaped starting from early developmental stages and become firmer and more resistant to change as individuals proceed in their life span development.

An attitude is considered as an individual's general and enduring evaluation of an object or concepts. These evaluations can be about almost anything, including persons, social groups, physical objects, behaviors, and abstract concepts (Fabrigar, MacDonald, & Wegener, 2005). Allport (1935) described attitude as "a mental and neural state of readiness, organized through experience" (p.810). According to Allport (1935), attitudes exert a directive or dynamic effect on an individual's reaction to all the objects and situations to which it relates. In this definition the expression of "a mental and neural state of readiness" particularly highlights the basis of attitudes. A remarkable number of models have been developed to identify attitude formation and change, but most of these models focus on cognitive processes (Maio, Haddock, & Verplanken, 2018). The cognitive component of attitudes indicates beliefs and thoughts related to an object. Indeed, the amount of knowledge on which the attitude is based affects the function of the attitude. The content and the breadth of knowledge toward the object are the associative links making up the attitude (Fabrigar & Wegener, 2010). Attitudes can influence individuals' learning (Brewer, Kramer, & O'Brien, 2009; Perkins, Adams, Pollock, Finkelstein, & Wieman, 2005), perception (Ajzen, 1989; Hinner, 2019), reasoning and thinking (Yinger, 1980). Attitudes also influence individuals' interpretation of information and memory processes (Blackton, 1986; Fabrigar & Wegener, 2010). Thus, in the current study, while forming the items of the scale a great deal of emphasis was placed on the cognitive aspect of attitudes toward women's working.

There have been a number of studies with Turkish samples exploring attitudes toward women's work. Most of these studies (Çiçek & Çopur, 2018; Koca, Arslan, & Aşçı, 2011) have used the scale developed by Kuzgun and Sevim (2004). The scale was developed with a sample of 112 adults (Kuzgun & Sevim, 2004). The authors began their scale development study with a form consisting of 27 items. Then, their exploratory factor analysis resulted in a scale made of five factors. Then, the authors eliminated 12 items. The remaining 15 item – form gathered in one factor with an internal consistency coefficient of .92. Another widely used scale was developed by Köseoğlu (2017) who attempted to measure male university students' attitudes toward women's work. Her initial scale made of 30 items was given to a sample of 251 male students. Exploratory factor analysis yielded in a four-factor structure. After eliminating 9 items from the scale, with the remaining 21 items, Köseoğlu (2017) obtained a single-factor. The last form explained 57.88% of the total variance and had an internal consistency coefficient of .93. There were some limitations of these two scale development studies. One, they worked with relatively small samples. Two, in their initial forms, exploratory factor analysis with both scales resulted in multi-factorial structures, but then the scales were transformed into a single-factor structure by eliminating items from the scale.

However, if an attitude toward women's working is a multidimensional construct, there will be some drawbacks to merging its dimensions in a single component. In such a case, the conceptualization of the construct will be insufficient. Besides, there will be a deficient understanding of the construct's antecedents and consequences (Fredricks, Blumenfeld, & Paris, 2004). In addition, if the measures of attitudes operate differently between the comparison

groups, the item or/and groups of items that cause this difference must be identified. If existing differences are not taken into account in the measurement process, comparisons of levels of attitudes or its effects across groups are invalid. Therefore, it is essential to provide evidence that the given construct works similarly between groups before the scores obtained from the relevant construct are used for comparison purposes. Furthermore, exploratory factor analysis (EFA) is generally recognized as initial phases of scale development. However, further statistical techniques should be applied to confirm or disprove the results obtained in the exploratory phase (Rentz, Shepherd, Tashchian, Dabholkar, & Ladd, 2002). Due to the methodological and conceptual limitations of the existing scales, it was deemed appropriate to develop a new instrument. Thus, the scale is intended to contribute to the related literature.

Meta-analyses have indicated that there is a significant and positive correlation among the different dimensions of attitudes and these attitudes predict behavior (Glasman & Albarracín, 2006). Baron and Bryne (2000) also stated that attitudes are an important factor that should be investigated because attitudes have a strong effect on thought and have an important effect on individual behaviors (as cited in Noor & Saad, 2016). Individuals receive messages on gender stereotypes and attitudes at early ages from various sources such as peers, the media, family and school in both overt and covert ways. Such differential approach leads to sex differences in activities persons partake, in areas they pursue to explore their abilities and even in their career aspirations. Therefore, identifying individuals' attitudes toward women at as early ages as possible will make it more likely for interventions geared toward changing negative attitudes. Such change will not only impact the existing generations but will perhaps be passed on future generations. Thus, the current study aimed at developing a scale for assessing young adults cognitive attitudes toward women's work. In line with this main purpose, two studies were conducted to develop a tool and investigate its psychometric properties on separate samples. The study 1 started with generates item pool and then proceeded with EFA to reveal underlying factor structure of the latent variable. The study 2 utilized a different sample and involved use of confirmatory factor analysis (CFA) and intended to test measurement invariance according to gender. Initial reliability was also investigated.

## **Study 1: Scale Development, Exploratory Factor Analysis**

### **2. METHOD**

In the current study was aimed to develop an item pool and to search out the underlying structure of the items.

#### **2.1. Participants**

A cross-sectional sample of 364 students from a state university located in central Anatolia in Turkey was involved for Study 1. This was a convenience sample consisting of 201 (55.2%) females and 163 (44.8%) males. Participants' ages ranged from 19 to 24 ( $M_{age}=21.43$ ,  $SD=1.03$ ). Twenty-five percent of the respondents stated that they studied in the faculty of education, 22% in faculty of sciences and literature, 19.5% in faculty of economics and administrative sciences, 16.5% in faculty of engineering, 8.8% in school of physical education and sports and 8.2% in school of health. Fifty-eight percent of the respondents' mothers and 45% of the respondents' fathers had only primary school education. Thirty-four percent of the participants stated that they come from the Central Anatolia region and 44% of the Mediterranean region of Turkey.

#### **2.2. Instrument: Scale development - Item pool generation and expert review**

At the outset of the current study an in-depth review of literature was performed to specify the conceptual boundaries and dimensions of the construct. Then, an initial pool of items was generated based on a literature review of existing measures assessing attitudes toward women's

working. At the same time 13 university students were asked to write an essay in which they expressed their thoughts about the women's work. Based on the literature and these essays, the author wrote 32 draft items. Instead of carefully selecting, if all the items are included in the form for the pilot study this will lead to response contamination (Erkuş, 2012). This recommendation by Erkuş (2012) was kept in mind; in other words, special care was given in selecting items most likely to capture the trait. Therefore, at the stage of item writing process, redundancy of items was not tolerated. In order to ensure the face-validity and the content-validity, two independent sociologists reviewed these items of the draft scale. Then, face-to-face interviews were conducted with four individuals inquiring their opinions about the items. Ambiguous items, items with similar meaning and irrelevant items were eliminated. After the assessment, eight items were removed and the number of items in the scale was reduced to 24 according to experts' opinions. Subsequently, a Turkish language specialist reviewed the remaining 24 items and according to her feedback changes were made in some items. Participants' level of agreement on each item was determined with a five-point Likert-type. The responses vary from *strongly disagree* (1) to *strongly agree* (5).

### 2.3. Procedure

In the present study, all respondents were informed about the aim of the study and were told they were free to leave the study at any time. Then, the scale was distributed to volunteers. Application was group administered during one class session. They received no payment or extra credit for their participation. Application took approximately 20 min.

### 2.4. Data Analysis

In order to explore the dimensions and purify the item pool of the ATWWS, exploratory factor analysis (EFA) using the principal axis factoring (PAC) extraction was performed with SPSS 22. If needed in proceeding stages an oblique rotation would be preferred. An oblique rotation allows factor to correlate (Worthington & Whittaker, 2006) and factor inter-correlations are the norm in social sciences (Costello & Osborne, 2005).

## 3. FINDINGS

### 3.1. Data Screening

Prior to conducting the analysis, data were subjected to monitor for missing values and outliers. Six missing values were detected. The cases having missing values were removed from the data set. Outliers were not detected in the data.

### 3.2. Scale Refinement

Exploratory factor analysis (EFA) was performed on attitude toward women's working scale (ATWWS) items. The Kaiser-Meyer-Olkin (KMO) analysis was carried out to examine sample size criteria. Since KMO index was .87 the sample size was found to be adequate. Factorability of the scores was assessed based on the *Barlett's test of sphericity* test that was significant ( $\chi^2_{(36)} = 1443$ ;  $p = 0.00$ ). Based on these findings it was concluded that factor analysis could be performed.

The underlying structure of the 24-item ATWWS scale was evaluated using the principal axis factoring (PAC) without rotation at first. When the eigenvalues were examined, there were six factors greater than 1. These initial eigenvalues were 7.1, 3.4, 2.3, 1.4, 1.2 and 1.1 respectively. The variances explained by these six factors were as follows: 29.6%, 14.4%, 9.6%, 5.9%, 5.2% and 4.8%. On the other hand, as shown in [Figure 1](#) the scree plot test proposed a three-factor solution. According to the scree plot it was clear that the slope after point third changes to a more straight line. In progress, many analyzes including three, four, five and six factor solutions were performed. It is desirable to maintain sufficient factor for adequate fit, "but not so many

that parsimony is lost” (Tabachnick & Fidell, 2001, p.620). Therefore, when deciding factor retention, item loadings, eigenvalue, scree plot test, explained variances but especially the interpretability of the items under the factors were taken into consideration. Taken together, in this case the number of optimal factors was considered to be three. Once it was decided to number of factor, the EFA with oblique rotation for three-factor restriction was performed. After conducting EFA distinct three factors were emerged.

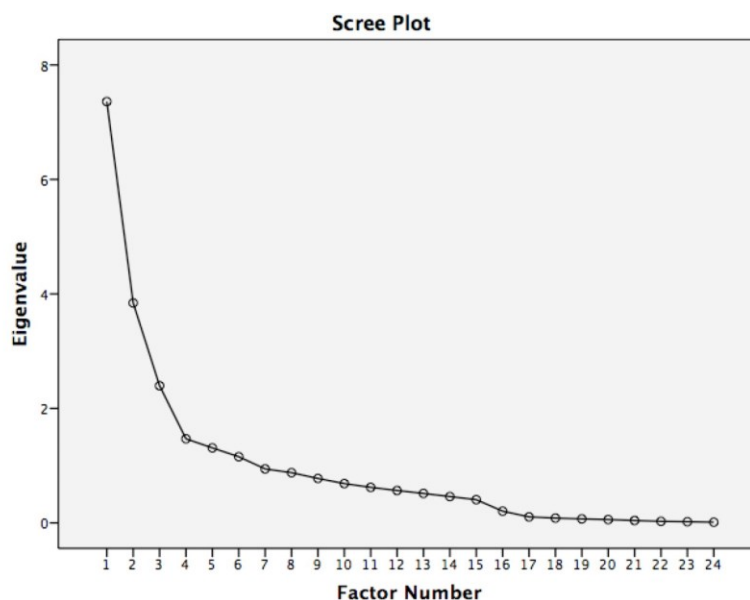


Figure 1. Scree Plot

In order to purify the scale, items with low communalities (less than .40), low factor loadings (less than .40) and/or cross-loadings (less than .20) were eliminated. This resulted in removing 15 items from the scale. Thus, nine items remained on the scale. The factor loadings and communalities of the scale were presented in Table 1.

Table 1. Principle axis factoring with oblique rotation Pattern matrix for the final ATWWS items

Item	Factor loadings			$h^2$	Item-Total $r$
	F1	F2	F3		
Due to their biological cycles (menstruation, birth, menopause etc.) women cannot be productive in the work force.	<b>.80</b>	.32	.29	.59	.63
Women are not resilient to long and hard work conditions.	<b>.76</b>	.36	.50	.62	.63
Woman cannot tolerate pressures at work as much as men.	<b>.63</b>	.38	.51	.57	.64
Women should only work at jobs that suitable for women.	.41	<b>.88</b>	.40	.49	.57
Since men are breadwinners they should be given priority in hiring.	.24	<b>.82</b>	.25	.52	.70
Domestic work is more suitable for women.	.39	<b>.61</b>	.41	.67	.66
Presence of women at work places will decrease overall productivity.	.38	.33	<b>.94</b>	.66	.60
Women’s use of their femininity for personal gain will cause unfair conditions at work.	.34	.32	<b>.73</b>	.74	.69
Men should be preferred for administrative position.	.46	.38	<b>.59</b>	.77	.69

The three-factor explained 63.9% of the total variance. The first factor consisted of three items, explained 41.7% of the total variance. This factor was called as “Gender Discrimination”. The explained variance by the second factor was 12.1% and was called “Patriarchal Values”. The third factor explained 10.1% of the total variance and labeled “Work Environment”. The factor

loadings of the items ranged between .59 and .94. These findings provided evidence that the tri-factor scale had satisfactory construct validity.

The item-total test correlations took values between .57 and .70 (see Table 1). Item-total correlations of .30 or higher are evidence of the items' validity (Nunnally & Bernstein, 1994). This also indicates that items in the scale measure the properties that they need to measure.

#### 4. CONCLUSION

In summary, 32 draft items were created first. After qualitative evaluation of these items, it was decided to keep 24 items in the form. The EFA reduced the initial 24 items into 9 items formed three factors. Thus, Study 1 provided preliminary evidence for the structure and coherence of a measure of attitudes toward women's working.

#### Study 2: Further Construct Validity, Reliability and Measurement Invariance across Gender

In Study 1, the ATWWS was demonstrated to have distinguishable factor structure and sufficient convergent validity. To prove further evidence of its validity the factor structure extracted from previous study (in study 1) tested on a new sample. To confirm factor structure confirmatory factor analysis was performed. However, providing model fit does not guarantee that the scores obtained from the scale are comparable between the groups (Messick, 1995). Therefore, measurement invariance test was conducted across gender in Study 2.

#### 5. METHOD

##### 5.1. Participants

Participating in this study 2 were 600 undergraduate students. The convenience sample included 308 (51.33%) females and 292 (48.67%) males. Participants' age ranged between 19 and 26 years ( $M_{age}=21.67$ ,  $SD=1.43$ ).

##### 5.2. Instruments

In the first study a 9-item scale was yielded. This scale was named as the attitude toward women's working scale (ATWWS). The ATWWS was applied to participants in the study 2.

##### 5.3. Data Analysis

Confirmatory factor analysis (CFA) was performed to investigate whether the factor structure obtained in the previous study fits to the data obtained from another sample. Confirmatory factor analysis is a psychometric assessment that permits comparing *a priori* factor structure based on multiple fit assessment procedures (Morin, Arens, & Marsh, 2016). In the literature, it has been recommended that CFI, RMSEA, TLI and GFI should be preferred to evaluate model data-fit in CFA (Hu & Bentler, 1999; Weston, Gore, Chan, & Catalano, 2008). CFI, TLI and GFI values above .90 are acceptable, although values above .95 are more preferred (Kline, 2011). RMSEA values up to .06 (Brown & Cudeck, 1992; Yuan, 2005) and SRMR values up to .08 (Brown, 2006) are reasonably good fit. The chi-square test for model fit is expected to be insignificant, however, a significant value may not necessarily mean that there is poor model fit. Because of the large sample size, it is often inflated, so  $\chi^2/df$  less than 3 (or even 5) considered acceptable for good model fit. In order to assess the reliability of ATWWS' subscales Cronbach's alpha coefficients were calculated.

##### 5.3.1. Measurement Invariance tests.

Multiple-group confirmatory factor analysis (MGCFAs) was performed to examine gender invariance. In this procedure, the equality of model parameters is tested using a nested hierarchy model comparison based on the chi-square tests (Brown, 2006; Byrne, 2004). A more restrictive



hypothesis is proposed at each stage, thereby increasing the evidence for measurement invariance is provided. First, two CFAs were conducted for male and female participants separately. Next, procedure involved observing for significant changes in chi-square test values after constraining namely configural, metric, scalar, and strict invariance. If the chi-square difference across the models is not statistically significant then invariance is achieved (Dimitrov, 2010). This procedure, referred to as the forward approach because the analysis launches with the baseline model and goes towards to the more constrained model. In addition to chi-square difference test, a change of in CFI (e.g.  $\Delta CFI = CFI_{M1} - CFI_{M0}$ ) value is assessed for the nested models.  $\Delta CFI < -0.01$  would show a deficiency of invariance (Dimitrov, 2010). That is, a positive  $\Delta CFI$  indicates fit improvement; this result points out that invariance has been achieved (Dimitrov, 2010). Reporting  $\Delta CFI$ , along with  $\Delta \chi^2$ , assessing a change in RMSEA is also proposed.  $\Delta RMSEA \geq 0.015$  would indicate lack of invariance (problematic values) (Chen, 2007). All tests were carried out using maximum likelihood estimation in LISREL.

## 6. FINDINGS

### 6.1. Data Screening

Prior to conducting the analysis, data were examined for missing values and outliers. Eleven missing values and eight outliers were dropped from the data set. The analyzes were continued with 581 (291 female and 290 male) data.

### 6.2. Confirmatory Factor Analysis

The nine items selected from the exploratory phase were used in CFA to verify the tri-factor structure of the ATWWS. The measurement model summarized in Table 2 was tested to verify the relationship between observable variables and latent constructs. The  $\chi^2$ -to-*df* ratio was in the acceptable range ( $\chi^2_{(24)}=47.39$ ,  $p=.003$ ,  $\chi^2/df = 1.97$ ), and all fit indices were highly satisfactory (CFI=.99, TLI=.98, GFI=.98; NFI=.98, AGFI=.95, SRMR=.033, RMSEA=.049) for a first-order CFA. Then, a second-order model was evaluated. The  $\chi^2$ -to-*df* ratio was fairly well ( $\chi^2_{(24)}=47.13$ ,  $p=.003$ ,  $\chi^2/df = 1.96$ ), and a quite enough fit was obtained (CFI=.98, TLI=.97, GFI=.98; NFI=.98, AGFI=.95, SRMR=.028, RMSEA=.048) for the second-order CFA. Findings demonstrated that the second-order model provided also a good fit to the data.

**Table 2.** The CFA measurement model for the tri-factor ATWWS

Latent variables	Observed variables	Coefficients	Error Terms
Gender Discrimination	Due to their biological cycles (menstruation, pregnancy, childbirth, menapous etc.) women cannot be productive in the work force.	.62	.62
	Women are not resilient to long and hard work conditions.	.61	.63
	Woman cannot tolerate pressures at work as much as men.	.60	.64
Patriarchal Values	Women should only work at jobs that suitable for women.	.83	.31
	Since men are breadwinners they should be given priority in hiring.	.69	.52
	Domestic work is more suitable for women.	.62	.62
Work Environment	Presence of women at work places will decrease overall productivity.	.63	.60
	Women's use of their femininity for personal gain will cause unfair conditions at work.	.78	.39
	Men should be preferred for administrative position.	.71	.50

### 6.3. Measurement Invariance across Gender

At this stage firstly confirmatory factor analyzes were conducted for male and female participants separately. Both the first-order and the second-order CFAs for ATWWS were showed one by one for females and males in Table 3. The first-order and the second-order three-factor solutions yielded superior fit indices for both samples, with the model fitting the females slightly better.

**Table 3.** Fit indices of the 9-item three-factor ATWWS across Gender

Group	CFA Model	$\chi^2$	<i>df</i>	<i>p</i>	$\chi^2/df$	CFI	TLI	GFI	RMSEA
Females	1 <sup>st</sup> order	26.53	24	.327	1.10	.99	.98	.98	.019
	2 <sup>nd</sup> order	22.47	24	.551	.936	1.0	1.0	.98	.000
Males	1 <sup>st</sup> order	37.80	24	.036	1.57	.98	.98	.97	.044
	2 <sup>nd</sup> order	39.04	24	.027	1.63	.98	.97	.97	.047

Multiple-group analyzes for each group were performed to establish baseline model. Subsequent analyzes were conducted by adding each more constraint to the next model. As shown in Table 4, configural invariance had acceptable fit to the data. This indicated that the correlated three-factor structure held across males and females. Since the configural invariance was achieved, then the factor loadings were constrained. Metric invariance model appeared fit to the data well, and also better compared to the configural model ( $\Delta\chi^2=16.61$ ,  $\Delta df=6$ ,  $p=.011>.01$ ). Chi-square difference value was insignificant;  $\Delta CFI$  (.00) more than  $-.01$  and  $\Delta RMSEA$  (.002) less than .015 indicated that model had metric invariance across gender. Scalar model fitted adequately to the data. When two models compered it appeared that the scalar model indicated worse fit than the metric model ( $\Delta\chi^2=70.69$ ,  $\Delta df=14$ ,  $p=.000$ ). Besides,  $\Delta CFI=-.04$  less than  $-.01$  pointed to evidence for the lack of scalar invariance. Likewise, even though the strict invariance model yielded adequate fit to the data, but could not achieve better fit according to the scalar model ( $\Delta\chi^2=66.71$ ,  $\Delta df=9$ ,  $p=.000$ ). Compared to scalar model, strict model resulted in a change in CFI (-.02) less than  $-.01$ , thus evidence for strict invariance was not attained.

**Table 4.** Tests of Measurement Invariance

Model	$\chi^2$	<i>df</i>	<i>p</i>	$\chi^2/df$	CFI	TLI	RMSEA	$\Delta\chi^2$	$\Delta df$	<i>p</i>	$\Delta CFI$
Configural	80.57	48	.002	1.67	.98	.97	.042	-	-	-	-
Metric	97.18	54	.000	1.79	.98	.97	.046	16.61	6	.011	0.0
Scalar	167.87	60	.000	2.79	.94	.93	.069	70.69	14	.000	-.04
Strict	234.58	69	.000	3.39	.92	.91	.080	66.71	9	.000	-.02

### 6.4. Convergent Validity

In order to evaluate convergent validity of the ATWWS-9, Pearson correlations between ATWWS-9 total score and its subscales were computed. The correlations between subscales were presented in Table 5. Pearson correlations between the factors were significant and positive. Each subscale had moderate correlations with others. Moderate correlations indicate that each subscale is related to the others, but still sufficiently different. The Pearson correlations between each factor and the total scale score were found positive, strong, and significant (see Table 5).

**Table 5.** Correlations between ATWWS subscales

ATWWS	F1	F2	F3	Total
F1	–	.580**	.481**	.827**
F2		–	.467**	.812**
F3			–	.814**

\*\*Correlation is significant at the 0.01 level.

### 6.5. Internal Consistency

The *Cronbach's alpha* values found as .70, .72 and .74, for *Gender Discrimination*, *Patriarchal Values* and *Work Environment* respectively. The overall *Cronbach's alpha* for the scale was calculated as .81. Seventy percent or higher internal consistency coefficient is considered to be sufficient for the reliability.

## 7. GENERAL DISCUSSION

The main aim of this current multi-study investigation was to develop a scale measuring attitudes of young adults toward women's working. The attitude toward women's working scale was developed by the researcher. It was a tri-factor scale consisting nine Likert-type items. Participants rated each item on a scale from 1 (strongly disagree) to 5 (strongly disagree). In Study 1 the scale was administered to 364 young adults. The initial construct validity of the scale was determined by EFA. All items displayed moderate to high loadings on their respective factors, in a sense that all items contribute similarly to the latent variables. The 9-item tri-factor scale accounted for 64.9% of the total variance.

In Study 2, both the first and the second-order CFA's were performed to investigate whether the data support the proposed model of the scale on a different sample. Findings pointed out strong support for both the first-order and the second-order model consistent with the exploratory factor analysis in the whole group and in the gender groups. These results shored up the theoretical conceptualization of attitude toward women working as a sole construct comprising of the three related but independent dimensions. In short, the fit index values of the structural model confirmed the further construct validity of the scale. The *Cronbach's alphas* values were .70 for *Gender Discrimination*, .72 for *Patriarchal Values* and .74 for *Work Environment*. These reliabilities demonstrated sufficient internal consistency considering the few number of items included in each sub-scale. In addition, Pearson correlations between the factors were calculated. The three subscales demonstrated moderate, positive and significant correlations among each other. This means that although each of the three factors seems to share a common essence, each represents a separate dimension. Thus, three subscales demonstrated modest evidence of convergent validity.

Then, gender invariance of the latent construct was evaluated with MGCFA. The MGCFA findings indicated that configural and metric invariance is completely achieved for the three-factor structure of ATWWS across gender. Configural invariance means that the scale had the same number of factors in both females and males. Obtaining the configural invariance also shows that the items under each factor are the same across the groups. If the factor structures are the same between both groups, this shows that male and female participants use a similar conceptual domain (Riordan & Vandenberg, 1994). Providing metric invariance implies that the equality of factor loadings is accepted between gender. Establishing the invariance of factor loadings means that participants calibrate the intervals used on the measurement scale in similar ways (Riordan & Vandenberg, 1994). In other words, the intercourses between the latent factor and external variables can be compared among gender because a one-unit change in females would be equal to one-unit change in the males (Dimitrov, 2010).

Although scalar invariance was not met, the constraints resulted in a slightly decrease but still

acceptable model fit. Failure to support scalar invariance means item intercepts may be different. Since women are traditionally thought to have different attitudes from men (Dex, 1988), it is not unexpected that women and men have different reference points in regarding the construct examined. Chen (2008) stated that intercept lack of invariance could take place due to social norms. Where dominant social and cultural norms exist (such as Turkey), gender differences in attitudes towards women's participation in social and economic life are expected (Koca, Arslan, & Aşçı, 2011). On the other hand, although having the same factor mean, the fact that a particular group tends to react more strongly to an item can lead to scalar noninvariance (Chen, 2008).

In literature, scalar invariance was discussed less frequently because location parameters (intercepts) are often treated as being arbitrary and sample specific (Vandenberg & Lance, 2000). Lubke and Muthen (2004, p.516) stated “Threshold differences between groups indicate that groups use a given Likert scale in a group-specific way and are a violation of MI (Millsap & Tein, 2003), whereas threshold differences between the observed indicators of a factor do not violate MI” and they added “The MI model may be rejected because threshold differences between observed indicators can lead to a distorted factor structure or because indexes of goodness of fit based on the assumption of normally distributed data do not work properly” (p. 516). They also conclude such a case “would lead a researcher to believe that MI is violated when in fact it is not” (Lubke & Muthen, 2004, p.516). In sum, because of scalar invariance could not be achieved, it would be concluded that differences in the intercepts across the gender could exist. Since the main purpose of this study was not determined to make group comparisons and the proposed modifications on items did not improve the model fit, the investigation was not continued. However, if future research is planning to be compared in gender groups, scalar invariance should be examined. According to findings, strict invariance was not met. The lack of strict invariance however does not indicate that the scale is inconvenient for utilization among the groups, as the critical prerequisite for cross-group comparisons is metric and scalar invariance (Cheung & Rensvold, 2002).

## 8. GENERAL CONCLUSION

In sum, a brief 9-item, tri-factor scale for the assessment of attitudes toward women's working is developed. This scale reflects the multifaceted nature of the latent construct; with a factor structure revealed through EFA and verified conducting CFA. According to results obtained in the second level CFA, it is possible to state a total score can be obtained regarding the attitudes towards women's working. However, depending on the purpose of the prospective studies, the scores obtained from the subscales can also be used separately. The responses collected by a five-point Likert-type scale ranging from *strongly disagree (1) to strongly agree (5)*. Therefore, the high score obtained from the sub-scales and overall of the scale shows that negative attitudes towards women's working are high. Based on the findings, it can be stated that the ATWWS has satisfactory psychometric features. This study also supports the use of the ATWWS in its current configural and metric invariance for females and males. In other words, the scores of males and females obtained from the scale can be compared in terms of factor form and factor loadings.

Since the participants of these studies were recruited with convenience sampling, this procedure may limit the generalizability of the findings. It is recommended to use probability based sampling methods (such as simple random or stratified random sampling) for future research. Although successive studies were conducted in two different samples in this current study, instrument validation is an ongoing process and future psychometric studies are needed to further investigate the psychometric properties of the ATWWS and improve its generalizability. Furthermore, the results are limited to young Turkish adults due to the nature of the study from which the data were obtained. Studies for individuals in different developmental stages and in

larger samples may be able to provide further robust validation. Even though the present evidence revealed that the ATWWS is a psychometrically strong instrument, further investigation is necessary to warrant its use over time.

Measurement invariance was conducted only for gender. Future research on the invariance of the construct across age, parent education level and/or socio-economic status would be concerning. Because scalar invariance is not provided, it is recommended that researchers who want to compare scale scores on gender groups should be cautious in making interpretations.

In spite of the limitations, the current study has some implications. First, the ATWWS is a short and easy-to-administer self-report measure. Second, multidimensional nature of the scale allows researchers to make more clear interpretations of the test scores as well as the construct. Finally, assessment instruments having strong psychometric properties are critical for advancing social research. Thus, this study helps refine the understanding of conceptualization of attitudes toward women's working in the labor market.

This study also has educational implications. Determining the attitudes of young generations toward women's working is crucial in shedding light for the efforts geared toward facilitating positive attitudes. Determining the level of attitudes toward women's working will contribute to the awareness on this issue for all actors in economic and social life as well as for educators and policy makers. This kind of awareness can help to increase initiatives to improve social justice. Educators also play vital roles in developing and transforming attitudes. Therefore, developing attitudes toward women's working through education will make society more accessible to a stronger and fair labor distribution.

#### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

#### **ORCID**

Devrim Erdem  <https://orcid.org/0000-0003-1810-2454>

#### **9. REFERENCES**

- Agapiou, A. (2002). Perceptions of gender roles and attitudes toward work among male and female operatives in the Scottish construction industry. *Construction Management & Economics*, 20(8), 697-705. DOI: [10.1080/0144619021000024989](https://doi.org/10.1080/0144619021000024989)
- Ajzen, I. (1989). Attitude Structure and Behavior, In A. R. Pratkanis, S. J. Breckler, and A. G. Greenwald (Eds.), *Attitude Structure and Function*, Lawrence Erlbaum Associates, Hillsdale, NJ, 241-274.
- Akın, F. (2002). Kadınların işgücüne katılımı ve işteki durum tercihinin Nested Logit Model ile Analizi [Women's labor force participation and job status preference analysis with Nested Logit Model]. *METU/ERC International Conference in Economics VI*, Ankara, Turkey, September 2002.
- Aktaş, G. (2013). Feminist söylemler bağlamında kadın kimliği: Erkek egemen bir toplumda kadın olmak [Female identity in the context of feminist discourses: Being a woman in a male-dominated society]. *Journal of Faculty of Letters*, 30(1), 53-72.
- Allport, G.W. (1935). *Attitudes*. In C. Murchison (Ed) *Handbook of Social Psychology*, Worcester, Mass: Clark University Press.
- Alwin, D.F., Braun, M., & Scott, J. (1992). The separation of work and the family: Attitudes towards women's labour-force participation in Germany, Great Britain, and the United States. *European Sociological Review*, 8(1), 13-37. <https://www.jstor.org/stable/522315>

- Bianchi, S.M., Milkie, M.A., Sayer, L.C., & Robinson, J.P. (2000). Is Anyone Doing the Homework? *Trends in the Gender Division of Household Labor. Social Forces*, 79(1), 191-228.
- Bingöl, O. (2014). Toplumsal cinsiyet olgusu ve Türkiye’de kadınlık [The concept of gender and femininity in Turkey]. *KMU Journal of Social and Economic Research*, 16(1), 108-114.
- Blackton, R.R. (1986). A study of the correlation between the degree of acculturation and scholastic achievement and English gain of ESL students, grades 2-5, Beach School, Portland, Oregon. Dissertations and Theses. Paper 3560.
- Bora, A. (2011). *Kadınların sınıflı ücretli ev emeği ve kadın öznelliğinin inşası* [Women's class paid home labor and the construction of female subjectivity]. İstanbul: İletişim Publications.
- Bölükoğlu, A. (2018). Yedek işgücü olarak kadın emeği: Türkiye Örneği:1988-2013 [Women's labor as a reserve labor force: The Case of Turkey]. *Journal of Economics, Policy & Finance Research*, 3(1), 50-67.
- Brewe, E., Kramer, L., & O’Brien, G. (2009). Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. *Physical Review Special Topics-Physics Education Research*, 5(1), 013102.
- Britton, D.M. (2000). The epistemology of the gendered organization. *Gender and Society*, 14(3), 418-434.
- Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Browne, M., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, 21(2), 230-258.
- Byrne, B.M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling*, 11(2), 272-300. DOI: [10.1207/s15328007sem1102\\_8](https://doi.org/10.1207/s15328007sem1102_8)
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504.
- Chen, F.F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005-1018.
- Cheung, G.W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. DOI: [10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Coltrane, S. (2000). Research on household labor: Modeling and measuring the social embeddedness of routine family work. *Journal of Marriage and the Family*, 62, 1208-1233.
- Corrigan, E.A., & Konrad, A.M. (2007). Gender role attitudes and careers: A longitudinal study. *Sex Roles*, 56(11-12), 847-855. <https://doi.org/10.1007/s11199-007-9242-0>
- Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, 10(7), 1-9.
- Çiçek, B., & Çopur, Z. (2018). Bireylerin kadınların çalışmasına ve toplumsal cinsiyet rollerine ilişkin tutumları [Attitudes of individuals towards women's work and gender roles]. *International Journal of Eurasian Education and Culture*, 4, 1-21.
- Dayioğlu, M., & Kirdar, M.G. (2010). Determinants of and trends in labor force participation of women in Turkey (English). Welfare and social policy analytical work program; working paper no. 5. Washington DC: World Bank. <http://documents.worldbank.org/cur>

[ated/en/466591468316462301/Determinants-of-and-trends-in-labor-force-participation-of-women-in-Turkey](https://doi.org/10.1501/466591468316462301/Determinants-of-and-trends-in-labor-force-participation-of-women-in-Turkey).

- Dex, S. (1988). *Women's attitudes towards work*. Springer.
- Dimitrov, D.M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development*, 43(2) 121–149.
- Dökmen, Z.Y. (2004). *Toplumsal Cinsiyet: Sosyal Psikolojik Açıklamalar* [Gender: Social Psychological Explanations]. İstanbul: Sistem Publishing.
- Erkuş, A. (2012). Varolan ölçek geliştirme yöntemleri ve ölçme kuramları psikolojik ölçek geliştirmede ne kadar işlevsel: Yeni bir öneri [Existing scale development methods and measurement theories how functional are psychological scale development: A new proposal]. *Journal of Measurement and Evaluation in Education and Psychology*, 3(2), 279-290.
- Ersoy, E. (2009). Cinsiyet kültürü içerisinde kadın ve erkek kimliği: Malatya Örneği [Male and female identity in gender culture: Malatya Case]. *Firat University Journal of Social Sciences*, 19(2), 209-230.
- Fabrigar, L.R., MacDonald, T. K., & Wegener, D. T. (2005). The Structure of attitudes from: The Handbook of Attitudes Routledge. Accessed on: 21 May 2019. <https://www.routledgehandbooks.com/doi/10.4324/9781410612823.ch3>
- Fabrigar, L. R. & Wegener, D. T. (2010). Attitude structure. Em R. F. Baumeister & E. J. Finkel (Orgs.), *Advanced social psychology: The state of the science* (pp. 177-216). New York: Oxford University Press.
- Ferrao, V. (2010). *Paid work. Women in Canada: A Gender-Based Statistical Report*. Sixth edition. Statistics Canada. Catalogue 89-503X.
- Forsythe, N., Korzeniewicz, R.P., & Durrant, V. (2000). Gender inequalities and economic growth: A longitudinal evaluation. *Economic Development and Cultural Change*, 48(3), 573-617. <https://www.jstor.org/stable/10.1086/452611>.
- Fredricks, J.A., Blumenfeld, P.C, & Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59-109.
- GGGR (2018). <https://www.weforum.org>
- Glasman, L.R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin*, 132(5), 778–822.
- Gubernskaya, Z. (2010). Changing attitudes toward marriage and children in six countries. *Sociological Perspectives*, 53, 179-200.
- Gürler, Ö., & Üçdoruk, Ş. (2007). Türkiye’de cinsiyete göre gelir farklılığının ayrıştırma yöntemiyle uygulanması [Applying the decomposition method of income differences by gender in Turkey]. *Journal of Yasar University*, 2(6), 571-589.
- Heilman, M.E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, 57(4), 657-674.
- Hersch, J., & Stratton, L. (2002). Housework and Wages. *The Journal of Human Resources*, 37(1), 217-229. DOI:10.2307/3069609
- Himmelweit, S. (2002). Making visible the hidden economy: The case for gender-impact analysis of economic policy. *Feminist Economics*, 8(1), 49-70. DOI: 10.1080/13545700110104864
- Hinner, M.B. (2019). The cultural perspective of mergers & acquisitions: An exploratory study. In J. Stumpf-Wollersheim & A. Horsch (Eds.), *Forum mergers & acquisitions*. Springer
- Gabler, Wiesbaden.Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>.

- Kakıcı, H., Emeç, H., & Üçdoğruk, Ş. (2007). Türkiye’de çalışan kadınların çocuk bakım tercihleri [Child care preferences of working women in Turkey]. *Istanbul University Faculty of Economics Journal of Econometrics and Statistics*, 5, 20-40.
- Kaufman, G., & Uhlenberg, P. (2000). The Influence of Parenthood on the Work Effort of Married Men and Women. *Social Forces*, 78(3), 931-947. DOI:10.2307/3005936
- Kılıç, D., & Öztürk, S. (2014). Türkiye’de kadınların işgücüne katılımı önündeki engeller ve çözüm yolları: Bir ampirik uygulama [Barriers to women's labor force participation and solutions in Turkey: An empirical application]. *Journal of Public Administration*, 47(1), 107-130.
- Kıral, G., & Karlılar, S. (2017). Türkiye’de kadın işgücüne katılımını etkileyen faktörler: Adana ili üzerine bir uygulama [Factors affecting the labor force participation of women in Turkey: Adana province on an application]. *Journal of Çukurova University Institute of Social Sciences*, 26(3), 272-286.
- Kırkpınar, L. (2001). *Türkiye’de toplumsal değişme ve kadın [Social change and women in Turkey]*. Ankara: Ministry of Culture Publications.
- Kızılırmak, A.B. (2005). *Labor market participation decisions of married women: Evidence from Turkey*. Oxford University Press: Oxford.
- Kline, R.B. (2011). *Principles and practice of structural equation modeling*. New York: The Guilford Press.
- Koca, C., Arslan, B., & Aşçı, F.H. (2011). Attitudes towards women's work roles and women managers in a sports organization: The case of Turkey. *Gender, Work & Organization*, 18(6), 592-612. DOI:10.1111/j.1468-0432.2009.00490.x
- Koray, M. (1992). Çalışma yaşamında kadın gerçekleri [Woman facts in working life]. *Journal of Public Administration*, 25(1), 93-122
- Köseoğlu, S. (2017). Erkek üniversite öğrencilerinin kadınların çalışmasına yönelik tutum ölçeği [The attitude scale of male university students towards the work of women]. *TBB Journal*, 30, 303-312.
- Kuzgun, Y., & Sevim S.A. (2004). Kadınların çalışmasına karşı tutum ve dini yönelim arasındaki ilişki [The relationship between attitude towards women's work and religious orientation]. *Ankara University Journal of Faculty of Educational Sciences*, 37(1), 14-27.
- Lips, H.M. (2001). *Sex and Gender: An Introduction*. California: Mayfield Publishing Company.
- Lubke, G.H., & Muthén, B.O. (2004) Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(4), 514-534. DOI: 10.1207/s15328007sem1104\_2
- Maio, G.R., Haddock, G., & Verplanken, B. (2018). *The psychology of attitudes and attitude change*. Sage Publications Limited.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- MLSSF (2014). Analysis of the female labor force profiles and statistics in Turkey. <http://kadininstatusu.aile.gov.tr/>
- Morin, A.J.S., Arens A.K., & Marsh, H.W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct- relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 116–39.
- Negiz, N., & Yemen, A. (2011). Kamu örgütlerinde kadın yöneticiler: Yönetici ve çalışan açısından yönetimde kadın sorunsalı [Women executives in public organizations:



- Women's problem in management in terms of managers and employees]. *SDU Faculty of Arts and Sciences Journal of Social Sciences*, 24, 195-214.
- Noor, A.M., & Saad, R.A.J. (2016). The mediating effect of trust on the relationship between attitude and perceived service quality towards compliance behavior of zakah. *International Journal of Economics and Financial Issues*, 6(7S), 27-31. available online at: [www.econjournals.com](http://www.econjournals.com)
- Nunnally, J.C., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.
- OECD (2016). Labor force statistics: Labor force participation rate. [www.oecd/data](http://www.oecd/data)
- Özar, Ş. (2005). *GAP Bölgesi'nde kadın girişimciliği [Women entrepreneurship in the GAP Region]*. Ankara, GAP-GİDEM Publications, January.
- Özçatal, E.Ö. (2011). Ataerkillik, toplumsal cinsiyet ve kadının çalışma yaşamına katılımı [Patriarchy, gender and women's participation in working life]. *Çankırı Karatekin University Journal of Economics and Administrative Sciences*, 1(1), 21-39.
- Özkaplan, N. (2013). Kadın akademisyenler: Cam tavanlar hâlâ çok kalın! [Female academics: Glass ceilings are still too thick!]. *Journal of Women's Studies*, 12(1), 1-23.
- Palaz, S. (2015). The reasons for women's labor force non-participation: *Empirical evidence from Bandırma*. *Journal of Management and Economics Research*, 13(3), 435-449.
- Perkins, K. K., Adams, W.K., Pollock, S.J., Finkelstein, N.D., & Wieman, C.E. (2005). Correlating student beliefs with student learning using the Colorado Learning Attitudes about Science Survey. *In AIP Conference Proceedings*, 790(1), 61-64.
- Powell, A. (2019). *Women and the Economy*. Briefing paper. Number CBP06838.
- Rentz, J.O., Shepherd, C. D., Tashchian, A., Dabholkar, P.A., & Ladd, R.T. (2002). A measure of selling skill: Scale development and validation. *Journal of Personal Selling & Sales Management*, 22(1), 13-21. <https://doi.org/10.1080/08853134.2002.10754289>
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner. *Journal of Management*, 20, 643-671.
- Scott, J. & Clery, E. (2013). Gender roles: An incomplete revolution? In Park, A., Bryson, C., Clery, E., Curtice, J. and Phillips, M. (Eds.) (2013), *British Social Attitudes: the 30th Report*, London: NatCen Social Research. Available online at: [www.bsa-30.natcen.ac.uk](http://www.bsa-30.natcen.ac.uk)
- Şengül, S., & Kıral, G. (2006). Türkiye'de kadının işgücü pazarına katılım ve doğurganlık kararları [Participation of women in the labor market and fertility decisions in Turkey]. *Atatürk University Journal of Economics and Administrative Sciences*, 20(1), 89-99.
- Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Tan, M. (2000). *Eğitimde kadın-erkek eşitliği ve Türkiye gerçeği. Kadın-Erkek Eşitliğine Doğru Yürüyüş-Eğitim, Çalışma Yaşamı ve Siyaset içinde*. [Gender equality and the reality of Turkey in education. Walking towards equality between men and women-in education, working life and politics]. Ankara: TÜSİAD Publications.
- Tansel, A. (2002). Economic development and female labor force participation in Turkey: Time series evidence and cross-province estimates. *Economic Research Center Working Paper in Economics*, 1(5), 1-37.
- TUIK (2016). Labor force statistics: 2007-2016. <http://www.tuik.gov.tr>
- TUIK (1990). <http://www.tuik.gov.tr>
- TUIK (2000). <http://www.tuik.gov.tr>
- UN (2012). United Nations. *The Millennium Development Goals Report 2012*. Retrieved August 17, 2019, <https://www.un.org/millenniumgoals/pdf/MDG%20Report%202012.pdf>
- U.S. Bureau of Labor Statistics (2019). Retrieved September 3, 2019, from <https://statusofwomendata.org/explore-the-data/employment-and-earnings/>

- 
- Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Weston, R., Gore Jr, P. A., Chan, F., & Catalano, D. (2008). An introduction to using structural equation models in rehabilitation psychology. *Rehabilitation Psychology*, 53(3), 340-356. DOI: [10.1037/a0013039](https://doi.org/10.1037/a0013039)
- Worthington, R., & Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practices. *Counseling Psychologist*, 34, 806-838.
- Yamak, R., Abdiođlu, Z., & Mert, N. (2012). Türkiye’de işgücüne katılımı belirleyen faktörler: Mikro ekonomik analiz [Determinants of labor force participation in Turkey: Micro-economic analysis]. *Anadolu University Journal of Social Sciences*, 12(2), 41-58.
- Yıldırım, K., & Dođrul, G. (2008). Çalışmak ya da Çalışmamak: Türkiye’de Kentsel Alanlarda Yaşayan Kadınların İşgücüne Katılmama Kararlarının Olası Belirleyicileri [Work or not work: Possible Determinants of Women Living in Urban Areas in Turkey's decision not to participate in the labor force]. *Anadolu University Journal of Social Sciences*, 8(1), 239-262.
- Yinger, R.J. (1980). *Can we really teach them to think?* In R. E. Young (Ed.), *Fostering critical thinking*. San Francisco: Jossey-Bass Inc
- Yuan, K. H. (2005). Fit statistics versus test statistics. *Multivariate Behavioral Research*, 40, 115-148.

## 10. APPENDIX: Turkish form of the scale

**Table A1.** Kadınların Çalışmasına yönelik Tutum Ölçeği

Boyut	Madde no	Türkçe Form
Cisiyete dayalı Ayrımcılık	M1	Kadınlar biyolojik döngüleri (regl, hamilelik, doğum, menopoz vb.) dolayısıyla iş yerinde verimli olamaz.
	M2	Kadınlar uzun ve ağır çalışma koşullarına erkekler kadar dayanıklı değildir.
	M3	Kadınlar psikolojik olarak iş baskısını erkekler kadar tolere edemez.
Ataerkil Değerler	M4	Kadınlar sadece kadınlara has işlerde çalışmalıdır.
	M5	İşe alımlarda erkeklere öncelik verilmelidir çünkü erkek, ailenin temel geçiminden sorumludur.
	M6	Kadınların çalışma ortamı evi olmalıdır.
İş Ortamı	M7	İş yerinde kadınların olması verimi düşürür.
	M8	Kadınların iş yerinde dişiliklerini kullanması haksız rekabete yol açar
	M9	Yönetici pozisyonlara erkekler tercih edilmelidir.

## Physics Course Attitudes Scale for High School Students: A Validity and Reliability Study

Hulya Cermik<sup>1</sup>, Izzet Kara<sup>1,\*</sup>

<sup>1</sup>Pamukkale University, Faculty of Education, Kınıklı Campus, 20070, Denizli, Turkey

### ARTICLE HISTORY

Received: 08 May 2019

Revised: 24 December 2019

Accepted: 06 January 2020

### KEYWORDS

Physics course,  
Physics course attitudes  
scale,  
Validity,  
Reliability

**Abstract:** The purpose of this study is to develop an up-to-date scale with high validity and reliability that could reveal the attitudes of high school students towards the physics course. In the process of developing the scale in question, three independent samples were formed, and the data obtained from a total of 1118 high school students were analyzed. Firstly, the opinions of 152 high school students on the physics course were collected in written form, and a 58-item pool was formed. Afterwards, the draft scale which was designed as a 5-point Likert-type scale whose items were reduced to 43 based on expert opinions was applied on 602 high school students. Based on the data obtained, an exploratory factor analysis (EFA) was carried out. With the EFA, it is determined that 22 items of the scale have factor loads between 0.490 and 0.816, while they explain 66.276% of the total variance and are distributed under four factors. These factors are named as *interest*, *unwillingness*, *academic self* and *necessity*. Additionally, these four factors are significantly correlated, and there is no autocorrelation problem. For all items in the scale, item-factor and item-test correlation coefficients were calculated, and it is determined that each item is consistent with not only the factor it is under but also the entire test.

## 1. INTRODUCTION

Today, individuals who go through the education and training process are well-equipped in the field of physics, which has, undoubtedly, an important effect on their personal lives, on their professional development, and, moreover, on scientific developments in national and international arenas. This fact is also considered by educational institutions in determining the learning outcomes of the "physics course". Since knowledge of physics is of high importance during their education, students take physics courses usually in the first year of their high school education. However, physics course is considered to be difficult in both learning and teaching it (Angell, Guttersrud, Henriksen & Isnes, 2004; Mualem & Eylon, 2007; Mulhall & Gunstone, 2008). In addition, students find physics course difficult and also, they think it is boring (Williams, Stanisstreet, Spall & Boyes, 2003). It is also stated that the academic success level of students in physics education is rather lower than that in other disciplines (Rivard & Straw, 2000). That students see physics and physics classes difficult and boring or have lower success in comparison to other disciplines is a significant problem as well. Therefore, it is definitely

CONTACT: Izzet Kara ✉ [ikara@pau.edu.tr](mailto:ikara@pau.edu.tr) 📧 Pamukkale University, Faculty of Education, Department of Mathematic and Sciences Education, Kınıklı Campus, 20070, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

necessary to determine and eliminate the source of this problem. Determining students' attitudes towards the physics course can be one of the best ways to solve such a problem.

İnceoğlu (2010) defined attitude as a preliminary tendency of a mental, emotional and behavioral reaction organized by the person towards any object, social issue or event based on their experience, knowledge, emotions and urges. Developing a positive attitude towards a subject, in addition to willingness to participate in the class, involves a set of behaviors in the form of satisfaction from responding, acceptance of the value of the course and being an advocate of this acceptance (Özçelik, 1998). Papanastasiou and Zembylas (2002) stated that a positive attitude towards science increases the success of students in this field, whereas success does not guarantee positive attitudes. Another study emphasized that physics teachers should focus more on the attitudes of their students towards learning physics rather than focusing merely on their learning of physics (Veloo, Nor, & Khalid, 2015). This situation reveals the necessity of measurement instruments that may demonstrate the attitudes of students towards the physics course.

In the related literature, it is seen that scales have been developed to determine attitudes towards physics for different objectives (Douglas, Yale, Bennett, Haugan, & Bryan, 2014; Faour & Ayoubi, 2018; Moll & Milner-Bolotin, 2009; Olusola & Rotimi, 2012). The common characteristic of these studies is determination of the effects of attitude on the physics course. For example, the scale that was developed by Kurnaz and Yiğit (2010) aimed to determine the attitudes of high school students towards physics, physics-related topics and studies that are conducted. The 4-point Likert-type scale that was applied on 841 students who were enrolled at seven different types of high schools was given its final form by conducting an EFA. In the scale with the Cronbach's Alpha internal consistency coefficient of 0.95 and three factors, the factors were named as 'valuing physics', 'turning physics into a behavior' and 'point of view towards physics. In another study Tekbıyık and Akdeniz (2010) developed an up-to-date physics attitudes scale for high school students. The sample of the study consisted of 166 ninth-grade students. The 30-item, 5-point Likert-type scale that was given its final form by EFA had 4 factors named as importance, comprehension, necessity and interest. In their study Nalçacı, Akarsu and Kariper (2011) developed a 30-items scale for measuring the attitudes of high school students towards the field of physics by reviewing a number of physics attitude scales. The scale was developed as a 5-point Likert-type scale consisting of 30 items including 12 negative and 18 positive statements. The analyses that were conducted on the data collected from 303 students in total were highly limited. As a result of these analyses, it was stated that no item was removed from the scale as there was no item with a correlation value under 0.20, and the Cronbach's Alpha reliability coefficient of the scale was reported as 0.94. In another study, Kaur and Zhao (2017) developed a physics attitude scale by using data obtained from 624 students at the ages of 15-18 in India. Their scale consisted of five dimensions as Enthusiasm toward Physics, Physics Learning, Physics as a Process, Physics Teacher and Physics as a Future Vocation. In addition, it is possible to encounter a set of studies in the literature on the development or adaptation of physics attitudes scales (Özyürek & Eryılmaz, 2001; Taşlıdere, 2002; Taşlıdere & Eryılmaz, 2009; Tekbıyık, 2010; Uz & Eryılmaz, 1999). As seen, the scales in question usually focused on revealing the attitudes of students towards physics or physics-related topics. There is, on the other hand, a limitation in the scales that were developed to reveal attitudes of high school students towards the physics course. In the literature, under this title only the 36-item, 5-point Likert-type "Physics Course Attitudes Scale" that was developed by Akpınar (2006) can be encountered. This scale consists of six factors; namely, interest in the physics course, the concept of self in relation to the physics course, willingness to work on physics outside the school, thoughts on the relationship between the physics course and life, thoughts on the work required by the physics course and general

thoughts on the physics course. It is not possible to say that this scale, which was prepared as a physics course attitudes scale, is up to date.

This study, which was carried out based on the importance of revealing the attitudes of high school students towards the physics course, was conducted to develop an up to date, valid and reliable measurement instrument that could reveal students' attitudes towards the topic.

## 2. METHOD

### 2.1. Samples

In this study, three independent samples were formed out of 1118 high school students studying at the city center of the province of Denizli in Turkey and as stated by Seer (2015) different samples were created for conducting scale development studies.

The first sample consisted of 152 high school students whose opinions were consulted to create the item pool of the scale. The students in this group answered the open-ended question directed towards themselves during the process of creating attitude items. Among the students, 82 (53.9%) were female, and 70 (46.1%) were male. The second sample completed the 43-item draft scale that was prepared to reveal the attitudes of high school students towards the physics course. There was a total of 602 high school students in this sample. Among these students, 330 (54.8%) were female, and 272 (45.2%) were male. 167 (27.5%) of the students were enrolled at Science High Schools, 368 (61.1%) were students of Anatolian High Schools, and 67 (11.1%) were students of Religious Vocational High Schools. An exploratory factor analysis (EFA) was conducted on the data obtained from the second sample. For a sufficient sample size for factor analysis, Comrey and Lee (1992) stated that 300 is good, and 500 is very good, and while Kline (1994) stated that 200 individuals are sufficient for a sample size with reliable factors and they also recommended the sample size to be 10 times more than the number of items. Considering these issues, it was concluded that 602 high school students in the second sample were sufficient.

The third sample consisted of 364 high school students and was formed to confirm the construct to be analyzed. A confirmatory factor analysis (CFA) was carried out on the data obtained from this group. In addition to a sample size of larger than 300, as the scale that was applied on the group consisted of 22 items, it was ensured that the number of observations per item was higher than 10 individuals. Additionally, a particular attention was given to form a similar group of participants in the third sample by keeping in mind the percentage values of the genders and types of high schools in the second sample. Among the students in this group, 202 (55.5%) were female, and 162 (44.5%) were male. Of these students, 103 (28.3%) were Science High School students, 219 (60.2%) were Anatolian High School students, and 42 (11.5%) were Religious Vocational High School students.

### 2.2. Developing the measurement instrument

Within the scope of the development of the Physics Course Attitudes Scale, firstly, the opinions of the high school students in the first sample about the physics course were collected in writing. The data that were collected in written form from the 152 students in this group were examined, and an item pool of 58 items related to attitudes towards the physics course was created. The items in question were submitted for the opinions of a total of seven experts including two measurement and assessment, two curriculum development, two physics and one linguistics experts. Based on all the feedback received from the experts, the items that did not express attitudes towards the physics course and those that expressed similar meanings were removed. By eliminating the problems in the linguistic and semantic aspects of the items, a draft scale form consisting of 43 items including 21 positive and 22 negative statements was created. The scale items were designed to be scored as 5-point Likert-type items; namely, (1) Absolutely

agree, (2) Agree, (3) Somehow agree, (4) Disagree and (5) Absolutely disagree. The lowest possible score from this scale was 43, while the highest was 215.

### **2.3. Data analysis**

To conduct the validity and reliability analyses of the measurement instrument, the data obtained from the second and third samples were uploaded onto the SPSS 22.00 and AMOS 16 software and analyzed. Firstly, for the purpose of determining the construct validity of the scale, KMO and Bartlett's tests were carried out on the data obtained from the second sample to see the data's suitability for factor analysis. Based on the obtained values, an EFA was carried out on the data. Additionally, for each item in the scale, the item-factor and item-test correlation values were calculated with a purpose to see whether each item was consistent with its factor and the entire scale. Afterwards, a CFA was conducted on the data obtained from the third sample. To determine the reliability of the scale, the Cronbach's Alpha reliability coefficient method was used.

## **3. FINDINGS**

### **3.1. Findings on validity**

Construct validity was applied to the scale in order to determine the extent to which the attitude scale as the measurement instrument can measure the variable it aims to measure without confusing it with other variables (Balci, 2009). To determine the construct validity of the Physics Course Attitudes Scale, firstly, Kaiser-Meyer-Olkin (KMO) and Bartlett's test analyses were conducted on the data collected from the second sample, and the values were obtained as KMO= 0.945; Bartlett's test value  $\chi^2=7782.179$ ;  $df=231$  ( $p=0.000$ ). As KMO values of higher than 0.60 are seen to be sufficient for factor analysis in the social sciences (Büyüköztürk, 2002), it was decided that factor analysis could be conducted on the 43-item scale.

In Exploratory Factor Analysis, Principal Component Analysis (PCA) is a technique that is used to reveal whether or not the items in a scale could be divided into a lower number of factors that exclude each other (Büyüköztürk, 2002). Moreover, to clarify the factors that are formed by gathering the items, the Varimax orthogonal rotation technique was applied. Accordingly, PCA was carried out on the data, the Varimax orthogonal rotation technique was applied to see whether or not the scale could be divided into independent factors, and the factor loads were examined. Items that have factor load values under 0.30 and those that are distributed under more than one factor with less than a difference of 0.10 between their factors loads need to be removed from the scale (Balci, 2009; Büyüköztürk, 2002). As a result of the analyses in this study, the eigenvalues of the items had to be at least 1.00, while their factor loads at least 0.45. Items that were distributed under multiple factors were eliminated, 21 items were removed, and the analyses were carried out on the remaining 22 items.

A total of 22 items remaining in the scale were found to be distributed under four factors. Among these items, 12 had positive and 10 had negative statements. Without subjecting the remaining 22 items to rotation, it was found that the factor loads varied between 0.477 and 0.823. After subjecting the items to the Varimax orthogonal rotation technique, these factor loads were found to vary between 0.490 and 0.816. Moreover, it was determined that the items and factors in the scale explained 66.276% of the total variance. As it was stated that this ratio needs to be at least 52% (Henson & Roberts, 2006), the obtained value was found sufficient. This finding obtained by EFA is shown in [Figure 1](#) based on the eigenvalues. In [Figure 1](#), it is seen that there were steep drops in the first four factors, and therefore, these factors had significant contribution to the variance.

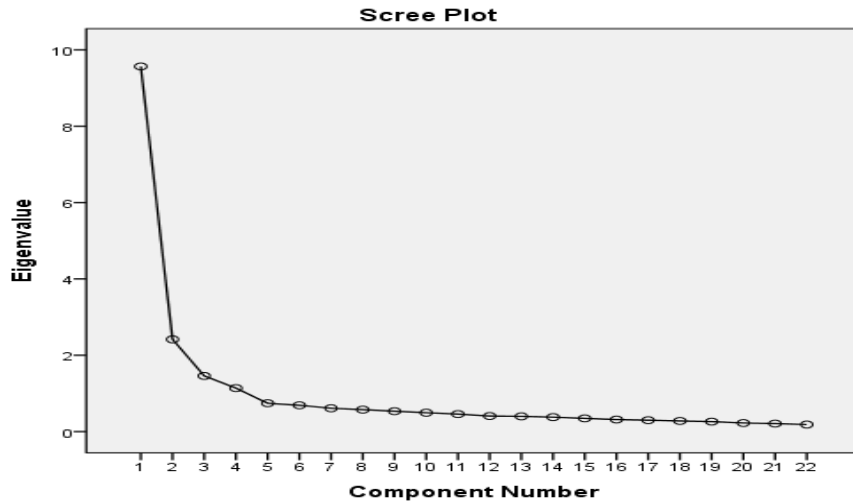


Figure 1. Eigenvalues based on the factors

Furthermore, the factors were named by examining the contents of the items gathered under these four factors. There were six items in each of the factors named *interest* and *unwillingness*, while there were five items in each of the factors named *academic self* and *necessity*. Table 1 presents findings on the item loads of the remaining 22 items based on the factors, factor eigenvalues and variance explanation ratios.

Table 1. Physics Course Attitudes Scale common variances, item factor loads, variances explained by sub-scales and item analysis results

Items	Common variance	F1 Interest	F2 Unwillingness	F3 Academic self	F4 Necessity
I37	0.766	0.805			
I43	0.632	0.780			
I38	0.702	0.750			
I32	0.685	0.710			
I33	0.762	0.677			
I30	0.673	0.655			
I20	0.733		0.816		
I19	0.715		0.807		
I22	0.754		0.785		
I23	0.775		0.767		
I27	0.725		0.760		
I26	0.519		0.633		
I2	0.686			0.773	
I1	0.720			0.750	
I5	0.635			0.688	
I3	0.662			0.576	
I7	0.662			0.566	
I9	0.652				0.765
I10	0.670				0.748
I11	0.628				0.711
I14	0.510				0.578
I12	0.507				0.490
Eigenvalue		9.566	2.417	1.458	1.140
Explained variance		43.480	10.987	6.629	5.180
Total variance				66.276%	



As seen in Table 1, the factor loads of the items in the factor *interest* of the scale varied between 0.655 and 0.805. The eigenvalue of this factor in the general scale was 9.566, and its contribution to the general variance was 43.480%. The factor loads of the items in the factor *unwillingness* varied between 0.633 and 0.816. The eigenvalue of this factor was 2.417, and its contribution to the general variance was 10.987%. The factor loads of the items in the factor *academic self* varied between 0.566 and 0.773. The eigenvalue of this factor was 1.458, and its contribution to the general variance was 6.629%. The factor loads of the items in the factor *necessity* varied between 0.490 and 0.765. The eigenvalue of this factor was 1.140, and its contribution to the general variance was 5.180%.

In addition, the relationship between the four factors in the Physics Course Attitudes Scale was determined and for this reason, the correlations among the factors were checked. The findings are shown in Table 2.

**Table 2.** Correlation analysis results among the factors of the Physics Course Attitudes Scale

Factors	Interest	Unwillingness	Academic self	Necessity
Interest	1			
Unwillingness	0.520**	1		
Academic Self	0.627**	0.600**	1	
Necessity	0.645**	0.487**	0.488**	1

\*\* $p < 0.01$

As seen in Table 2, based on the correlation values among the factors of the Physics Course Attitudes Scale, the four factors were found to be significantly related, while there was no problem of autocorrelation.

The correlation coefficients between the scores obtained from all items and the scores obtained from the factors and the scale were also calculated, and the discrimination rate of each item was determined in order to reveal the degree to which each item served the general purpose of the factor it was in and the entire scale (Balçı, 2009; Korkmaz, Şahin, & Yeşil, 2011). Table 3 presents the items of the scale: the first column shows the initial numbers of the items, the second column shows the updated numbers and negatively worded statements; and the other remaining columns present the items, item-factors, item subscale correlations and item-test correlations.

As seen in Table 3, the item-factor correlations were in the ranges of 0.620-0.817 for the first factor, 0.603-0.812 for the second factor, 0.599-0.707 for the third factor and 0.442-0.652 for the fourth factor. Each item had a significant and positive relationship with the general scale ( $p < 0.001$ ). When the item-test correlation coefficients for the entire scale were examined, the lowest correlation value was found as 0.438, while the highest one was 0.789. These coefficients that were obtained were the validity coefficients of all items, and they showed the consistency of the items with both their factor and the entire scale. In other words, these referred to the degree to which the scale served its general objective (Baykul, 2000).

The dimensions of the 'Physics Course Attitudes Scale' were determined to consist of four factors as a result of the EFA. To confirm these factors, the scale that consisted of 22 items was applied on the third sample that was selected independently of the second sample, and a CFA was carried out on the data. CFA is based on the relationship among observable and unobservable variables and testing them as hypotheses (Pohlmann, 2004).

**Table 3.** Item-Test correlation analysis results

Initial item no	Updated item	Items	Factor	Item-Subscale Correlation	Item-Test correlation
37	22	I look forward to the physics course.	F1	0.797	0.699
43	13	I enjoy daily repetition of what I learn in the physics course.	F1	0.620	0.505
38	4	I am more willing to study for the physics course than other courses.	F1	0.749	0.686
32	8	I enjoy conducting in-depth research on what I learn in the physics course.	F1	0.774	0.667
33	21	Topics of the physics course attract my interest.	F1	0.817	0.789
30	17	What I learn in the physics course excites me.	F1	0.756	0.692
20	6*	I see the physics course as waste of time.	F2	0.742	0.581
19	15*	I do not want to go to school on days of the physics course.	F2	0.760	0.601
22	9*	I am very bored during the physics course.	F2	0.802	0.691
23	3*	It is a torture for me to study for the physics course.	F2	0.812	0.722
27	12*	I would not attend the physics course if I were able to.	F2	0.762	0.690
26	19*	I feel very nervous during the physics course.	F2	0.603	0.514
2	11*	Physics is not a subject I can learn by my own effort without receiving special support.	F3	0.661	0.547
1	1*	I do not believe I could be successful however much I study for the physics course.	F3	0.707	0.617
5	7*	I believe it is a miracle for me to understand the topics of the physics course.	F3	0.599	0.545
3	16	I see myself as a successful student in the physics course.	F3	0.685	0.700
7	14	The physics course is among the courses I can learn easily.	F3	0.610	0.614
9	2	I believe physics is an important subject that needs to be learnt by everyone.	F4	0.624	0.517
10	5	I believe our education would be lacking if there were no physics courses in high school curricula.	F4	0.652	0.585
11	18	I believe what I learn in the physics course makes my daily life easier.	F4	0.625	0.558
14	10	The physics course is necessary for me to have a good occupation.	F4	0.468	0.438
12	20*	I do not think the physics course will be useful for me after I graduate from high school.	F4	0.442	0.501

\*Items with negative statements

According to the results that were obtained, the  $\chi^2/df$  ratio was calculated as 2.380. A  $\chi^2/df$  ratio of 5 or lower is considered to be sufficient for model data fit (Schumacker & Lomox, 2004; Wang, Lin & Luarn, 2006). Moreover, a  $\chi^2/df$  ratio of smaller than 3 shows a high model-data fit (Schumacker & Lomox, 2004). The  $\chi^2/df$  value obtained as 2.380 in this study was a significant indicator that the measurement instrument had four dimensions. Another important index, the RMR value was calculated as 0.084. It is known that the RMR index needs to be between 0 and 1 (Golob, 2003). Other fit indices were also calculated to assess the fit of the model. The calculated goodness of fit indices values were as: IFI=0.922; CFI=0.921;

GFI=0.897; NFI=0.872; AGFI=0.866, and RFI=0.849. While it is generally acceptable for the aforementioned indices to be in the range of 0.80-0.90, values higher than 0.90 refer to a good fit (Yap & Khong, 2006; Wang et al., 2006). The RMSEA analysis result was determined as 0.062. RMSEA values of lower than 0.10 show an acceptable level of model-data fit, while those lower than 0.05 are an indicator of a good fit (Bayram, 2013). Based on the  $\chi^2/df$ , RMSEA and RMR values obtained from the data in the study, it may be stated that the measurement instrument consisted of four factors. Figure 2 below shows the standardized Structural Equation Modelling parameter values on the obtained findings.

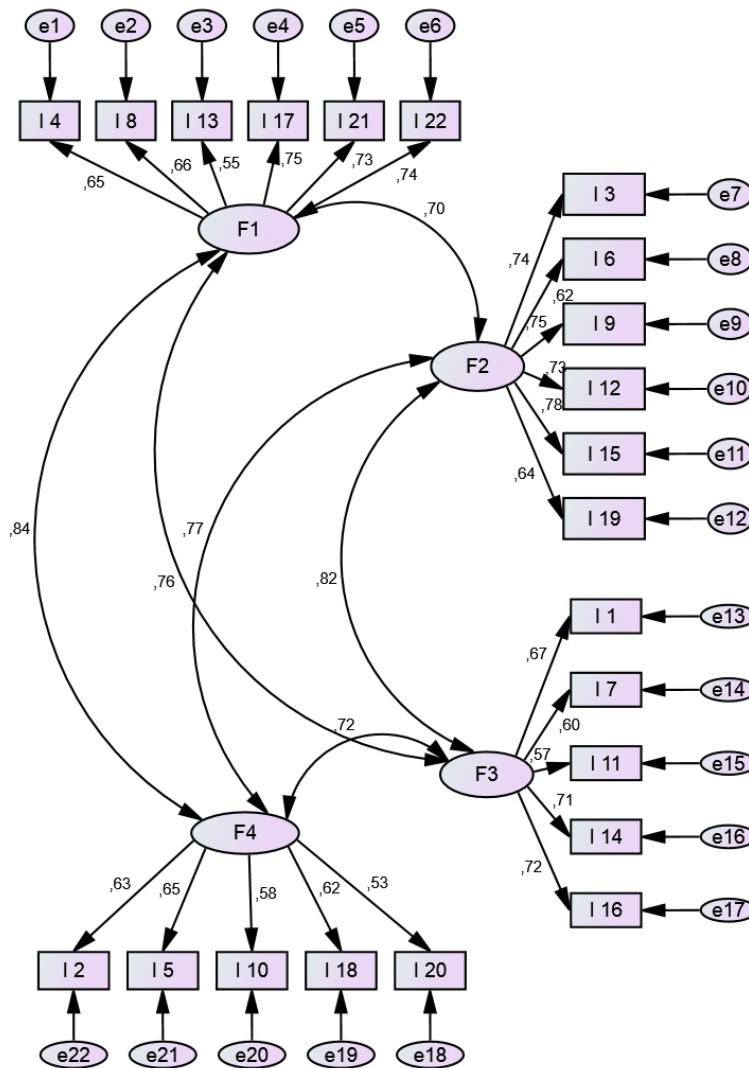


Figure 2. Confirmatory Factor Analysis results of the Scale

As a result of the Confirmatory Factor Analysis, it was confirmed that the ‘Physics Course Attitudes Scale’ consisted of 22 items and four factors.

### 3.2. Findings on reliability

Reliability is a concept that is related to whether or not a measurement instrument provides the same results in times of repeated application (Balçı, 2009; Baykul, 2000). As a result of the EFA, it was determined that the ‘Physics Course Attitudes Scale’ consisted of a total of 22 items and four factors. To determine the reliability rates of these factors in relation to internal consistency, their Cronbach’s Alpha reliability coefficients were obtained. The Cronbach’s

Alpha reliability coefficients of the factors were as 0.911 for Interest, 0.906 for Unwillingness, 0.845 for Academic self and 0.782 for Necessity. The Cronbach's Alpha value for the entire scale was 0.936. The Cronbach's Alpha coefficient takes values in the range of 0.00 to 1.00. As the coefficient gets closer to 1.00, the reliability of the measurement instrument increases, while as it gets closer to 0.00, the reliability decreases. In the social sciences, in general, Cronbach's Alpha coefficients of 0.60 or higher are seen to be sufficient. On the other hand, the reliability coefficient used for preparing and applying psychometric tests is expected to be 0.70 or higher (Büyüköztürk, 2002). According to the findings obtained, the internal consistency coefficients for the factors and the entire scale were quite high.

#### 4. DISCUSSION and CONCLUSION

One thousand one hundred and eighteen high school students participated in this study, which was conducted to develop an up-to-date, valid and reliable 'Physics Course Attitudes Scale'. At the first stage of the study, to reveal the opinions of high school students on the physics course, a draft form with 58 items was obtained based on the data collected from the 152 high school students in the first sample. From this draft form, based on the opinions of seven experts including experts on 'measurement and assessment', 'curriculum development', 'physics' and 'linguistics', a 43-item scale form was prepared. The 43-item, 5-point Likert-type 'Physics Course Attitudes Scale' was applied to a total of 602 high school students studying at the city center of the province of Denizli in Turkey who constituted the second sample of the study. At this stage, 21 items that were found to be statistically unsuitable were removed. An EFA was conducted on the data of the remaining 22 items, and item-factor and item-test correlations were calculated. According to the results, the scale consists of four factors which are named as *interest*, *unwillingness*, *academic self* and *necessity*. For the remaining 22 items in the scale, all findings on the item factor loads, factor eigenvalues and ratios of explaining the total variance were examined. Furthermore, it was found that there was a significant relationship among these four factors, and there was no problem of autocorrelation. The final scale was applied on the third sample consisting of 364 high school students, and a CFA was carried out on the obtained data. The EFA results were also confirmed with the results of the CFA.

The reliability degrees of the entire scale and the four factors in relation to their internal consistency were obtained by calculating Cronbach's Alpha coefficients. The Cronbach's Alpha value for the entire scale was obtained as 0.936. According to the findings, the internal consistency coefficients of the entire scale and the factors of the scale were high. The 'Physics Course Attitudes Scale' that was developed in this study, gathered under four factors and included 22 items containing 12 positive and 10 negative statements was found to be a valid and reliable scale based on the statistical data.

This scale which was developed with the purpose of revealing the attitudes of high school students towards the physics course is not only an up-to-date scale, but also a valid and reliable measurement instrument. For these reasons, this scale is believed to be an effective measurement instrument to determine and also monitor high school students' attitudes towards the physics course. The increases in the scores obtained from the scale is interpreted as a positive change towards the physics course.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## ORCID

Hülya Çermik  <https://orcid.org/0000-0002-5343-5441>

İzzet Kara  <https://orcid.org/0000-0002-9837-2819>

## 5. REFERENCES

- Akpınar, M. (2006). *Öğrencilerin fizik dersine yönelik tutumlarının fizik dersi akademik başarısına etkisi* (Unpublished master's thesis). Gazi University, Institute of Educational Sciences, Ankara.
- Angell, C., Guttersrud, Ø., Henriksen, E. K., & Isnes, A. (2004). Physics: Frightful, but fun, Pupils' and teachers' views of physics and physics teaching. *Science Education*, 88, 683–706.
- Balcı, A. (2009). *Sosyal bilimlerde araştırma: Yöntem, teknik ve ilkeler*. Ankara: PegemA Yayıncılık.
- Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik Test Teorisi ve uygulaması*. Ankara: ÖSYM Yayınları.
- Bayram, N. (2013). *Yapısal eşitlik modellemesine giriş* (2. baskı). Bursa: Ezgi Kitabevi.
- Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi el kitabı*. Ankara: PegemA Yayıncılık.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2<sup>nd</sup> ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Douglas, K. A., Yale, M. S., Bennett, D. E., Haugan, M. P., & Bryan, L. A. (2014). Evaluation of Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 10(2), 1-10/020128.
- Faour, M.A., & Ayoubi, Z. (2018). The effect of using virtual laboratory on grade 10 students' conceptual understanding and their attitudes towards physics. *Journal of Education in Science, Environment and Health*, 4(1), 54-68.
- Golob, T.F. (2003). Structural equation modeling for travel behavior research. *Transportation Research*, 37(1), 1-25.
- Henson, R.K., & Roberts, J.K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practise. *Educational and Psychological Measurement*, 66(3), 393-416.
- İnceoğlu, M. (2010). *Tutum, algı, iletişim*. İstanbul: Beykent Üniversitesi Yayınları, No: 69.
- Kaur, D., & Zhao, Y. (2017). Development of Physics Attitude Scale (PAS): An instrument to measure students' attitudes toward physics. *Asia-Pacific Educational Researcher*, 26(5), 291-304.
- Kline, P. (1994). *An easy guide to factor analysis*. New York: Routledge
- Korkmaz, Ö., Şahin, A., & Yeşil, R. (2011). Bilimsel araştırmaya yönelik tutum ölçeği geçerlilik ve güvenilirlik çalışması. *İlköğretim Online*, 10(3), 961-973.
- Kurnaz, M.A., & Yiğit, N. (2010). Fizik tutum ölçeği: Geliştirilmesi, geçerliliği ve güvenilirliği. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, 4(1), 29-49.
- Moll, R., & Milner-Bolotin, M. (2009). The effect of interactive lecture experiments on student academic achievement and attitudes towards physics. *Canadian Journal of Physics*, 87(8), 917-924.
- Mualem, R., & Eylon, B.S. (2007). 'Physics with a smile'-Explaining phenomenon with a qualitative problem-solving strategy. *Physics Teacher*, 45(3), 158-163.

- Mulhall, P. J., & Gunstone, R. (2008). Views about physics held by physics teachers with differing approaches to teaching physics. *Research in Science Education*, 38(4), 435-462.
- Nalçacı, İ.Ö., Akarsu, B., & Kariper, İ. A. (2011). Orta öğretim öğrencileri için fizik tutum ölçeği derlenmesi ve öğrenci tutumlarının değerlendirilmesi. *Journal of European Education*, 1(1), 1-6.
- Olusola, O. O., & Rotimi, C.O. (2012). Attitudes of students towards the study of physics in College of Education Ikere Ekiti, Ekiti State, Nigeria. *American International Journal of Contemporary Research*, 2(12), 86-89.
- Özçelik, D.A. (1998). *Ölçme ve değerlendirme*. Ankara: ÖSYM Yayınları.
- Özyürek, A., & Eryılmaz, A. (2001). Öğrencilerin fizik dersine yönelik tutumlarını etkileyen etmenler. *Eğitim ve Bilim Dergisi*, 26(120), 21-28.
- Papanastasiou, E.C., & Zembylas, M. (2002). The effect of attitudes on science achievement: a study conducted among high school pupils in Cyprus. *International Review of Education*, 48(6), 469-484.
- Pohlmann, J. T. (2004). Use and interpretation of factor analysis in the journal of educational research: 1992-2002. *The Journal of Educational Research*, 98(1), 14-23.
- Rivard, L.P., & Straw, S.P. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education*, 84, 566-593.
- Schumacker, R.E., & Lomax, R.G. (2004). *A Beginner's guide to Structural Equation Modeling* (2<sup>nd</sup> ed.). NJ: Lawrence Erlbaum Associates, Mahwah.
- Seçer, İ., (2015). *Psikolojik test geliştirme ve uyarlama süreci*. Ankara: Anı Yayıncılık.
- Taşlıdere, E. (2002). *The effect of conceptual on students' achievement and attitudes toward physics* (Unpublished master's thesis). Middle East Technical University, Institute of Science and Technology, Ankara.
- Taşlıdere, E., & Eryılmaz, A. (2009). Alternative to traditional physics instruction: effectiveness of conceptual physics approach. *Eurasion Journal of Educational Research*, 35, 109-128.
- Tekbıyık, A. (2010). *Bağlam temelli yaklaşımla ortaöğretim 9. sınıf enerji ünitesine yönelik 5E modeline uygun ders materyallerinin geliştirilmesi* (Unpublished doctoral dissertation) Karadeniz Technical University, Institute of Science, Trabzon.
- Tekbıyık, A., & Akdeniz, A.R. (2010). Ortaöğretim öğrencilerine yönelik güncel fizik tutum ölçeği: Geliştirilmesi, geçerlik ve güvenilirliği. *Türk Fen Eğitimi Dergisi*, 7(4), 134-144.
- Uz, H., & Eryılmaz, A., (1999). Effects of socioeconomic status, locus of control, prior achievement, cumulative gpa, future occupation and achievement in mathematics on students' attitudes toward physics. *Hacettepe University Journal of Education*, 17(17), 105-112.
- Veloo, A., Nor, R., & Khalid, R. (2015). Attitude towards physics and additional mathematics achievement towards physics achievement. *International Education Studies*, 8(3), 35-43.
- Wang, Y., Lin, H., & Luarn, P. (2006). Predicting consumer intention to use mobile service. *Information Systems Journal*, 16(2), 157-179.
- Williams, C., Stanisstreet, M., Spall, K., & Boyes, E. (2003). Why aren't secondary students interested in physics? *Physics Education*, 38(4), 324-329.
- Yap, B.W., & Khong, K.W. (2006). Examining the effects of customer service management (CSM) on perceived business performance via Structural Equation Modelling. *Applied Stochastic Models in Business and Industry*, 22, 587-605.

## When interviewing: how many is enough?

William W. Cobern <sup>1,\*</sup>, Betty AJ Adams <sup>1</sup>

<sup>1</sup>The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

### ARTICLE HISTORY

Received: 17 December 2019

Accepted: 11 February 2020

### KEYWORDS

Research methodology,  
Sample size,  
Generalization,  
Interview research,  
Qualitative research,  
External validity

**Abstract:** Researchers need to know what is an appropriate sample size for interview work, but how does one decide upon an acceptable number of people to interview? This question is not relevant to case study work where one would typically interview every member of a case, or in situations where it is both desirable and feasible to interview all target population members. However, in much of qualitative and mixed-methods research and evaluation, the researcher can only reasonably interview a subset of the target population. How big or small should that subset be? This paper provides a brief explanation of why the concept of generalization is inappropriate with respect to the findings from qualitative interviewing, what wording to use in place of generalization, and how one should decide on sample size for interviews.

## 1. INTRODUCTION

Researchers need to know what is an appropriate sample size for interview work, but how does one decide upon an acceptable number of people to interview? This question is not relevant to case study work where one would typically interview every member of a case, or in situations where it is both desirable and feasible to interview all target population members. However, in much of qualitative and mixed-methods research and evaluation, the researcher can only reasonably interview a subset of the target population. How big or small should that subset be?

We raise this issue because we have seen sample size, or interview numbers, questioned by both graduate students and faculty, but without much validation for their opinions. For example, a doctoral student of ours proposed to interview 20 parents of primary, middle, and high school students. The proposed 20 parents would thus be divided over three grade bands. This proposal was challenged by a few faculty and other graduate students for having too few parents in each band. These dissenters objected, argued that dividing 20 interviews across three grade bands would mean too small of an N per group, too few subjects in each band for generalization purposes. Subsequently the student and his committee decided he should focus on only one grade band, but for reasons unrelated to generalizability or N-size as voiced by the dissenters during the public presentation.

---

CONTACT: William W. Cobern ✉ [bill.cobern@wmich.edu](mailto:bill.cobern@wmich.edu) 📍 The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

ISSN-e: 2148-7456 /© IJATE 2020

With respect to interview work, the concept of generalization is misapplied, so on this point the student's objectors were mistaken. As it happens, efforts to misapply generalizability standards to purposive, qualitative research sampling is not uncommon among people who primarily do probabilistic, quantitative research. Still, there is a valid underlying question: what is an acceptable number for interview work? What follows is a brief explanation of why the concept of generalization is inappropriate with respect to the findings from qualitative interviewing, what wording to use in place of generalization, and how one should decide on interview number.

## 2. GENERALIZATION IS A STATISTICAL CONCEPT

The related concepts of generalization and sample size (N size) are from quantitative work (see for example, Teo, 2013). They have no counterparts in qualitative research including qualitative interviewing. Generalizability is a statistical concept that is often defended partially on the basis of finding a low enough resultant p-value, or probability value. If one uses the common significance level (alpha) or threshold of  $p < 0.05$  for statistical significance, it suggests about a 5% chance of getting this (or a more extreme) result by chance instead of as an accurate representation of a larger population. However, many conditions apply, including assumptions regarding data distribution modes, variance, and normality, for both the sample population and the larger population you might like to "generalize" about. There is rarely any certainty involved, and this is arguably even more true in education research than in medical trials or physics experiments, for instance, when comparing a control student group to a treatment student group for an instructional innovation. A resulting low p-value suggests that the null hypothesis (no difference between) is not true, but this does not necessarily mean that the treatment hypothesis is perfectly true.

Sample size expressed as an N value is related to the statistical concept of generalization through power calculations. Admittedly, researchers often neglect this calculation (typically because they are using convenience samples), but power calculations are used for estimating the N size needed to show statistically significant difference if such a difference exists.

In plain English, statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error, or concluding there is no effect when, in fact, there is one, goes down... Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples (Ellis, 2010).

As you can see, the ability to detect a true effect is sensitive to sample size. Hence, the ability to generalize is sensitive to sample size (Royall, 1986). However, statistical significance does not necessarily mean practical significance. A large enough sample size may allow the researcher to determine statistically that a very small difference between treatment and control conditions is significant where the difference is too small to have any practical value.

In qualitative work, such calculations do not exist and therefore the concept of generalization should not be applied to qualitative work. Nevertheless, it is good news for qualitative researchers that size isn't everything, not even in quantitative research. Indeed, years ago Cronbach offered the following advice on generalizing from quantitative data, advice insufficiently heeded by quantitative researchers:

Instead of making generalization the ruling consideration in our research, I suggest that we reverse our priorities. An observer collecting data in one particular situation is in a position to appraise a practice or proposition in that setting, observing effects in context. In trying to describe and account for what happened, he will give attention to whatever variables were controlled, but he will give equally careful attention to uncontrolled



conditions, to personal characteristics, and to events that occurred during the treatment and measurement. As he goes from situation to situation, his first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that local of series of events (cf. Geertz, 1973, chap. 1, on "thick description"). As results accumulate, a person who seeks understanding will do his best to trace how the uncontrolled factors could have caused local departures from the modal effect. That is, generalization comes late, and the exception is taken as seriously as the rule. (Cronbach, 1975, p. 124-125)

In this quote, Cronbach refers to Clifford Geertz and his notion of "thick description" which is a notion well-known amongst qualitative researchers. The point is that even in quantitative research, the qualitative description of relevant factors is essential to the understanding of practical significance.

### **3. EXTERNAL VALIDITY AND QUALITATIVE INTERVIEWING**

One can find discussions about sampling and generalizability in the literature on qualitative research (e.g., Gobo, 2007), but rather than speaking about generalization one should think in terms of external validity (Kukul & Ganguli, 2012). We can say that qualitative findings will be externally valid for situations similar to the one in which the study was conducted. Hence, rather than talking about how generalizable the qualitative data is, the qualitative researcher is well advised to use forms of the word "indicative" and similar words such as "suggest." The qualitative researcher should say something like "the findings of this study are indicative of what one would find in other situations given similar characteristics." Or, "this study indicates that in other situations..." Or, "this study suggests that in other situations..." Using wording such as this highlights the importance of context, which, as per Cronbach, is something that even quantitative researchers should be heeding. The qualitative researchers are saying that these findings are likely to be valid for similar situations. It is then up to consumers of the research to judge to what extent the research findings are valid for the particular circumstances of interest to that consumer. Furthermore, we do not advise that qualitative researchers use the word generalization when addressing the limitations of their work. Again, it is the language of "indication" and "suggestion" that is appropriate. The true limitation is that qualitative findings are indicative only for situations having similar characteristics.

### **4. SAMPLE SIZE AND THE CONCEPT OF 'SATURATION'**

But we still haven't answered the question of how many to interview. The number does matter, though not for the reasons that numbers matter in quantitative work. Take for example an opinion survey where the subjects respond to items such as Likert items. The researcher needs a sample size ample enough to allow accurate estimation of how likely (probable) it is that people (of similar characteristics) will hold the opinions represented by the items. The situation is not much different with test scores. If achievement scores from treatment and control conditions are to be compared, researchers need numbers so that they can accurately estimate how likely (probable) it is that the outcome will be the same for other students (of similar characteristics). In contrast, an interview is used to determine what opinions are held by interviewees. Hence, you need to interview enough people so that you learn most if not all possible opinions (among people of similar characteristics). Of course, researchers often want to know which opinions are more popular or more frequent, but that's not the primary aim of qualitative work. Those questions are better answered quantitatively.

For qualitative interviewing there is a critical assumption: the number of unique opinions is not very large. For example, if we asked professors what they thought about working at their university there would be a limited number of opinions; from 100 professors you are not going to get 100 unique opinions. What you will find is that several opinions get repeated over and

over, which means that the researcher does not need to interview all 100 professors in order to discover all of the unique opinions in this group of people, especially not all the most common, unique opinions. Clearly, judgement is called for (see Baker & Edwards, 2012, for a variety of opinions). Here is a counter example. We were interested in how students understood a common claim about the nature of science: Scientific knowledge is durable but can change in light of new evidence or new perspectives. We particularly wanted to know how in this context students interpreted the word ‘durable.’ We reasoned that students could easily have more unique opinions than the number of students we could reasonably interview. Hence, we used a survey method; and the survey results validated our judgment: student opinions were many. No reasonable number of interviews would have so efficiently disclosed such a large number of opinions.

The research student we spoke of earlier, however, wanted to know what local parents thought of the new Next Generation Science Standards (NGSS) being implemented in the area schools. Knowing that parents did not have much experience with this new curriculum, he reasoned that there would be a limited number of unique opinions, and that these could be adequately identified by interviewing a subset of parents. He reasonably expected that as he went down his list of parents, a few opinions would begin reoccurring; because opinions on most topics do not run in the hundreds; they do not even run in the dozens. Unless a topic is vague, lacking focus, or poorly defined, there just are not that many distinct opinions that one could hold about most topics. The goal of qualitative interviewing is to capture most if not all of those opinions, however many opinions there are. And this is where the number of people needed for interviewing comes into question.

Clearly, the likelihood of capturing most if not all opinions increases with the number of people one interviews. The thing is, once you have captured the possible range of opinions, to whatever level of detail you seek, there is little reason to continue interviewing more people. You have reached “saturation” (Seidman, 2006). Interviewing more people will not result in more opinions because very likely there are no more opinions. The probability that a unique opinion exists is inversely related to how long it takes to find that unique opinion. But still we have to ask how many interviews are enough. One approach to deciding, and it is one that we’ve used, is that you don’t estimate ahead of time how many people to interview. You keep interviewing until you reach a point where you stop getting unique opinions and all that you are hearing is what you have heard from previous interviewees. At that point you interview perhaps one, two, or three more for insurance; but you have reached the number you need. In a Cobern, Gibson & Underwood (1999) study, the researchers quit at 16 interviews having reached saturation.

On the other hand, oftentimes for logistical reasons, time constraints, and financial ability to pay honoraria, a researcher must decide ahead of time the maximum number of people to interview. This is the situation in which many researchers find themselves, and it calls for judgment. Researchers have to consider how many opinions on any given topic the people of interest to them might hold. Is the topic like defining ‘durable’ in the context of science, or asking parents their opinion of a newly implemented science curriculum? Only two opinions? Three? Three to five? Could there be 10 distinct opinions on the topic of interest? The literature can help because it can suggest what opinions might be out there, but conventional wisdom (maybe we would even say common sense) is that for most well-defined topics there are not 10 unique opinions among similar people. If we assume that there will be no more than 10 unique opinions on most of the topics we would want to ask people about, then we have to ask how many people we would need to interview to get those 10 opinions. That is the question the researcher must answer. Rather, the researcher must estimate an answer for that question. That estimation gives you the number of people you should plan to interview. Conventional wisdom suggests that the number is between 15 and 20 insofar as the topic is of limited scope. It was a

good bet that the high schoolers' parents our doctoral student was interested in would have fewer than 20 unique opinions about NGSS, and that those opinions might or might not be equally common. By the 20th interview, he could expect to have reached saturation – and he did (Channell, 2019) (see Appendix for how this approach might be worded for a research proposal.)

Our point is that for qualitative interviewing, the number of people one plans to interview is not the first question that needs to be answered. For our graduate student, the important question was, how likely are parents of students, across the three grade bands, to have such differing opinions that the domain of unique opinions across the three grade bands exceeds the number of unique opinions in any one grade band. If it can be argued that grade band is unimportant, then his original plan was fine. On the other hand, if different grade bands are likely to result in different opinions, then six or seven interviews per grade band would not likely be enough to reach saturation per grade band, and too few would probably be too risky.

## **5. CONCLUSION**

All research requires judgement. It does not matter whether the research is quantitative or qualitative; judgment is required. Not even a power calculation can be run without judgment, because the input values are not self-evident. A good quantitative researcher describes the situation in which the research takes place and defends value judgments and assumptions. A qualitative researcher does the same. Deciding on how many subjects to interview is a value judgment and requires an explanation. We knew from other research that students very likely had a poor understanding of what it meant that scientific knowledge is 'durable.' Hence, we could reasonably expect that a large number of students would hold a number of unique opinions almost as large, thus making an interview approach not only impractical but nigh impossible. On the other hand, NGSS is a new curriculum in our area and parents simply had not had much time to form many opinions. Moreover, the focus on NGSS was specific to the science classrooms where the parents' children attended. An interview approach was reasonable. The choice between administering quantifiable surveys or conducting qualitative interviews does not usually require elaborate explanation. However, for interview work, we advise explaining the general basis for sample size, and also whether or not informational redundancy or saturation was achieved.

Finally, we urge qualitative interviewers to exchange the rhetoric of generalizing for the rhetoric of external validity. Some research is designed to simply provide specific and actionable information about the sample population. More often, consumers of research want to know whether qualitative findings are applicable to their own situation of interest or indicative of what might be the case in a different but similar situation. This is the judgment that consumers of research have to make and that they can only make if the original researchers adequately describe the context in which the research was conducted. If you're going to interview parents about their children's education, we need descriptive information about the parents and about the schools that their children attend. Only then can the consumer judge whether or not aspects of the findings are likely to be valid elsewhere, that is, judge to what extent the findings have external validity. Generalization, however, is a term best left for quantitative, probability-based research where, even then, generalizing applies to adequately similar situations or populations. Qualitative findings can be usefully indicative of what one might find in similar situations and contexts, and also of how different aspects/elements studied may relate to one another.

## Acknowledgements

Not applicable.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## ORCID

William W. Cobern  <http://orcid.org/0000-0002-0219-203X>

Betty AJ Adams  <http://orcid.org/0000-0002-8554-8002>

## 6. REFERENCES



- Baker, S. E., & Edwards, R. (2012). How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods Review Paper. Retrieved December 28, 2019 from [http://eprints.ncrm.ac.uk/2273/4/how\\_many\\_interviews.pdf](http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf)
- Channell, A. C. (2019). Teacher and Parent Perspectives on Alignment to The Next Generation Science Standards Following Teacher Professional Development. (PhD), Western Michigan University, Kalamazoo, MI.
- Cobern, W. W., Gibson, A. T., & Underwood, S. A. (1999). Conceptualizations of Nature: An Interpretive Study of 16 Ninth Graders' Everyday Thinking. *Journal of Research in Science Teaching*, 36(5), 541-564. [DOI.org/10.1002/\(SICI\)1098-2736\(199905\)36:5<541:AID-TEA3>3.0.CO;2-1](https://doi.org/10.1002/(SICI)1098-2736(199905)36:5<541:AID-TEA3>3.0.CO;2-1)
- Cronbach, L. J. (1975). Beyond the Two Disciplines of Scientific Psychology. *American Psychologist*, 30(2), 116-127. [DOI:10.1037/h0076829](https://doi.org/10.1037/h0076829)
- Ellis, P. D. (2010). Effect Size FAQs. Retrieved December 20, 2019 from <https://effectsizefaq.com/about/>
- Gobo, G. (2007). Sampling, representativeness and generalizability. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice*. SAGE Publications: Thousand Oaks, CA, p. 405-426. ISBN-13: 978-0761947769.
- Kukull, W. A., & Ganguli, M. (2012). Generalizability: the Trees, the Forest, and the Low-Hanging Fruit. *Neurology*, 78(23), 1886-1891. DOI:10.1212/WNL.0b013e318258f812.
- Royall, R. M. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician*, 40(4), 313-315. [DOI:10.2307/2684616](https://doi.org/10.2307/2684616)
- Seidman, I. E. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and The Social Sciences, 3rd Edition*. Teachers College Press: Columbia University, New York. ISBN-13: 978-0807746660.
- Teo, T. (2013) (Ed.). *Handbook of Quantitative Methods for Educational Research*. Sense Publishers: Rotterdam, The Netherlands. ISBN: 978-94-6209-404-8.

## **7. APPENDIX**

Here is an example of how this approach might be worded for a research proposal. For this example, we are indebted to our colleague Dr Brandy Pleasants.

Based on research with a similarly homogenous group it seems that about 10 participants is sufficient to cover all reasonable responses I might get. I therefore plan to interview no less than 10 participants, with a goal of 15 (even if saturation is reached); however, I also plan to continue interviewing if at 10 I still seeing variation in the data, continuing until I reached saturation.

## Comparison of Passing Scores Determined by The Angoff Method in Different Item Samples

Hakan Kara <sup>1,\*</sup>, Sevda Cetin <sup>2</sup>

<sup>1</sup>Ministry of National Education, 06930, Ankara, Turkey

<sup>2</sup>Hacettepe University, Faculty of Education, Measurement and Evaluation Department,06800, Ankara, Turkey

### ARTICLE HISTORY

Received: 01 October 2019

Revised: 14 February 2020

Accepted: 05 March 2020

### KEYWORDS

Standard-setting,  
Angoff,  
Random sampling methods,  
Minimum passing scores

**Abstract:** In this study, the efficiency of various random sampling methods to reduce the number of items rated by judges in an Angoff standard-setting study was examined and the methods were compared with each other. Firstly, the full-length test was formed by combining Placement Test 2012 and 2013 mathematics subsets. After then, simple random sampling (SRS), content stratified (C-SRS), item-difficulty stratified (D-SRS) and content-by-difficulty random sampling (CD-SRS) methods were used to constitute different length of subsets (30%, 40%, 50%, 70%) from the full-test. In total, 16 different study conditions (4 methods x 4 subsets) were investigated. In data analysis part, ANOVA analysis was conducted to examine whether minimum passing scores (MPSs) for the subsets were significantly different from the MPSs of the full-length test. As a follow-up analysis, RMSE and SEE (Standard Error of Estimation) values were calculated for each study condition. Results indicated that the estimated Angoff MPSs were significantly different from the full-test Angoff MPS (45.12) only in the study conditions of 30%-C-SRS, 40% C-SRS, 30% D-SRS and 30%-CD-SRS. According to RMSE values, the C-SRS method had the smallest error while the SRS method had the biggest one. Moreover, SEE examinations revealed that to achieve estimations similar to the full-test Angoff MPS (within one SEE), it is sufficient to get 50% of items with the C-SRS method. C-SRS method was the more effective one compared to the others in reducing the number of items rated by judges in MPS setting studies conducted with the Angoff method.

## 1. INTRODUCTION

Defined as the process of determining one or more passing scores in a test, standard setting has recently become necessary in order to make important decisions in many areas. These decisions include selection, classification, licensing or certification decisions in the fields such as health, law, and especially education. The accuracy of these decisions depends on the accurate specification of the measure (standard). The correct setting of the standard also depends on the selection and use of appropriate standard setting methods, in other words, it depends on effective monitoring of the process (Downing, 2006; Kane, 2001). There are more than 50 methods in the literature to set standards (Smith, 2011). Many studies have examined whether different methods give similar standards for the same exam and concluded that method selection

---

CONTACT: Hakan Kara ✉ [hakankaraodtu@gmail.com](mailto:hakankaraodtu@gmail.com) 📍 Ministry of National Education, 06930, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

has an effect on passing scores, and that different methods may produce different passing scores on the same exam (Berk, 1996; Çetin, 2011; Irwin, 2007; Jaeger, 1989; Kane, 1998; Mehrens, 1995). For this reason, the simultaneous use of multiple methods has been proposed in standard setting studies. While similar results support the acquired passing score, different results provide a suggestion to review the results (Cizek, 2001; Irwin, 2007).

In addition, Behuniak, Archambault, and Gable (1982) detected in their study comparing the Angoff and Nedelsky methods that standards do not only vary between methods but also between judges who use the same method. This result can be interpreted as that the same method can give different results on the same exam when standard setting methods based on judge opinion are used. In addition, Smith (2011) stated that standard setting methods that require judges to make judgments about a hypothetical individual can be cognitively exhausting for these individuals. The cognitive effort expected from judges has been an important source of criticism especially for the Angoff method (Lewis, Green, Mitzel, Baum, & Patz, 1998). Judges are expected to estimate the performance of the individual at the minimum competence level for each item and do the same thing for each performance level. Therefore, the procedures expected from judges can become time consuming, exhausting and cognitively challenging. As a result, if the number of questions to be assessed by a judge can be reduced, the judges will be able to make more accurate evaluations because they will evaluate less questions, resulting in less time consuming and tiring procedures (Ferdous & Plake, 2007; Smith, 2011).

Reducing the number of questions that judges will evaluate is possible in two different ways. The first one is to reduce the number of items in a standard setting study in a way which will form a subtest representing the whole test, thereby reducing the total number of items reviewed by judges (Buckendahl, Ferdous, & Gerrow, 2010; Ferdous & Plake, 2005; 2007). The second is to divide the test into smaller subtests representing the whole test and allocate an equal number of judge subgroups to evaluate these subtests (Norcini, Shea, & Ping, 1988; Plake & Impara, 2001; Sireci, Patelis, Rizavi, Dillingham, & Rodriguez, 2000). In the studies conducted in this way, the total number of items considered does not change while the number of items to be considered by each judge reduces.

In the light of the above given information, it is observed that standard setting is important in terms of forming the basis for decisions taken in education and that the accuracy of the decisions given depends on setting the right standard. Item reduction is recommended to be used, especially given that the standard setting processes using the Angoff and similar methods are very time consuming, very exhausting and require more cognitive effort. When the related literature is examined, it is seen that there are studies on reducing the number of items in standard setting studies using the Angoff method (Ferdous & Plake, 2005; Ferdous & Plake, 2007; Kannan, Katz, Sgammato, & Tannenbaum Katz, 2015; Plake & Impara, 2001; Smith, 2011); however, it was detected that in terms of reducing the number of items, studies which analyze the effectiveness of stratified random sampling methods (content stratified [C-SRS], difficulty stratified [D-SRS], content-difficulty stratified [CD-SRS], content-difficulty-discrimination stratified [CDD-SRS] etc.) are limited in number. Within the scope of the research, passing scores related to the Math sub-test of Placement Test were tried to be determined by using the Angoff method. Because of the low number of items, in this study, the sub-tests were created by considering only item difficulty indexes and content areas and it was analyzed to determine what percentage of test items could be sufficient to obtain similar predictions for the passing score of the whole test.

In this study, it is aimed to analyze the effectiveness of random sampling methods which can be used to reduce the total number of items evaluated in standard setting studies using the Angoff method and to compare different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], difficulty stratified [D-SRS], content-difficulty stratified

[CD-SRS]) with each other. The subtests were created by considering only content categories in C-SRS method; considering only difficulty categories in D-SRS method; and both content and difficulty categories in CD-SRS method.

In the process of standard setting, the recommended minimum passing scores are generally obtained in two rounds (Hambleton, 1998; Reckase, 2001). It may take a lot of time for judges to think of the student with the minimum qualification level and make individual estimates for each item in each round. This challenging and long process can prevent judges from making decisions in a healthy way. Therefore, it is thought that reducing the number of items to be evaluated by each judge will improve the scoring quality of judges (Ferdous & Plake, 2005). One of the methods of reducing the number of items is to create subtests which represent the whole test using random item sampling methods. In the scope of this study, Placement Test 2012 and Placement Test 2013 math tests were combined and a whole test containing 40 items were created and the subtests were derived from this test by using different random sampling methods. Minimum passing scores (MPS) for all tests and the subtests were determined and compared according to the Angoff method. In this way, effectiveness of different random sampling methods was also analyzed. From this point of view, it is thought that the study will provide important information to standard setting institutions and individuals about which method can be used especially in large scale exams. In addition to this, the study may give an idea as to what percentage of test items would be sufficient to obtain estimates similar to the passing score of a test; in this way, it is thought that reducing the number of questions will help judges to make healthier decisions by reducing their workloads.

To accomplish this purpose, the research questions are as follows:

1. Is there any significant difference between the Angoff passing scores for the whole test and the different subtests generated from the whole test, with respect to;
  - a. simple random sampling (SRS),
  - b. content stratified random sampling (C-SRS),
  - c. item difficulty stratified random sampling (D-SRS), and
  - d. content and item difficulty stratified random sampling (CD-SRS) method?
2. How do the average Angoff passing scores differ for each subtest generated by different random sampling methods?

## 2. METHOD

Research models which do not have any intervention affecting variables and analyze the relationship between two or more variables are relational type of research models (Fraenkel, Wallen, & Hyun, 2012). In this research, a variety of random sampling methods that can be used to reduce the number of items in a standard setting study are compared. In this respect, the research is one of relational research models. At the same time, this study is a descriptive study in terms of obtaining descriptive statistics related to the Angoff method.

### 2.1. Research Population and Sample

The research population consisted of 1,075,533 and 1,112,604 8<sup>th</sup> grade students who took the Placement Test in 2011-2012 and 2012-2013 academic years, respectively. The sample of the study consisted of two different groups as being students and judges. In the student group, a total of 20611 students were selected by random sampling method among the students who entered the Placement Test 2012 and Placement Test 2013 as being 10,187 and 10,424 students respectively; and in the judge group, a total of 28 judges including 12 academicians and 16 secondary school math teachers were included. In this study, goal-oriented sampling method was used to determine the judges and voluntariness was taken as the basis for their selection. In goal-oriented sampling method, researchers can use their personal evaluations to form a



sample according to the prior knowledge of the study group and the purpose of the research (Fraenkel, Wallen, & Hyun, 2012). The academicians in the judge group were selected among the academicians who graduated from the undergraduate programs of Elementary Mathematics Education and have postgraduate education in the fields of Educational Sciences or Mathematics Education; and the teachers in the judge group were selected among the secondary school mathematics teachers with at least five years of experience in the profession. Within the scope of the study, each judge evaluated 40 items in accordance with the Angoff method, and the Angoff passing scores were calculated by considering these evaluations.

## **2.2. Data Collection Process**

Student answers to the Placement Test 2012 and Placement Test 2013 math subtests used in the study were obtained from the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education. The evaluations of the judges on the items were collected with the data collection form prepared by the researcher. Data collection from judges was conducted by the researchers herself.

## **2.3. Data Collection Tools**

Two different data were used in this research. Student data used in the research are the data about the results of the exams (Placement Test 2012 and Placement Test 2013) applied by the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education; the other data were obtained from 28 judges through the “Volunteer Participation Form for Judge Opinions” which was prepared by the researcher.

**Placement test.** The exams conducted by the Ministry of National Education for the transition to secondary education have varied in terms of method and content over the years. These exams, which were held under different names until 2008, were conducted under the name of Placement Test for the 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> grades from the 2008-2009 academic year to the 2013-2014 academic year. Between these years, it was gradually implemented in all classes and was gradually abolished in the following years. In this study, student answers to the 8<sup>th</sup> grade math subtests (20 questions for each test) of Placement Tests which were conducted in 2011-2012 and 2012-2013 academic years were used.

**Volunteer Participation Form for Judge Opinions.** Volunteer Participation Form for Judge Opinions was prepared in order to set the passing score. In this form, the definition of “minimum proficiency level” is clearly defined based on level 2 (PISA, 2007) in mathematics proficiency levels of International Student Assessment Program. The judges were asked to carefully examine the multiple-choice test questions before starting the assessment, consider what percentage of students with minimum qualification would answer the question correctly, and estimate a percentage value for each item separately. The individual passing score of each judge was calculated by converting the scores obtained by adding the difficulty estimations determined by the judges for each item to the 100-point scale, and the final passing score of the test was determined by the average of the individual passing scores.

## **2.4. Data Analysis**

The whole test which includes 40 items was formed by combining math subtests of the Placement Tests 2012 and 2013. Considering that the items in both tests measure the same gains and that they are equivalent in terms of skill levels of the group that took the exam in two years, it was not inconvenient to combine the two tests. During the analysis of data, firstly, the student answers for the items of the two different tests (the Placement Tests 2012 and 2013) were converted to 1-0 data by coding “1” for correct answers and “0” for incorrect and blank answers, and then the test and item statistics were calculated. Passing score of the whole test was calculated in accordance with the Angoff method.

Four different subtests, which are considered to represent the whole test in terms of passing score, (30%, 40%, 50%, and 70% of the total item number) were created by using four different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], item difficulty stratified [D-SRS], content-item-difficulty stratified [CD-SRS]). In other words, 4x4 pattern including four different random sampling methods and four different subtests were used, and 16 study cases were examined in total. During the creation of the subtests, 1000 replications were done for each study case. For example, for the 30% subtest, which consists of 30% of 40 items with simple random sampling method, item selection procedures were repeated 1000 times and 1000 different subtests and passing points were obtained. Thus, the results obtained were tried to be more consistent and reliable.

One-way analysis of variance (ANOVA) was performed to determine whether there was a significant difference between the passing score of the whole test and the mean passing scores of the subtests. Assumptions of normality and homogeneity of variances were checked for each ANOVA analysis. It was observed that normality assumption was provided for each case. The kurtosis and skewness values of all cases were within the range of (-2, 2), and Shapiro-Wilk  $p$  value was greater than .05. The assumption of homogeneity of variances was checked by Levene test and it was found that the assumption could not be provided for any case ( $p < .05$ ). In this case, the results of Welch test which was suggested to be used (Pallant, 2005) were applied.

Root Mean Square Error (RMSE) and SEE values were used in order to make more detailed analyses. While interpreting SEE values, it was examined that what percentage of passing scores of the 1000 subtests created remained within the total test passing score  $\pm 1$  SEE. At this point, 95% was taken as the similarity criterion and it was interpreted that if more than 5% of 1000 passing scores were out of range, no result similar to the whole test passing score was obtained.

**Test and item statistics.** Descriptive statistics of the Placement Test 2012 and Placement Test 2013 Math subtests are presented in Table 1. The statistics were calculated considering the answers of 10187 and 10424 students who took the exam in 2011-2012 and 2012-2013 academic years, respectively.

**Table 1.** Descriptive statistics of the math subtests

	Placement Test 2012 Math Subtest	Placement Test 2013 Math Subtest
Number of Items	20	20
Number of students	10187	10424
Mean	6.41	4.97
Variance	19.35	16.89
Standart Deviation	4.40	4.11
Reliability (KR20)	.84	.83
Average Difficulty	.32	.25

Values in the table show that the Placement Test 2012 Math subtest ( $\bar{X}=6.41$ ) and Placement Test 2013 Math subtest ( $\bar{X}=4.97$ ) are difficult. Average difficulty of the math subtests was calculated as .32 and .25, respectively. The reliability of the tests whose results are used to make important decisions, should be .80 and over when the number of items is low (Özçelik, 2013). The reliability coefficients of the mathematics subtests were .84 and .83. Accordingly, it can be stated that the scores for these tests are reliable.

Difficulty values ( $p$ ) of 40 items in total, as being items between 1 and 20 are from the Placement Test 2012 Math subtest and items between 21 and 40 are from the Placement Test 2013, are given in Table 2.

**Table 2.** Item difficulty indices for the whole test

Item No.	$p$	Item No.	$p$	Item No.	$p$	Item No.	$p$
1	.28	11	.24	21	.20	31	.20
2	.39	12	.30	22	.43	32	.22
3	.41	13	.23	23	.22	33	.19
4	.63	14	.17	24	.22	34	.41
5	.18	15	.53	25	.32	35	.21
6	.25	16	.53	26	.13	36	.20
7	.40	17	.34	27	.27	37	.23
8	.34	18	.20	28	.30	38	.14
9	.30	19	.43	29	.18	39	.49
10	.13	20	.15	30	.25	40	.19
Mean	.29						

Item difficulty values for 40 items of the whole test formed by combining the Placement Test 2012 and 2013 Math subtests ranged from .13 to .63. Accordingly, it is observed that the test has difficult and moderately difficult items but not easy items. While the most difficult items ( $p = .13$ ) of the test were items 10 and 26, the easiest item is item 4 ( $p = .63$ ). The overall average difficulty of the whole test was calculated as .29 and it can be said to be a difficult test.

**Formation of the subtests.** Simple random and stratified random sampling methods were used to create the subtests which were considered to represent the whole test in terms of passing score. For each sub-problem, 30%, 40%, 50% and 70% of the total 40 items were selected, and four separate subtests containing 12, 16, 20, 28 items were created with 1000 replications, respectively.

**Simple random item sampling method.** In the subtests created using this method, the items were randomly selected from 40 items.

**Stratified random item sampling.** In the stratified random sampling method, the items were selected according to content and item difficulty categories when the subtests were created. In this context, content-stratified random sampling (C-SRS), item difficulty stratified random sampling (D-SRS), and content and item difficulty stratified random sampling (CD-SRS) were used.

All test items are divided into 5 categories according to their contents by the field judge before the item selection for the subtests by C-SRS method; learning areas (numbers, geometry, measurement, probability and statistics, algebra) in secondary school mathematics curriculum were taken into consideration while determining the categories (MEB, 2009). In the selection of the items for the subtests, content categories were used as strata, and the items selected for the subtests were randomly selected from each category, proportional to the total number of items in each content category of the test. The categories of all test items which were classified with regard to their contents, and the figures and number of the items in each category are presented in [Table 3](#).

According to [Table 3](#), 22.5% of all test items are in numbers, 25% in geometry, 17.5% in measurement, 12.5% in probability and statistics, and 22.5% in algebra category. In the light of this information, it was ensured that the items selected for the subtests were also in the same proportions in each category. For example, in order to create a 20-item subtest, 5 items from numbers, 5 items from geometry, 3 items from measurement, 2 items from probability and statistics, and 5 items from algebra were randomly selected. The number of items that are expected to be selected for the subtests according to content categories is given in Appendix-J.

**Table 3.** Figures and number of items in content categories of the whole test

Content Categories	No. of Items	Item No.
Numbers	9 (%22.5)	10, 18, 26, 33, 1, 2, 25, 3, 22
Geometry	10 (%25)	21, 36, 6, 7, 9, 23, 28, 35, 4, 39
Measurement	7 (%17.5)	14, 38, 11, 12, 13, 30, 32
Probabilityandstatistics	5 (%12.5)	17, 24, 15, 16, 34
Algebra	9 (22.5)	5, 20, 29, 31, 40, 8, 27, 37, 19

In the item difficulty stratified random sampling (D-SRS) method, firstly, all the test items were divided into 3 categories according to their difficulty values; the items with difficulty parameters in the range of .00-.20 were included in Category 1, the items in the range of .20-.40 were included in Category 2, and the items in the range of .40-.63 were included in Category 3. Difficulty categories were used as strata in subtest item selection, and the items selected for the subtests were randomly selected from each category as being proportional to the total number of items in each difficulty category of the whole test. The item parameter value ranges of the categories and the figures and number of the items in each difficulty category of the whole test are given in [Table 4](#).

**Table 4.** Figures and number of items in difficulty categories of the whole test

Difficulty Categories	No. of Items	Item No.
Category 1 ( $.00 < p \leq .20$ )	13 (%32.5)	5, 10, 14, 18, 20, 21, 26, 29, 31, 33, 36, 38, 40
Category 2 ( $.20 < p \leq .40$ )	19 (%47.5)	1, 2, 6, 7, 8, 9, 11, 12, 13, 17, 23, 24, 25, 27, 28, 30, 32, 35, 37
Category 3 ( $.40 < p \leq .63$ )	8 (%20)	3, 4, 15, 16, 19, 22, 34, 39

When [Table 4](#) is examined, it is observed that 32.5% of the items are in Category 1, 47.5% are in Category 2, and 20% are in Category 3 according to item difficulty values. In this case, the difficulty distribution of the selected items to a sub-test which was desired to be formed is also ensured to be the same as the whole test. For example, to create a 20-item subtest, 6 items from Category 1, 10 items from Category 2, and 4 items from Category 3 were randomly selected.

In the content and item difficulty stratified sampling method (CD-SRS), the items were selected considering both content areas and difficulty values. The items were first divided into 5 categories according to their content, then the items in each content category were divided into 3 groups according to their difficulties and 15 strata were formed in total. The items selected for the sub-tests were randomly selected from each content-difficulty stratum in proportion to the total number of items in each stratum of the whole test. The figures and number of the items in each content-difficulty stratum of the whole test are given in [Table 5](#).

In [Table 5](#), each content category was divided into groups according to difficulty values and 15 separate strata were formed. When the distribution of the items in the table is examined, it is seen that the highest number of items is found in Geometry-Group2 (6 items) and the least number of items is found in Algebra-Group3 (1 item). In addition, no items were included in Measurement-Group3 and Probability and Statistics-Group1 levels. It was attempted to ensure that the items selected for the sub-tests were proportional to represent the distribution of items in these 13 strata of the whole test. For example; since 4 (10%) of the 40 items in the whole test are in Numbers-Group1, 10% of the total number of items to be selected for each subtest was selected randomly from the items in the Numbers-Group1 stratum.

**Table 5.** Figures and number of items in content-difficulty strata of the whole test

Content Categories	Difficulty Categories	No. of Items	Item No.
Numbers	Group 1 (.00<p≤.20)	4 (%10)	10, 18, 26, 33
	Group 2 (.20<p≤.40)	3 (%7.5)	1, 2, 25
	Group 3 (.40<p≤.63)	2 (%5)	3, 22
Geometry	Group 1 (.00<p≤.20)	2 (%5)	21, 36
	Group 2 (.20<p≤.40)	6 (%15)	6, 7, 9, 23, 28, 35
	Group 3 (.40<p≤.63)	2 (%5)	4, 39
Measurement	Group 1 (.00<p≤.20)	2 (%5)	14, 38
	Group 2 (.20<p≤.40)	5 (%12.5)	11, 12, 13, 30, 32
	Group 3 (.40<p≤.63)	0	
Probability and statistics	Group 1 (.00<p≤.20)	0	
	Group 2 (.20<p≤.40)	2 (%5)	17, 24
	Group 3 (.40<p≤.63)	3 (%7.5)	15, 16, 34
Algebra	Group 1 (.00<p≤.20)	5 (%12.5)	5, 20, 29, 31, 40
	Group 2 (.20<p≤.40)	3 (%7.5)	8, 27, 37
	Group 3 (.40<p≤.63)	1 (%2.5)	19

**Setting and interpretation of passing scores.** A whole test consisting of 40 items was created by combining the Placement Test 2012 and Placement Test 2013 Mathematics subtests, and during the process of setting the passing score for the whole test through the Angoff method, firstly, whether there was a concordance between judges were analyzed by Kendall's *W* coefficient of concordance. In this non-parametric technique, Kendall's coefficient of concordance is calculated by the following formula (Siegel, 1956).

$$W = \frac{12\sum R_i^2 - 3k^2N(N + 1)^2}{k^2N(N^2 - 1)}$$

In the equation;

*k*: represents the number of raters, *N*: represents the number of items rated, *R*: represents the sum of the scores given by all raters for each item.

For the cases in which the number of raters is equal to seven or more,  $\chi^2$  is used; and  $\chi^2_{(N-1)} = k(N - 1)W$  value shows the distribution of  $\chi^2$  in *N*-1 degree of freedom (Siegel, 1956).

Afterwards, the passing score was calculated for the whole test by the Angoff method. The individual passing score of each judge was calculated by converting the scores obtained by summing the difficulty estimations determined by the judges for each item to the 100-point scale, and the final passing score of the test was obtained by averaging the individual passing scores. The final passing score of the test was determined by taking the average of the individual passing scores. Passing score for each subtest with regard to the Angoff method was calculated by following the above given steps and placed on the same scale with the whole test. One-way ANOVA was carried out in order to determine whether there was a significant difference between the scores of the whole test and the scores of the sub-tests.

In the light of the above analyzes, RMSE and SEE reviews were performed with the aim of analyzing the generalizability of the results. In this way, the effectiveness of different sampling methods in reducing the number of items was examined and more systematic and stable findings were tried to be obtained about the generalizability of the results.

SEE for each passing score is calculated with the following formula.

$$TSH = \frac{SS}{\sqrt{N}}$$

In this formula;

SEE: Standard error

SS: Standard deviation of individual passing scores of judges

N: The number of judges.

RMSE values which were analyzed in addition to SEE were calculated with the help of the following formula.

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (\hat{X} - X_j)^2}{N}}$$

In this formula;

$\hat{X}$ : Minimum passing score of the whole test,  $X_j$  : Passing score of j. replication, N: Total replication number (=1000).

### 3. RESULT

According to the Angoff method, 28 judges made assessments for each item in the test. Before calculating the minimum passing scores (MPS) of the whole test, the consistency between the judges was analyzed by Kendall's coefficient of concordance. According to conducted analyses, Kendall's coefficient of concordance was found to be .30 ( $\chi^2=322.99$ ,  $sd=39$ ,  $p<.05$ ). This result shows that there is a significant concordance among the judges.

*Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the SRS method?*

In the analyses of this sub-problem, first of all, the minimum passing score (MPS) of the entire test consisting of 40 items was calculated. Later on, 1000 replications were performed for each subtest (30%, 40%, 50% and 70%). Thus, it was aimed to increase the consistency and reliability of the MPS values obtained from the subtests. Descriptive statistics of the MPSs of 1000 replications for each subtest by simple random sampling (SRS) method are given in [Table 6](#).

**Table 6.** Descriptive statistics of the Angoff MPSs of the subtests formed by SRS

SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	45.03	2.47	38.38	52.94
%40	16	45.15	1.96	38.94	51.36
%50	20	45.15	1.51	40.76	49.76
%70	28	45.16	1.02	41.82	48.13
Whole Test	40	45.12			

As observed in [Table 6](#), the MPS for the whole test (40 items) was calculated as 45.12 according to the Angoff method. When the values related to the subtests were examined, the mean MPS of the 1000 different subtests, which were formed through the selection of 30% (12 items) of the items by simple random method, was found to be 45.03 and the standard deviation was found to be 2.47. It was observed that MPSs of these tests ranged between 38.38 and 52.94 points. While MPSs of the 16-item subtests formed by selecting 40% of the items varied

between 38.94-51.36, the mean was found to be 45.15 and standard deviation was found to be 1.96. When 50% of the items were randomly selected 1000 times according to the SRS method, the mean of the subtests was calculated as 45.15 and the standard deviation was calculated as 1.51. The MPSs of these 20-item subtests ranged from 40.76 to 49.76 points. In addition, when 70% of the whole test items were selected, the mean of the sub-tests was found to be 45.16 and the standard deviation was found to be 1.02. MPSs of these subtests including 28 items ranged from 41.82 to 48.13. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, as the number of items increased, MPSs of the subtests approximated the MPS of the whole test.

The normality and homogeneity assumptions of variances were tested before the ANOVA analyses were carried out to determine whether MPS means for all tests and the subtests differ significantly from each other. It was observed that the assumption of normality was achieved but the assumption of homogeneity of variances was violated; therefore, Welch test results were examined. The acquired results showed that there was no significant difference between the MPSs of the subtests ( $F(4,1998) = 0.824, p = .510$ ).

SEE and RMSE values were reviewed in order to conduct a more detailed analysis and to determine the subtest that best represents the whole test in terms of MPS.

**Table 7.** RMSE and subtest percentages of the Angoff MPSs of the subtests formed by SRS

SRS	RMSE	Subtest Percentages	
		Mean±1SEE	Mean±2SEE
%30	2.47	%70.1	%96.4
%40	1.96	%81.3	%99.3
%50	1.51	%91.5	%100
%70	1.02	%98.6	%100
Whole Test	SEE = 2.58 Mean±1SEE= (42.54 - 47.70) Mean±2SEE = (39.96 – 50.28)		

As observed in Table 7, the value of standard error of the estimate (SEE) for the whole test was calculated as 2.58. The values given in the percentage of subtest column give information about what percentage of MPSs of 1000 different subtests generated for each subtest remained within the specified limits. According to this information, the percentages of MPS remained in 1 SEE and 2 SEE values of the passing score of the whole test were the lowest for the 30% test and the largest for the 70% test.

When RMSE values were analyzed, as expected, error value decreased with the increase in the number of items. The lowest error was obtained from the 70% subtest (1.02) and the most error was obtained from the 30% subtest (2.47). Additionally, only the Angoff MPSs of the 70% subtests remained within the desired criteria (within 1 SEE of the final passing score with at least 95% possibility). Absolute values of the difference between MPSs of the subtests formed by 70% of the items and the MPS of the whole test were less than 1 SEE in 98.6% of the 1000 subtests. As a result, the use of at least 70% of the test items can be suggested through SRS method in order to obtain a passing score similar to the MPS of the whole test.

*Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the C-SRS method?*

In Table 8, descriptive statistics related to 1000 replications generated for each subtest by content stratified random sampling (C-SRS) are given.

**Table 8.** Descriptive statistics of the Angoff MPSs of the subtests formed by C-SRS

C-SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	44.81	2.05	37.92	50.93
%40	16	45.43	1.66	40.25	49.89
%50	20	45.08	1.37	41.42	49.11
%70	28	45.20	0.86	42.63	48.16
Whole Test	40	45.12			

When [Table 8](#) is analyzed, it is seen that the average of the 1000 different subtests created by selecting 30% of the items according to the C-SRS method is 44.81 and the standard deviation is 2.05. It was observed that the MPSs of these tests ranged from 37.92 to 50.93 points. While the MPSs of the 16-item sub-tests formed by selecting 40% of the items ranged between (40.25-49.89), the mean was found to be 45.43 and standard deviation was found to be 1.66. When 50% of the test items were selected 1000 times, the mean of the generated subtests was 45.08 and the standard deviation was 1.37. The MPSs of these subtests ranged from 41.42 to 49.11 points. In addition, when 70% of the test items were selected, the mean of the generated subtests was 45.20 and the standard deviation was 0.86. The MPSs of these subtests varied between 42.63 and 48.16 points. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all tests and the subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. According to the assumption controls made before ANOVA analysis, normality assumption was provided, but homogeneity of variances was violated. Therefore, the Welch test results were interpreted, and the analysis results showed that the means of MPS differed significantly between the tests ( $F(4,1998) = 16.866, p=.000$ ).

Multiple comparison (Post-Hoc) was performed to determine which subtests' MPS means differed significantly from the means of MPS of the whole test. According to the comparison results, MPS means of both the 30% ( $\bar{X} = 44.81$ ) and the 40% ( $\bar{X} = 45.43$ ) subtests were significantly different from the MPS of the whole test ( $\bar{X} = 45.12$ ) ( $p < .05$ ). Therefore, it cannot be interpreted that 30% and 40% subtests formed by the C-SRS method represent the whole test in terms of passing score. In addition, the percentage values and RMSE values of the subtests remaining within the limits determined for each subtest (mean  $\pm$  1SEE; mean  $\pm$  2SEE) were calculated and the values obtained are presented in [Table 9](#).

**Table 9.** RMSE and subtest percentages of the Angoff MPSs of the subtests formed by C-SRS

C-SRS	RMSE	Subtest Percentages	
		Mean $\pm$ 1SEE	Mean $\pm$ 2SEE
%30	2.07	%78.9	%99.3
%40	1.69	%86.8	%100
%50	1.37	%95.0	%100
%70	0.86	%99.9	%100
Whole Test	SEE = 2.58 Mean $\pm$ 1SEE = (42.54 - 47.70) Mean $\pm$ 2SEE = (39.96 - 50.28)		



According to Table 9, the percentages of MPS in the 1 SEE and 2 SEE values of the passing score of the whole test were the lowest for the 30% test and the largest for the 70% test. In addition to this, the Angoff MPSs for only the 50% and 70% subtests were within the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). MPSs of 95% and 99.9% of the 1000 different subtests, including 50% and 70% of the test items respectively, were within 1 SEE of the MPS of the whole test. In addition, the MPSs of almost all sub-tests of different sizes was within 2 SEE values of the MPS of the whole test. When RMSE values were analyzed, as expected, error value decreased with the increase in the number of items. Error value was found to be 1.37 for the 50% test and 0.86 for the 70% subtest. In the light of this information, it is recommended to use at least 50% of the test items with C-SRS method in order to obtain a passing score similar to the MPS of the whole test.

*Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the D-SRS method?*

In Table 10, descriptive statistics related to 1000 replications generated for each subtest by item-difficulty stratified random sampling (D-SRS) are given.

**Table 10.** Descriptive statistics of the Angoff MPSs of the subtests formed by D-SRS

D-SRS	No. of Items	MPS mean	Standard deviation	Minimum	Maximum
%30	12	44.85	2.20	38.50	51.16
%40	16	45.00	2.19	39.79	50.10
%50	20	45.13	1.43	40.36	49.81
%70	28	45.32	0.94	42.32	48.04
Whole Test	40	45.12			

When Table 10 is analyzed, it is seen that the average of the 1000 different subtests created by selecting 30% of the items (12 items) according to the D-SRS method is 44.85 and the standard deviation is 2.20. It was observed that MPSs of these tests ranged between 38.50 and 51.16. While MPSs of the 16-item subtests created by selecting 40% of the items ranged between (39.79-50.10), the mean was found to be 45.00 and the standard deviation was found to be 2.19. When 50% of the test items were selected 1000 times in accordance with D-SRS method, the mean of the generated subtests was 45.13, and the standard deviation was 1.43. MPSs of these 20-item subtests ranged between 40.36 and 49.81. Also, when 70% of the whole test items were selected, the mean of the generated subtests was found to be 45.32 and the standard deviation was found to be 0.94. MPSs of these 28-item subtests ranged between 42.32 and 48.04. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all tests and the subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. In the assumption controls made before ANOVA, it was seen that homogeneity of variances was not provided; for this reason, the results of Welch test were interpreted. Analysis results revealed that the means of MPS differed significantly between the tests ( $F(4,1998)=16.110, p=.000$ ).

Multiple comparison (Post-Hoc) was performed to determine which subtests' MPS means differed significantly from the means of MPS of the whole test. According to the comparison results, only MPS mean of the 30% subtests ( $\bar{X} = 44.85$ ) was significantly different from the MPS of the whole test ( $\bar{X} = 45.12, p < .05$ ). Therefore, it cannot be interpreted that the 30% subtests formed by the D-SRS method represent the whole test in terms of passing score.

In addition to above given analyses, the percentage values and RMSE values of the sub-tests remaining within the limits determined for each subtest (mean  $\pm$  1SEE; mean  $\pm$  2SEE) were calculated and the values are presented in [Table 11](#).

**Table 11.** RMSE and subtest percentages of the Angoff MPSs of the subtests formed by D-SRS

D-SRS	RMSE	Subtest Percentages	
		Mean $\pm$ 1SEE	Mean $\pm$ 2SEE
%30	2.21	%74.5	%98
%40	1.67	%87.8	%99.9
%50	1.43	%93.5	%100
%70	0.96	%99.1	%100
WholeTest	SEE = 2.58 Mean $\pm$ 1SEE = (42.54 - 47.7) Mean $\pm$ 2SEE = (39.96 - 50.28)		

According to [Table 11](#), the percentages of MPS of the whole test remaining within 1 SEE value increased as the number of items increased. However, only the Angoff MPSs of 70% subtests provided the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). The absolute value of the difference between MPSs of 1000 subtests formed by 70% of the items and the MPS of the whole test was less than 1 SEE in 99.1% of subtests. In addition to this, MPSs of almost all subtests with different sizes differed by no more than 2 SEE from the MPS of the whole test. As expected, RMSE error value decreased as the number of items increased. The least error was acquired in 70% subtest (0.96), and the most error was acquired in 30% subtest (2.21). As a result, the use of at least 70% of the test items can be suggested through D-SRS method in order to obtain a passing score similar to the MPS of the whole test.

*Is there a significant difference between the Angoff passing scores of the whole test and the different subtests generated from the whole test according to the CD-SRS method?*

In [Table 12](#), descriptive statistics related to 1000 replications generated for each subtest by content and item-difficulty stratified random sampling (CD-SRS) are given.

**Table 12.** Descriptive statistics of the Angoff MPSs of the subtests formed by CD-SRS

CD-SRS	No. of Items	MPS Mean	Standard deviation	Minimum	Maximum
%30	12	44.66	2.14	39.02	51.07
%40	16	45.12	1.64	40.67	50.02
%50	20	45.13	1.44	41.57	49.81
%70	28	45.16	0.93	41.93	48.05
WholeTest	40	45.12			

According to [Table 12](#), the average of the 1000 different subtests created by selecting 30% of the items (12 items) according to the CD-SRS method was found to be 44.66 and the standard deviation was found to be 2.14. It was observed that MPSs of these tests ranged between 39.02 and 51.07. While MPSs of 16-item subtests created by selecting 40% of the items ranged between 40.67-50.02, the mean was found to be 45.12 and the standard deviation was found to be 1.64. When 50% of the test items were selected 1000 times in accordance with CD-SRS method, the mean of the generated subtests was found to be 45.13, and the standard deviation was found to be 1.44. MPSs of these 20-item subtests ranged between 41.57 and 49.81. Also, when 70% of the whole test items were selected, the mean of the generated subtests was found

to be 45.16 and the standard deviation was found to be 0.93. MPSs of these 28-item subtests ranged between 41.93 and 48.05. Considering the standard deviation and minimum-maximum values of the subtests of different sizes, the MPSs of the subtests approximated the MPS of the whole test as the number of items increased.

Whether all the tests and subtests differed significantly in terms of passing score means was analyzed by one-way ANOVA. It was observed that the assumption of normality was achieved but the assumption of homogeneity of variances was violated. Therefore, the interpreted Welch test results revealed that MPS means significantly differed between tests ( $F(4,1998) = 12.133$   $p=.000$ ).

According to the results of the conducted multiple comparison (Post-Hoc), only the MPS mean of 30% ( $\bar{X} = 44.66$ ) subtests was significantly different from the MPS of the whole test ( $\bar{X} = 45.12$ ) ( $p < .05$ ). Therefore, it cannot be interpreted that 30% subtests formed by the CD-SRS method represent the whole test in terms of passing score.

With the aim of having a more detailed analysis, the percentage values and RMSE values of the sub-tests remaining within the limits determined for each subtest (mean  $\pm$  1SEE; mean  $\pm$  2SEE) were calculated and the values are presented in [Table 13](#).

**Table 13.** RMSE and subtest percentages of the Angoff MPSs of the subtests formed by CD-SRS

CD-SRS	RMSE	Subtest Percentages	
		Mean $\pm$ 1SEE	Mean $\pm$ 2SEE
%30	2.19	%84.9	%98.3
%40	1.64	%88.3	%100
%50	1.44	%93.0	%100
%70	0.93	%99.5	%100
Whole Test	SEE = 2.58 Mean $\pm$ 1SEE = (42.54 - 47.7) Mean $\pm$ 2SEE = (39.96 - 50.28)		

According to the table given above, only the Angoff MPSs of 70% subtests provided the desired criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility). The MPSs of almost all 1000 different subtests, each containing 70% of the total number of items, differed by no more than 1 SEE from the MPS of the whole test. In addition, almost all the MPSs of different size subtests were within 2 SEEs of the MPS of the whole test. When RMSE error values were checked, RMSE error value decreased as the number of items increased. The lowest error was obtained from 70% subtest (0.93) and the most error was obtained from 30% subtest (2.19). Accordingly, the use of at least 70% of the test items can be suggested through CD-SRS method in order to obtain a passing score similar to the MPS of the whole test.

*How do the Angoff passing scores differ from each other for subtests generated by different random item sampling methods?*

In order to compare different random sampling methods, 50% and 70% subtests which gave results similar to the passing score of the whole test were chosen. The percentage of passing scores and RMSE values which are within the 1 SEE difference from the average for different sampling situations are given in [Table 14](#). As can be understood from the above given table, as expected, RMSE error value decreased as the number of items increased in each sampling method. Apart from that, it is seen that RMSE error values are lower for SRS (Stratified Random Sampling) method compared to SRS (Simple Random Sampling) methods. The lowest error values (1.37;0.86) for subtests with the size of 50% and 70% of test items were acquired through C-SRS method.

**Table 14.** RMSE and subtest percentage values of sampling status of the Angoff MPSs

	RMSE				Subtest Percentage			
					Mean $\pm$ 1SEE			
	SRS	C-SRS	D-SRS	CD-SRS	SRS	C-SRS	D-SRS	CD-SRS
%50	1.51	1.37	1.43	1.44	%91.5	%95.0	%93.5	%93.0
%70	1.02	0.86	0.96	0.93	%98.6	%99.9	%99.1	%99.5
WholeTest	SEE = 2.58				Mean $\pm$ 1SEE= (42.54 - 47.7)			

When the percentage rates of the MPS of the whole test within 1 SEE (42.54 - 47.70) are analyzed, it is seen that the lowest one was acquired through SRS (Simple Random Sampling) and the highest rate was acquired through C-SRS (Stratified Random Sampling). Therefore, it can be stated that C-SRS method is more effective in acquiring the passing score which is the closest to the MPS of the whole test. Considering the additionally applied similarity criterion (within 1 SEE of the MPS of the whole test with at least 95% possibility), it is observed that again C-SRS method is more effective.

In order to obtain a cut-off score that is similar to the MPS of the whole test, it is enough to select 50% of the total number of the items with C-SRS method while at least 70% should be selected with SRS (Simple Random Sampling), D-SRS and CD-SRS methods. As a result, content stratified random sampling method (C-SRS) can be a more effective method in the selection of the items for subtest/tests which are expected to represent the whole test in terms of MPS.

#### 4. DISCUSSION and CONCLUSION

In this study, the effectiveness of random sampling methods which can be used to reduce the total number of items evaluated in standard setting studies using the Angoff method was analyzed and different random sampling methods (simple random sampling [SRS], content stratified [C-SRS], item-difficulty stratified [D-SRS], content-and-item-difficulty stratified [CD-SRS]) were compared with each other. Within the scope of the study, four different item sampling methods (SRS, C-SRS, D-SRS and CD-SRS) were used to create sub-tests with different sizes (30%, 40%, 50% and 70%) from the whole test, and 16 study status (4 methods x 4 subtest) were evaluated in total. As a result of this study, the mean Angoff passing scores of the 30% and 40% subtests formed by the C-SRS method, and the 30% subtests formed by the D-SRS and CD-SRS methods, differed significantly from the whole test. In contrast, no significant level of difference was observed in the other 12 cases. These results comparatively support the findings of Kannan, Sgammato, Tannenbaum and Katz (2015). Kannan et al. (2015) reported that the predicted mean Angoff MPSs did not change much for different sampling methods or different subtests. In his study, only the mean Angoff MPS of the subtest containing 30 items (approximately 30%) and generated by the CD-SRS method differed from that of the whole test.

In addition, this study indicated that stratified random sampling methods are more effective than simple random sampling method in terms of giving similar MPS estimations. This finding was in agreement with the similar studies (Ferdous & Plake, 2005, 2007; Kannan et al., 2015, Smith, 2011). However, the finding that the content stratified method (C-SRS) is more effective than the other methods contradicts with the finding of Ferdous and Plake (2005) and Kannan et al. (2015) studies. They found that content-difficulty stratified method (CD-SRS) is more effective than content stratified method (C-SRS) and difficulty stratified method (D-SRS).

More importantly, the results of this study suggest that it is sufficient to select 50% of the items with C-SRS method to obtain very similar estimates (at least 95% probability within 1 SEE) to the Angoff MPS of the whole test. This finding is also consistent with previous research results (Ferdous & Plake, 2005, 2007; Kannan et al., 2015; Smith, 2011). Ferdous and Plake (2005, 2007) and Smith (2011), argued that approximately 50% of the items would be sufficient to obtain estimates similar to the MPS of the whole test. Similarly, Kannan et al. (2015) indicated that about 45 items were sufficient to obtain generalizable MPS for the whole test containing about 100 items.

In summary, this study suggests using 50% of the items with the C-SRS method to obtain estimates similar to the Angoff MPS of the whole test. When setting the passing score, educators may be advised to reduce the number of items considering this information. Thus, both time and money and workload can be saved. However, the fact that the number of the items used in the study was limited to 40 items and as a result, the number of the items in some cells formed in the strata was very low or not at all may have negatively affected the results. A future study may include more items. In addition, the fact that the difficulty parameters of the items in the test were very low and close to each other may have reduced the effect of the difficulty stratum. A future research may be carried out with tests with a wider range of item difficulties and more content areas, such as proficiency tests. Also, the fact that difficulty and content classifications were carried out in different ways may have had an effect on results, too. A future study may focus on using different sampling methods and different classification techniques (various number of difficulty / content strata) such as multiple matrix sampling and balanced incomplete block design. Moreover, the effect of stratification according to content, item difficulty and item discrimination may be examined in different educational practices, and the other standard setting methods may be used.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### ORCID

Hakan Kara  <https://orcid.org/0000-0002-2396-3462>

Sevda Çetin  <https://orcid.org/0000-0001-5483-595X>

### 5. REFERENCES

- Behuniak, P., Gable, R. K., & Archambault, F. X. (1982). The validity of categorized proficiency test scores. *Educational and Psychological Measurement*, 42, 247-252.
- Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, 9, 215–235.
- Buckendahl, C. W., Ferdous, A. A. & Gerrow, J. (2010). Recommending cut scores with a subset of items: An empirical illustration. *Practical Assessment*, 15(6), 1-10.
- Cizek, G. J. (2001). *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Çetin, S. (2011). *İşaretleme ve angoff standart belirleme yöntemlerinin karşılaştırılması [Comparison of Bookmark and Angoff Standard Setting Methods]*. PhD dissertation, Hacettepe University, Ankara.
- Downing, S. M. (2006). Selected-Response item formats in test development. In T. M. Haladyna & S. M. Downing (Ed.), *Handbook of test development* (pp. 287-300). Mahwah, New Jersey: Routledge.

- Ferdous, A. A., & Plake, B. S. (2005). The use of subsets of test questions in an Angoff standard setting method. *Educational and Psychological Measurement*, 65(2), 185-201.
- Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. *Educational and Psychological Measurement*, 67(2), 193-206.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8<sup>th</sup> ed.). New York: McGraw Hill.
- Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. N. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: Council of Chief State School Officers.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 433-470). Westport, CT: Praeger.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353-366.
- Irwin, P. (2007). *An alternative examinee-centered standard setting strategy* (Doctoral dissertation). University of Nebraska, USA.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485-514). New York: American Council on Education/Macmillan.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kannan, P., Sgammato, A., & Tannenbaum, R. J. (2015). Evaluating the operational feasibility of using subsets of items to recommend minimal competency cut scores. *Applied Measurement in Education*, 28(4), 292-307.
- Kannan, P., Sgammato, A., Tannenbaum, R. J., & Katz, I. R. (2015). Evaluating the consistency of angoff-based cut scores using subsets of items within a generalizability theory framework. *Applied Measurement in Education*, 28(3), 169-186.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Ed.), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- MEB (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı ve kılavuzu. [Elementary mathematics course curriculum and guide of 6-8. classes]*. Retrieved November 29, 2019, from <https://ttkb.meb.gov.tr>.
- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Norcini, J., Shea, J., & Ping, J. C. (1988). A note on the application of multiple matrix sampling to standard setting. *Journal of Educational Measurement*, 25(2), 159-164.
- Özçelik, D. A. (2013). *Test Hazırlama Kılavuzu [Test Preparation Guide]*. Pegem Akademi Yayıncılık.
- Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows* (2nd ed.). Crows Nest, Australia: Allen & Unwin.
- Plake, B. S., & Impara, J. C. (2001). *The fourteenth mental measurements yearbook*. Lincoln, NB: Buros Institute of Mental Measurements.
- Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task. The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.159-174). Mahwah, NJ: Erlbaum.

- Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A. M., & Rodriguez, G. (2000). *Setting standards on a computerized-adaptive placement examination*. Laboratory or Psychometric and Evaluative Research Report No. 378.
- Smith, T. N. (2011). *Using stratified item selection to reduce the number of items rated in standard setting*. University of South Florida, USA.
- Siegel, S. (1956). *Nonparametric methods for the behavioral sciences*. New York.

## What You might not be Assessing through a Multiple Choice Test Task

Burcu Kayarkaya <sup>1,\*</sup>, Aylin Unaldi <sup>2</sup>

<sup>1</sup>School of Foreign Languages, Yıldız Technical University, Davutpaşa Campus, 34220, İstanbul, Turkey

<sup>2</sup>School of Education and Professional Development, University of Huddersfield, Huddersfield, HD1 3DH, UK

### ARTICLE HISTORY

Received: 04 November 2019

Revised: 14 February 2020

Accepted: 06 March 2020

### KEYWORDS

Textual reading comprehension,  
Macrostructure formation,  
Reading operations,  
Multiple choice task,  
Summary task

**Abstract:** Comprehending a text involves constructing a coherent mental representation of it and deep comprehension of a text in its entirety is a critical skill in academic contexts. Interpretations on test takers' ability to comprehend texts are made on the basis of performance in test tasks but the extent to which test tasks are effective in directing test takers towards reading a text to understand the whole of it is questionable. In the current study, tests based on multiple choice items are investigated in terms of their potential to facilitate or preclude cognitive processes that lead to higher level reading processes necessary for text level macrostructure formation. Participants' performance in macrostructure formation after completing a multiple choice test and a summarization task were quantitatively and qualitatively analyzed. Task performances were compared and retrospective verbal protocol data were analyzed to categorize the reading processes the participants went through while dealing with both tasks. Analyses showed that participants' performance in macrostructure formation of the texts they read for multiple choice test completion and summarization task differed significantly and that they were less successful in comprehending the text in its entirety when they were asked to read to answer multiple choice questions that followed the text. The findings provided substantial evidence of the inefficacy of the multiple choice test technique in facilitating test takers' macrostructure formation and thus pointed at yet another threat to the validity of this test technique.

## 1. INTRODUCTION

One requirement a second language (L2) reader in an academic context has to meet is to process a text thoroughly and carefully to extract complete meanings from the written material. Careful reading of extended texts is the basic skill for “learning” in academic environments (Weir, Hawkey, Green, & Devi, 2009). In tests of English for academic purposes (EAP), careful reading at whole text level should thus be assessed to ensure adequate construct representation. It is a general contention that through designing several test items on main ideas in a text, text level comprehension can be achieved even in multiple choice (MC) tests. However, there is research that generally points out that scores obtained in MC tests measuring reading comprehension may not truly represent test takers' understanding of the written material (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008). It is also known that test

---

CONTACT: Burcu Kayarkaya ✉ [burcudurak@hotmail.com](mailto:burcudurak@hotmail.com) 📍 School of Foreign Languages, Yıldız Technical University, Davutpaşa Campus, 34220, İstanbul, Turkey

ISSN-e: 2148-7456 /© IJATE 2020



formats may determine how readers perform reading activities and test tasks in different formats can invoke different reading skills and strategies (Lee, 1986; Shohamy, 1984; Wolf, 1991). Studies conducted to discover whether task format has an effect on the extent and depth of reading comprehension have produced empirical evidence for the existence of inconsistency in reader performance due to task format (Kobayashi, 2002; Pearson, Garavaglia, Lycke, Roberts, Danridge, & Hamm, 1999). Thus, whether all task formats can facilitate the assessment of targeted skills and whether a test with certain tasks can operationalize the relevant reading skills with adequate coverage is an important issue for discussion, namely, a discussion on the construct validity of the test (ALTE, 2011).

Albeit, to our knowledge, there are no studies that focus on whether extensively used MC items can facilitate text level comprehension. To what extent this commonly used assessment technique can be instrumental in enabling readers to form a coherent mental representation of the text they read for test taking purposes is a question yet to be answered. This study aims at providing evidence for the claim that an MC reading test may be assessing comprehension of certain parts of a text but this may not necessarily mean that high scores from such a test reflect a complete understanding of the test text. This is yet another issue that challenges the validity of the use of MC format in reading assessment and it is important that this question be probed.

Reading in a foreign language is a complex process with many underlying cognitive components (Gernsbacher, 1997; Graesser, Singer & Trabasso, 1994; Kintsch, 1998; Myers & O'Brien, 1998). Examining these cognitive components is necessary to understand reading comprehension processes that take place in a reader's mind. Explanations on how a reader comprehends texts generally point to a series of processes to eventually arrive at constructing textual meaning. In the bottom-up processing, the reader starts with decoding linguistic structures or units to unfold propositions of the text one by one (Gough, 1972), or the top-down processing suggests more global processes of activating background knowledge to predict the content of the text and confirmation of the prediction takes place as the reading goes along (Goodman, 1967). In modern views of reading, reading process is seen as an interaction between bottom-up and top-down processes: Readers go along a continuum of selection of processes while reading, changing their focus from linguistic units towards textual clues or the other way round, making use of top-down and bottom-up processes in different quantities and sequences (Grabe, 1991).

Khalifa and Weir (2009) hypothesized that difficulty in reading is a function of the level of processing required by reading purpose and the complexity of the text. Reading is conceptualized as having several types; expeditious versus careful and local versus global reading. Moreover, careful reading is further divided into four levels including within-sentence (propositional meaning), across sentences (mental model; ongoing meaning making as the reader proceeds in the text), text (text model) and texts (documents model) models. Careful reading is predominantly a bottom-up process, starting with linguistic processing of the elements of a sentence and establishing propositional meaning (the literal interpretation of what is printed on the page). Through inferencing, the reader relates the message to the context. Inferencing is also functional in establishing coherence, or meaning between propositions, as the reader integrates new information into a mental representation of the text so far. This is the stage at which the reader starts to identify main ideas and impose a hierarchical structure on the information in the text. According to Kintsch and van Dijk (1978), this is the stage where microstructure rules are at work to link the textual pieces and reduce the content to higher propositions to be stored in working memory. Background knowledge on the content of the text and the meaning formed on the text so far facilitate inferencing and control of coherence and consistency in the text. At the text level, micropropositions are collapsed into macropropositions. Macropropositions are derived from a text through the application of macrorules: less important portions of the text are deleted, instances are generalized, and

summaries of events are constructed (van Dijk, 1980). Macroproposition of a text is the skeleton that makes up the text body which can also be regarded as an organised form of the most important portions of the text in an hierarchical order. Recognition of the hierarchical structure of the text is of crucial importance in forming a unified understanding at the text level. The new information presented in the text is combined with what the reader already has in supply and eventually a situation model is produced (Kintsch & Kintsch, 2005; Kintsch & van Dijk, 1978). Whether a reader will construct a situation model of interpretation depends on what purposes the reader is engaged in the text for. Urquhart and Weir (1998) categorize types of reading that serve for different purposes as follows: a) expeditious reading (quick, selective and efficient reading to access desired information in a text- scanning, skimming and search reading), and b) careful reading (processing a text thoroughly with the intention of extracting complete meaning from presented material). They further make distinctions between global and local comprehension gains from reading texts. Global comprehension refers to reaching an understanding of the explicit information available in a text, including main ideas and the links between these ideas, through integrating and synthesizing information. The reader is then able to build logical relationships between ideas. Local comprehension is more related to an understanding of propositions within the sentence and is a process that involves word recognition, lexical access and syntactic parsing and maintaining meaning at the phrase, clause and sentence level (Bax, 2013, Khalifa & Weir, 2009, Weir & Bax, 2012). A reader, for example, looking for specific information in a text may favor search reading or expeditious reading to access the necessary information quickly (Guthrie & Kirsch, 1987). However, a reader comparing the arguments of a writer with those of another writer may embrace a global, careful style that would enable him/her to arrive at a deeper understanding of the text.

Elaborating on the purposes of reading, Enright, Grabe, Koda, Mulcahy-Ernt, and Schedl (2000) put forward four purposes for reading in L2: a) reading to find information (search reading), b) reading for basic comprehension, c) reading to learn, and d) reading to integrate information across multiple texts. Grabe (2009) built upon Enright's four purposes and added two further purposes: e) reading for quick understanding (skimming), f) reading to evaluate, critique and use information. For some of the purposes listed above (search reading or skimming), a reader does not necessarily form a text or situation model as deriving the macrostructure of a text is not the aim, but for some models of comprehension (reading to learn or reading for basic comprehension), this is required. Enright et al. (2000) explain that in the reader purpose perspective, a reader approaches a text depending on what he/she is supposed to do with the text, assuming that all readers read for a reason in certain contexts, be it for an exam purpose or orientation in real life. A reader's standard of coherence affects the depth of their comprehension and alters how a text is processed. Requirements of the reading task itself or the goals a reader sets for reading a text change the reading processes a reader goes through while reading. That is, readers shape and reshape their reading behavior, or the processes, to fit the requirements of the model of the task in mind (Britt, Rouet & Durik, 2017).

For L2 learners of English, reading ability in an academic context means that they can successfully perform a variety of reading skills including the ability to read and understand a text in its entirety with the purpose of learning from it. As mentioned before, this is usually referred to as "text/situation model formation" (Kintsch, 1998), "reading to learn" (Enright et al., 2000), and "reading at the whole text level" (Khalifa & Weir, 2009). The basic principle in this process is the formation of the macrostructure and thus text/situation model. Whether this can be adequately assessed in tests of academic reading and whether certain item formats are conducive or non-conducive to the assessment of text level comprehension is a worthy matter of inquiry in the field of language testing.

Reading comprehension is a multi-faceted complex activity in which features of input text, the

reader's competences and motivation and task demands interact with each other to determine the characteristics of reading behavior (Bachman & Palmer, 2010; Khalifa & Weir, 2009). As included in Bachman's (1990) framework, "the nature of the expected response to the input (the test format)", and its relationship with the input (the interaction between the written material and the test format)" determine the extent and depth of reading.

In test development, test items are aimed at operationalizing certain sub-skills of reading, presumably in adequate coverage. There is certainly no best method for testing reading since no single test method can fulfill all the varied purposes for which one might be testing (Alderson, 2000). Convenience, practicality and efficiency may become primary considerations while deciding on the most suitable method for assessment (Bachman, 2000). More often than not, objectively and economically scorable item formats such as MC and matching items are chosen instead of open-ended, extended response items (Prapphal, 2008; Watson-Todd, 2008). However, this brings up the issue of whether different item formats can measure the same ability or not – such as whether the MC and open-ended items assess equivalent reading skills. Besides, there is the question of whether test items can invoke reading skills at macro levels as well as they do at micro levels; and if they do, whether they can cover all the skills relevant to the test purpose. Pearson, Garavaglia, Lycke, Roberts, Danridge and Hamm (1999) investigate whether there are differences in the cognitive processes readers execute to complete an MC and a constructed response task. The results indicate that the MC task activates a significantly lower proportion of skills at the macro level (e.g. intertextuality). However, as Kintsch and Kintsch (2005) underline, questions requiring the readers to set off macro operations and questions targeting the macrostructure of a text are more instrumental in reflecting reading comprehension ability of the readers.

Obviously, some task types are more conducive to assessing text level comprehension processes whereas others may only assess it at local levels. MC items are widely used in reading assessment; therefore, it is important to understand the characteristics of these items types, and find out whether tests formed of MC items can tap into text level reading comprehension processes (Sheehan & Ginther, 2001).

There are several reasons why the MC technique is widely used in tests. First, relatively more content can be covered in MC tests when compared to other test formats (Haladyna & Downing, 2009). That is, MC technique provides more flexibility in covering larger bits of information. Second, it makes scoring easy and effortless – human raters or machine raters can be in charge and the answer key is fixed in that in a carefully planned MC test, there is usually only one correct option (Fuhrman, 1996). MC test format also economically represents whether and to what extent the reader can read and understand parts of a text.

However, while responding to MC questions, readers usually have to choose the best option from alternatives rather than verbalizing or producing answers themselves. This means that the question format may limit the understanding of the reader to the ideas as they are worded in the options by the item writer. MC questions may mostly encourage memorization and factual recall and may not promote high-level cognitive processes (Airasian, 1994; Scouller, 1998). Another serious problem concerning the MC format is that the rater simply does not know why readers respond to the questions the way they do (Lau, S. Lau, Hong, & Usop, 2011). There is also the risk of guessing effect, especially if the distractors are not written carefully (Kurz, 1999).

Research also suggests that test takers use various strategies which are not necessarily comprehension-based to answer MC items and critics claim that such test-taking strategies are executed to get an acceptable answer to the question rather than to understand the text (Anderson, Hiebert, Scott, & Wilkinson, 1985; Cohen, 1984). Shohamy (1984) and Wolf (1991) investigate the format effects comparing MC questions and open-ended questions and

conclude that MC items are easier to deal with for readers. Rupp, Ferne and Choi (2006) provide empirical evidence for the hypothesis that when readers respond to texts followed by MC questions, they go through different processes than they would while reading in non-testing contexts. Rupp et al. (2006) state that readers tend to approach reading tasks with MC questions as a problem solving task, rather than a comprehension task. That is, readers, as strategic test takers, use a number of techniques to “solve” the problems that appear as questions in tests and this makes the activity less similar to real-life reading. Similarly, the study by Cerdan, Vidal-Abarca, Martinez, Gilabert, and Gil (2009) point at the likelihood of readers’ following a question-to-text sequence when answering MC questions.

Remembering that cognitive (construct) validity examines the relationship between what a test aims to measure and what it actually elicits from test takers (Weir, 2005), if the cognitive demands of a task do not adequately represent the demands of the skills in the target domain, the cognitive validity of such a task would be questionable. If specific task characteristics in a test affect the cognitive processes employed by test takers, then our inferences based on the scores from that test would be undermined (Smith, 2017). MC test format has been scrutinized from several aspects as discussed above (Martinez, 1999; Martinez & Katz, 1995; Rupp et al., 2006). However, although it is widely used in reading assessment, there is no study focusing on whether it facilitates comprehension at whole text level. With several MC items in a test, it is possible to cover all or most of the main ideas in a text. However, if test takers cannot form a successful macrostructure of the text they have dealt with during the test, then we can assume that processes leading to coherent mental representation formation are not facilitated or even precluded in reading tests with MC questions.

## **2. METHOD**

This study aims at providing evidence as to whether MC items can assess text level reading by comparing the performances of test takers in an MC test and an oral summary task by addressing two research questions:

RQ1: To what extent can textual level comprehension be attained upon the completion of multiple choice and oral summary reading tasks?

RQ2: How do test takers’ reading styles and preferences differ according to multiple choice and oral summary tasks?

### **2.1. Participants**

A total of 32 (15 female, 17 male) students were selected for the study through convenience sampling. In selection of the students, a number of factors were taken into consideration. As the materials used in the study target learners of English above a certain level of proficiency (at B2 level), participants were chosen from a pool of almost 300 students taking an Advanced English mass course at a state university based in İstanbul, Turkey. When forming the participant group, the grades students got from the midterm exams were checked and those with scores at or above 80 out of 100 were shortlisted and a further elimination was made according to the performance those had in the reading section of the exam. Eventually, the 32 students who were regarded as eligible for the study were asked for consent and all agreed to participate in the study.

### **2.2. Instruments**

#### **2.2.1. Tasks and texts used in the study**

Reading comprehension performance of the participants is measured by multiple choice (MC) and summarization tasks. Summarization is taken as a strong indicator of text level comprehension because summarizing a text requires the ability to identify main ideas in the text, integrate them into a text model, and develop a proper situation model of interpretation

(Grabe, 2009; Taylor, 2013). In order to understand main ideas, readers need to have a large receptive vocabulary, basic grammar, effective comprehension strategies, and strategic processing abilities to maintain a high level of comprehension and an awareness of discourse structure (Grabe, 2009; Pressley, 2002).

The summary technique provides a solid representation of how mental processes operate in the reader’s mind, how they prioritize, construct and organize information as well as the retrieval strategies they use (Bernhardt, 1983). Khalifa and Weir (2009) state that global (text level) careful reading at the highest level requires the reader to understand the micro and macropropositions in a text and how these are interconnected, while integrating new information into a mental model to create a discourse level structure that is appropriate to their purpose. We can say that the cognitive processes a reader has to follow to summarize a text are text level macrostructure formation processes.

The participants in this study were asked to summarize the text they read to answer MC questions right after completing the task. A combined term, multiple choice summary (MCSUM) is used to refer to this task. MCSUM aimed to measure the extent of textual level comprehension that surfaced upon the completion of MC questions, and the performance in MCSUM is compared to that in oral summary (SUMONLY) task. The SUMONLY task is used as a baseline to assess the general summarization abilities of the participants so that we can identify whether the success or failure in summarization in the MCSUM task is due to what is comprehended after the text is processed or to the general comprehension abilities of the participants. The participants completed the tasks (MC and SUMONLY) designed on two different texts (Text A and Text B).

The texts and the set of MC questions accompanying them were taken from TOEFL preparation materials. TOEFL is a strong representative of MC EAP tests students take in college and university settings. To ensure comparability of the texts in different tests (Text A and Text B) in terms of textual features (vocabulary, topic, language use and level, cohesion, coherence, syntactic simplicity, narrativity, genre and interest), automatic text analysis tools were used along with ideas and suggestions from expert judgement.

For the MC test, there were eight questions in both versions intended to elicit a variety of reading subskills based on TOEFL (2014) test specifications. These questions were carefully matched in terms of subskills and question types across Texts A and B by the researchers (see [Table 1](#)). The SUMONLY task was a verbal instruction for the participants, informing that what they had to do with the text was to read it to summarize. The study counter-balanced task and text order in a four-way distinction. [Table 2](#) describes how and in what order the tasks and texts were assigned to the participants.

**Table 1.** Question Types and the Reading Subskills the Question Types Measured.

Text A Q2	Text insertion/Cohesion formation	Text B Q5
Text A Q3	Sentence simplification	Text B Q8
Text A Q4	Factual information	Text B Q2
Text A Q5	Inference	Text B Q7
Text A Q6	Rhetorical purpose	Text B Q6
Text A Q7	Reference	Text B Q3
Text A Q8	Vocabulary	Text B Q1

**Table 2.** The Distribution of Tasks in Four Groups

Group	First session	Second session
Group I N=8	Text A – MC MCSUM VP (How did the participant read for the MC task?)	Text B – SUMONLY VP (How did the participant read for the SUMONLY task?)
Group II N=8	Text B – SUMONLY VP (How did the participant read for the SUMONLY task?)	Text A – MC MCSUM VP (How did the participant read for the MC task?)
Group III N=8	Text A – SUMONLY VP (How did the participant read for the SUMONLY task?)	Text B – MC MCSUM VP (How did the participant read for the MC task?)
Group IV N=8	Text B – MC MCSUM VP (How did the participant read for the MC task?)	Text A – SUMONLY VP (How did the participant read for the SUMONLY task?)

### 2.2.2. The scoring of summaries

The researchers worked with three instructors working at the School of Foreign Languages of the university mentioned above for expert judgement both in the formation of the tests and in the scoring of the summaries. The instructors read the texts to evaluate their appropriateness for the participant profile. They were also asked to identify the parts of the texts that were essential to include in an accurate summary of the texts. After having them make their own lists of relevant text parts, a meeting was organized with all the instructors to compare the summary lists including one of the researchers' and to discuss discrepancies. Thus, a consensus rubric for the scoring of the summaries was formed by the instructors and one of the researchers. These rubrics contained seven statements which carried primary information; i.e. main ideas and topic sentences for each text; Text A and Text B. While scoring, one point is given to each statement in the participants' summary that matched a statement in the rubric. The participants were free to summarize the text either in their L1, Turkish, or in L2, English. Each summary was scored twice by one of the researchers using the rubrics: during the sessions when the participants completed the tasks and upon completion of data collection. The time interval between the two scorings was 15 days. The final scores were turned into percentages. The scores obtained from the first and second scoring of the summaries for both conditions, MCSUM and SUMONLY, were compared and the intra-rater agreement, Cohen's kappa, was found to be 0.76 and 0.73, respectively. The difference between the scores was analyzed through one-way ANOVA and the effects of test methods and texts were analyzed through two-way ANOVA between subjects.

### 2.2.3. Retrospective verbal protocols

In order to investigate the cognitive processes the participants went through while reading to answer MC questions and summarizing the text, retrospective verbal protocol (VP) technique was used. Following the summarization process both after MC task (MCSUM) and in SUMONLY condition, the participants were asked to reflect on their reading behavior, explaining how they read the texts to answer MC tasks and whether or not reading for such a

purpose affected their reading style. The participants were also asked to describe their reading processes when they read to summarize the text. All VP sessions were video-recorded. When necessary, prompt questions were asked to help the participants express how they handled the task. The questions were in participants' L1, that is to say, in Turkish.

#### **2.2.4. Coding verbal protocols**

In order to develop a coding scheme to classify the reading processes emerging from VPs, several coding schemes were examined (Cohen & Upton, 2007; Lim, 2014; Unaldi, 2004; Weir et al., 2009). As a result, a list of reading operations that fit the purpose of the current study was compiled. Two more operations that emerged from the VP data were added to the reading operations. A customized scheme of 11 reading operations (RO) that the participants stated they executed while accomplishing the tasks was formed.

The VPs for each participant (following MC and SUMONLY tasks) were coded and scored by one of the researchers using the coding scheme, identifying each reading operation as test takers verbalized them. Following this, a second coder coded both tasks using the coding scheme and watching the video-recorded sessions. The two coders' results were compared and inter-rater agreement on the reading operations in MC task was found to be 81% and for the SUMONLY task, the agreement on the reading processes were calculated as 73%. A paired-sample t-test was run to determine the statistical significance of the difference between the reading operations used in MC and SUMONLY tasks.

### **3. RESULT / FINDINGS**

#### **3.1. Research Question 1**

RQ1: To what extent can textual level comprehension be attained upon the completion of multiple choice and oral summary tasks?

To provide an answer for this question, the researcher used the summary rubric to count how many of the sentences in the rubric were produced by the participants while they summarized in two conditions. By doing so, the extent to which the macrostructure of the text had been successfully formed by the participant in both conditions was assessed. The scores for the two summaries were then compared to find out whether they pointed to a statistical difference in the formation of macrostructures.

A paired samples t-test was used to test the null hypothesis that there is no difference between the mean scores from two summaries (MCSUM and SUMONLY). In addition, a two-way analysis of variance (ANOVA) was used to investigate whether the texts or the tasks accounted for the main variance. Table 3 shows the distribution of the scores in the tasks. The mean scores (converted into percentages) the participants obtained showed that they had the highest scores in the SUMONLY task, when they read the text to make a summary of it. The lowest scores were observed in MCSUM task, when the participants summarized the text they had read to answer the MC questions (MC task). In other words, the participants received lower scores when they summarized the text upon the completion of the MC task in comparison to the case when they read a text to summarize it.

The success levels of the participants in textual level comprehension in MCSUM and SUMONLY were found to be statistically significantly different ( $p=.002$ ) through one-way analysis of variance given in Table 4. The participants performed better in macrostructure formation, that is to say, text level comprehension when they read for the SUMONLY task ( $M=65.59$ ,  $SD=21.14$ ). In the MCSUM, the macrostructures they produced of the texts were significantly less successful ( $M=48.6\%$ ,  $SD=22.46$ ).

**Table 3.** Descriptive Statistics

	MC	MCSUM	SUMONLY
M	63.35	48.6	65.69
Median	62.5	42.8	71.4
SD	19.71	22.64	21.13
Skewness	-.2	.36	.06
Kurtosis	-.7	-.64	-.93

**Table 4.** One-way Analysis of Variance

(I) Method	(J) Method	Mean dif. (I-J)	Std. E.	Sig.	95% CI	
					Lower bound	Upper bound
MC	MCSUM	14.75	5.22	.006	4.73	25.122
	SUMONLY	-2.23	5.22	.670	-12.6	8.14
MCSUM	MC	-14.75	5.22	.006	-25.12	-4.37
	SUMONLY	-16.98	5.22	.002	-27.35	-6.6
SUMONLY	MC	2.23	5.22	.670	-8.14	12.6
	MCSUM	16.98	5.22	.002	6.6	27.35

The results of the analyses (Tables 3 and 4) indicate that reading a text for the purpose of answering MC questions may not contribute to the formation of the macrostructure of the text as much as reading for summarization purposes does. It is important to note that overall performance difference among the MC and SUMONLY task scores was minimal. Although the scores participants got from the MC test were close to the grades they got from the SUMONLY task (see Table 3), the performance in MC task was not totally transferred to the task following immediately in which formation of the macrostructure of the text was required.

Besides, two-way ANOVA between-subjects analysis (see Table 5) indicates that the difference between the mean scores of the two tasks was merely a result of the test method and that the test results were not affected by the texts used in the study or by text and method interactions; the effect for method was significant ( $F=6.24, p=.003$ ).

**Table 5.** Two-way ANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7987.41	5	1597.48	3.66	.005
Intercept	336291.53	1	336291.53	771.15	.000
Method	5449.64	2	2724.82	6.24	.003
Text	944.38	1	944.38	2.16	.145
MethodText	1593.39	2	796.69	1.82	.167
Error	39247.85	90	436.08		
Total	383526.81	96			
Corrected Total	47235.27	95			



### 3.2. Research Question 2

RQ2: How do test takers' reading styles and preferences differ according to multiple choice and oral summary tasks?

The purpose of RQ2 was to identify the types and frequency of reading operations employed by the participants in the study while they completed an MC and a summarization task. By doing so, we aimed at examining the reading operations that were instrumental in the participants' performances during reading for the two tasks and whether the cognitive processes they engaged in while reading contributed to the formation of the macrostructure of the texts. For the MC task completion, eight reading operations (RO) were identified during the verbal protocols. Table 6 lists the reading operations that participants stated they operationalized during reading for the MC task.

**Table 6.** ROs the Participants Stated they Went Through during the MC Task

		f	%
RO3	I followed a question-to-text sequence, matching the keywords in text and questions	30	93.7%
RO6	I read the whole text from the beginning to the end carefully	12	37.5%
RO5	I read carefully only the selected part(s) of the text that might be relevant to the question	9	28.1%
RO7	During reading, I read a part of the text more than once to understand it	8	25%
RO1	I read the text carefully first, before attempting the task	7	21.8%
RO2	I read the text expeditiously to have a general idea before attempting the task	6	18.7%
RO4	I read expeditiously to find a relevant part that might include the answer	6	18.7%
RO9	I tried to understand how the text was organized, how the ideas and details were connected	1	3.1%

The compilation of reading operations that the participants stated they went through while completing the MC task emphasizes certain characteristics of MC tasks and make it clear how test takers approach these tasks. For the MC task, 93.7% of the participants reported that they “followed a question-to-text sequence, matching the keywords in the text and the questions, (RO3)”. Only 21.8% of the participants stated they “read the text carefully before attempting the task (RO1)”.

For the SUMONLY task, the reading operations the participants in the study stated they carried out are presented in Table 7. Eight reading operations arose from the participants' statements regarding their reading preferences for the summary task.

**Table 7.** ROs the Participants Stated they Executed during the SUMONLY Task

		f	%
RO1	I read the text carefully, before attempting the task	31	96.8%
RO6	I read the whole text from the beginning to the end carefully	31	96.8%
RO10	I read to get the main ideas and remember them	12	37.5%
RO7	During reading, I read a part of the text more than once to understand it	11	34%
RO8	I read to make connections between paragraphs or parts	10	31.2%
RO11	I paid further attention to introduction and conclusion paragraphs as they would include the main idea	7	21.8%
RO9	I tried to understand how the text was organized, how the ideas and details were connected	5	15.6%
RO2	I read the text expeditiously to have a general idea before attempting the task	1	3.1%

While reading for the SUMONLY task, 96.8% of the participants stated they “read the text carefully first, before attempting the task, (RO1)” and as a result, they “read the whole text carefully, (RO6)”. The second most frequently utilized reading operation was RO10, “I read to get the main ideas and remember them”, which was reported by 37.5% of the participants. 21.8% of them reported that they “paid further attention to the introduction and conclusion paragraphs as they would include the main idea of the whole text, (RO11)” when they were dealing with SUMONLY task. 31.2% of the participants asserted that they “read to make connections between paragraphs or parts, (RO8)” and 15.6% of them stated that they “tried to understand how the text was organized and how the ideas and details were connected, (RO9)” to understand the text. It is clear that as expected, these reading processes are genuinely careful reading operations that a test taker can make use of to attain comprehension at text level.

To compare the means of reading operations of MC and SUMONLY tasks, a paired-samples t-test was performed (see Table 8). It indicated that differences were statistically significant between the means of every reading operation executed for MC and SUMONLY tasks except for RO6. However, the effect size ( $d=0.27$ ) for this analysis was found to be small.

**Table 8.** Paired Samples T-test – The Comparison of ROs

		M	SD	Std. E.M.	95% CI of the D.		df	Sig.(2-tailed)
					Lower	Upper		
Pair 1	RO1MC-RO1SUM	.75	.44	-.9	.07	-.59	31	.000
Pair 2	RO2MC – RO2SUM	.15	.36	.06	.02	.28	31	.023
Pair 3	RO3MC – RO3SUM	.18	.39	.07	.04	.33	31	.012
Pair 4	RO4MC – RO4SUM	.28	.45	.08	.11	.44	31	.002
Pair 5	RO5MC – RO5SUM	-.59	.49	.08	-.77	-.41	31	.000
Pair 6	RO6MC – RO6SUM	-.09	.64	.11	-.32	.13	31	.414
Pair 7	RO7MC – RO7SUM	-.28	.45	.08	-.44	-.11	31	.002
Pair 8	RO8MC – RO8SUM	-.15	.36	.06	-.28	-.02	31	.023
Pair 9	RO9MC – RO9SUM	-.37	.49	.08	-.55	-.19	31	.000
Pair 10	RO10MC – RO10SUM	-.25	.44	.07	-.4	-.09	31	.003
Pair 11	RO11MC – RO11SUM	.93	.24	.04	.84	1.02	31	.000

#### 4. DISCUSSION and CONCLUSION

This study was set out to investigate whether MC tests of reading comprehension allow test takers to form an integral understanding of the texts they read during test taking process. The aim was to test whether through MC questions it might be possible to make the test takers to process all the information in the text so that they can form a coherent summary in their minds. Two equal forms TOEFL reading tests were used for two different summarization tasks in a counter-balanced way; summarization after taking an MC test and summarization without questions (SUMONLY). The comparison of the results have shown that test takers’ summarization was less effective after an MC test; they could remember fewer ideas from the text than they would normally remember if they read the text from the beginning to the end. The analyses conducted showed that the level of comprehension required for answering MC questions was not adequate for a satisfactory summary formation as the test takers could remember only half of the main ideas from the texts (48.6%). Verbal protocols confirmed that reading for MC test is strategic: An overwhelming percentage of participants (93.7%) stated that they “followed a question-to-text sequence” as a problem-solving activity where the problem is the question and the text is only a means to answer it. Therefore, questions are prioritized and reading is guided by the questions and maintained as much as the questions required. Following a question-to-text sequence in test taking means that reading the whole text

is not a requirement, which is supported by only a few participants (37.5%) “reading the whole text from the beginning to the end carefully (RO6)” in this study.

The reading skills that were uniquely operationalized in the MC task, but were not mentioned in the SUMONLY task, are also worth consideration as they help differentiate between the two tasks. RO3 “I followed a question-to-text sequence”, RO4 “I read expeditiously to find a relevant part that might include the answer” and RO5 “I read carefully only the selected part(s) of the text that might be relevant to the question” were strategic, expeditious reading operations that predictably did not emerge from how the participants in the study described their reading processes for the SUMONLY task. It is, therefore, fair to say that reading a text for the MC task and the SUMONLY task required the execution of different reading operations and that the MC task directed the participants towards the activation of local level selective reading operations, which served only for the answering of the questions but not necessarily understanding the text as a whole. Therefore, we can conclude that text level understanding might be hindered through task-specific strategic question answering in an MC task whereas in linear, careful reading to summarize a text, as there are no interfering processes, text level understanding is more possible. Thus, we can conclude that task specific MC reading operations do not contribute to a deeper understanding of a text; on the contrary, they may even be hindering it.

That is a critical observation to make in terms of reading activities. When the whole text is not read in a test, as proven by the remaining 62.5% of participants who did not mention reading the text in full, interpretations about a test taker’s ability based on a representative sample of reading operations cannot be made. Reading for answering questions, which is inherent in MC design, seems to contribute only to finding the answers to the questions as test takers may devote less effort for reading outside the scope of the questions. If in a reading test, test takers can complete tasks without reading the whole text, the test is more like a puzzle activity where test takers find out which information in a text fits the question.

Attaining comprehension at textual level and formation of macrostructure requires a careful, high-level reading process where readers are able to use their ability to integrate information and draw conclusions (Pressley, 2002). When a careful and a linear reading style is not adopted, and most importantly, when reading activity takes place in a segmented and selective manner, formation of macrostructure is not likely to take place because the processes necessary for it cannot be substantiated during such a reading activity as was confirmed by our MCSUM test results. A local and segmented style of reading, whether it is done in a careful or expeditious manner, leaves test takers with only pieces of information from the text. In this study, none but only one participant indicated that they read to understand organization of the text in MC task.

On the other hand, the responses to the summarization task reflected a careful, linear reading style from the beginning to the end of the text. This is quite similar to text processing when the reader needs to understand all the information in the text and learn from it. As Khalifa and Weir (2009) stated, during summarization, high level processes come into play, directing the reading activity so that it follows a global and holistic manner. Such reading operations were reflected in 96.8% of the participants who stated that they “have read the whole text from the beginning to the end carefully” (RO6). Our findings are also supportive of Taylor (2013), who states that summarization tasks represent “a view of text comprehension as the construction of a mental representation of the whole text and they therefore offer an appropriate format for assessing this” (p.56).

Considering the frequent use of MC tests in the assessment of academic reading ability, the question is whether or not an MC assessment tool is an efficient evaluation technique that can account for the relevant reading skills and sub-skills indicative of academic reading ability. Validity of our interpretations based on test results is closely contingent upon whether the test measures what it intends to measure (Brown, 1996). The task used for assessment purposes,

thus, should have a good representation of what a test taker can do in a real-life context. Reading in a real-life academic context requires readers to read at deeper levels and construct both a text model of comprehension and a situation model of interpretation (Grabe, 2009; Kintsch, 1998). More specifically, an important skill in an academic context is to be able to “read and understand a text in its entirety with the purpose of learning from it” and this skill is associated with high-level reading processes where “reading at the whole-text level” (Khalifa & Weir, 2009) and eventually the formation of the macrostructure of it is required. The formation of the macrostructure of a text necessitates the mastery of global comprehension skills and they are associated with understanding explicit information in the text and extracting main ideas and making connections between them to eventually integrate and synthesize information (Bax, 2013; Weir & Bax, 2012). Thus, tests that fall short of assessing ability to understand a long text and to form efficient macrostructure of it cannot help us to arrive at adequate and accurate interpretation of test takers’ readiness for academic life.

MC tasks are practical in terms of their administration and scoring and they are definitely useful in assessing several other reading skills. Note that reading is an umbrella term that covers many sub-skills a learner needs to develop to be able to cope with a written text that they encounter in real life. MC tasks, for instance, can be regarded as effective tools to teach and assess the reading sub-skills that require the mastery of sentence and paragraph level reading comprehension and the ability to search for information. In order to teach or to assess textual level comprehension skill, however, we have seen that asking several MC questions on a text is not a helpful format. Global level deeper comprehension should be emphasized both in teaching reading and in the assessment of it and appropriate techniques should be used for this. Classroom instruction and teaching programs preparing learners for academic life should incorporate practices that cultivate and enhance the required academic reading skills. The focus of classroom practices and materials should be to teach learners how to read texts to perform well in different reading types and also process a text fully to extract complete meanings from it (Weir et al., 2009) and create a text and a situation model (Kintsch, 1998) by reading carefully at the whole text level (Khalifa & Weir, 2009).

As mentioned above, accurate interpretations about a test taker’s reading ability cannot be made depending on a test that is not representative of relevant reading skills. Similarly, it is not realistic to expect desired outcomes from teaching programs that include practices that are weak in such coverage. As proven to be instrumental in the assessment of textual level comprehension in the current study, summaries, both oral or written, should be utilized more for teaching purposes. Summaries are good indicators of the formation of macrostructures and the macrostructure of the written material reflects reading comprehension more effectively (Kintsch & E. Kintsch, 2005).

In terms of assessment procedures, we need to make sure that decisions concerning the test format are not made at the expense of construct validity. Tests designed for academic assessment purposes, as well as research purposes, should be assessing test takers’ reading ability at several levels as required in real life settings where readers need to tackle with long texts which they have to read and understand from the beginning to the end. Test tasks should target such higher level integrative, interpretative reading skills as well as expeditious, selective ones to draw a full picture of a reader’s ability. Otherwise, they are risking their validity.

### **Declaration of Conflicting Interests and Ethics**

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## ORCID

Burcu Kayarkaya  <https://orcid.org/0000-0002-3801-6345>

Aylin Ünalı  <https://orcid.org/0000-0003-4119-6700>

## 5. REFERENCES

- Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK; New York, NY, USA.
- ALTE (2011). Manual for Language Test Development and Examining, Strasbourg, Council of Europe. Retrieved from <https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b>
- Airasian, P. W. (1994) Classroom assessment (2nd ed.). New York: McGraw-Hill.
- Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. (1985). *Becoming a nation of readers: the report of the Commission on Reading*. Pittsburgh, PA: National Academy of Education.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Bachman, L. F. (2000). *Language Testing in Practice*. Oxford: OUP.
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye tracking. *Language Testing*, 30(4), 441-465. [doi:10.1177/0265532212473244](https://doi.org/10.1177/0265532212473244)
- Bernhardt, E. B. (1983). Three approaches to reading comprehension in intermediate German. *The Modern Language Journal*, 67(2), 111-115. [doi:4781.1983.tb01478.x](https://doi.org/10.1177/0265532212473244)
- Britt, M., Rouet, J.F., Durik, A. M. (2017). *Literacy beyond Text Comprehension*. New York: Routledge, <https://doi.org/10.4324/9781315682860>
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Cerdan, R., Vidal-Abarca, E., Martinez, T., Gilabert, R., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Language and Instruction*, 19(1), 13-27.
- Cohen, A. (1984). On taking language tests. *Language Testing*. 1(1). 70-81. [doi:10.1177/026553228400100106](https://doi.org/10.1177/026553228400100106)
- Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing*, 24(2), 209-250. [doi:10.1177/0265532207076364](https://doi.org/10.1177/0265532207076364)
- Cutting, L.E., & Scarborough, H.S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, 10(3), 277-299. [doi:10.1207/s1532799xssr1003\\_5](https://doi.org/10.1207/s1532799xssr1003_5)
- Enright, M.K, Grabe, W., Koda, K., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: a working paper*. Princeton, NJ: ETS.
- Fuhrman, M. (1996). Developing Good Multiple-Choice Tests and Test Questions. *Journal of Geoscience Education*, 44(4), 379-384. [doi:10.5408/1089-9995-44.4.379](https://doi.org/10.5408/1089-9995-44.4.379)
- Gernsbacher M. A. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships in the production and comprehension of text* (pp. 3-21). Mahwah, NJ: Erlbaum.
- Graesser A., Singer M., & Trabasso T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly*, 25(3), 375-406. [doi:10.2307/3586977](https://doi.org/10.2307/3586977)

- Grabe, W. (2009). *Reading in a second language: moving from theory to practice*. Cambridge: Cambridge University Press.
- Goodman, K.S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6(4), 126-135. [doi:10.1080/19388076709556976](https://doi.org/10.1080/19388076709556976)
- Gough, P.B. (1972). One second of reading. *Visible Language*, 6, 290-320.
- Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, 79(3), 220–227. <https://doi.org/10.1037/0022-0663.79.3.220>
- Haladyna, T.M., & Downing, S.M. (2009). A Taxonomy of Multiple-Choice Item- Writing Rules. *Applied Measurement in Education*, 2(1), 37-50. [doi:10.1207/s15324818ame0201\\_3](https://doi.org/10.1207/s15324818ame0201_3)
- Khalifa, H., & Weir, C. (2009). Examining Reading: Research and Practice in Assessing Second Language Reading. *Studies in Language Testing*, 29. Cambridge: Cambridge University Press.
- Keenan, J.M., Betjemann, R.S., & Olson, R.K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281-300.
- Kintsch, W. (1998). *Comprehension: a paradigm for cognition*. Cambridge: Cambridge University Press.
- Kintsch, W., & Kintsch, E. (2005). Comprehension. In Paris, S. G. and Stahl, S. A. (eds.) *Children's Reading Comprehension and Assessment*, 71-92. Mahwah, New Jersey: Erlbaum.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394. <https://doi.org/10.1037/0033-295X.85.5.363>
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193-220.
- Kurz, T.B. (1999). *A review of scoring algorithms for multiple choice tests*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.
- Lau, P.N.K., Lau, S.H, Hong, K.S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple choice tests. *Educational Technology & Society*, 14, 99-110.
- Lee, J.F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition*, 8, 201-212.
- Lim, H. J. (2014). Exploring the validity evidence of the TOEFL IBT reading test from a cognitive perspective. *Unpublished PhD Thesis*. Michigan State University.
- Martinez, M. E., & Katz, I. R. (1995). Cognitive Processing requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment. *Educational Assessment*, 3(1), 83–98. [doi:10.1207/s15326977ea0301\\_4](https://doi.org/10.1207/s15326977ea0301_4)
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207–218. [doi:10.1207/s15326985ep3404\\_2](https://doi.org/10.1207/s15326985ep3404_2)
- Myers J. L., & O'Brien E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26, 131-157.
- Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing*, 25(1) 127-143.
- Pearson, P. D., Garavaglia, D., Lycke, K., Roberts, E., Danridge, J., & Hamm, D. (1999). The impact of item format on the depth of students' cognitive engagement. Washington, DC: *Technical Report*, American Institute for Research.

- Pressley, G.M. (2002). Metacognition and self-regulated comprehension. In Farstrup, A.E., & Samuels, S.J. (Eds.), *What research has to say about reading instruction*. Newark, DE: International Reading Association.
- Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, 23(4), 441-474.
- Scouller, K. (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. *Higher Education*, 35, 453-472.
- Sheehan, K.M. and Ginther, A. (2001). What do passage-based MC verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1(2), 147-170.
- Smith, M. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, 54(6), 1256-1287.
- Taylor, L. (2013). Testing reading through summary: investigating summary completion tasks for assessing reading comprehension ability. *Studies in Language Testing*, 39. Cambridge, England, UCLES/Cambridge University Press.
- Unaldi, A. (2004). *Componentiality of the reading construct: Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test*. Unpublished PhD Thesis. Faculty of Education, Boğaziçi University.
- Urquhart, S., & Weir, C. (1998). *Reading in a second language: process, product and practice*. London: Routledge.
- Watson Todd, R. (2008). The impact of evaluation on Thai ELT. In Ertuna, K., French, A., Faulk, C., Donnelly, D., and Kritprayoch, W. (Eds.), *Proceedings of the 12th English in South East Asia conference: Trends and Directions*, pp.118-127. Bangkok: KMUTT
- van Dijk, T.A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction and cognition*. Hillsdale (N.J.): Lawrence Erlbaum Associates.
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281-300. <https://doi.org/10.1191/0265532205lt309oa>
- Weir, C, Hawkey, R, Green, T., & Devi, S. (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. IELTS Research Reports, 9, 157–189, British Council, London and IELTS Australia, Canberra.
- Weir, C., Bax, S. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. Cambridge Research Notes, Cambridge ESOL, 47 (February 2012), 3–14. Retrieved from [www.cambridgeesol.org/rs\\_notes/rs\\_nts47.pdf](http://www.cambridgeesol.org/rs_notes/rs_nts47.pdf)
- Wolf, D.F. (1991). *The effects of task, language of assessment, and target language experience on foreign language learners' performance on reading comprehension tests*. Dissertation, University of Illinois, ProQuest Dissertations and Theses.

## Scaling of Mood-State and Sample Cases Causing Anger in a Relationship with Rank-Order Judgment and Classifying Judgment

Merve Yildirim Seheryeli <sup>1,\*</sup>, Duygu Anil <sup>2</sup>

<sup>1</sup>Hasan Kalyoncu University, Faculty of Education, Gaziantep, Turkey

<sup>2</sup>Hacettepe University, Faculty of Education, Ankara, Turkey

### ARTICLE HISTORY

Received: 10 November 2019

Revised: 11 February 2020

Accepted: 06 March 2020

### KEYWORDS

Anger in a relationship,  
Scaling techniques,  
Rank-order judgment,  
Classifying judgment,  
Measurement of affective  
features

**Abstract:** This study is a survey study which aims to determine underlying causes of anger and the anger levels of individuals, in the sample cases and mood-states defined in the research. 255 people participated by filling in forms developed by the researcher. They were asked to rank 6 mood-state expressions between 1 and 6, to classify 23 sample case expressions between 1 and 4. Using Microsoft Office Excel 2016, responses given to mood-state expressions were examined with rank-order and given to sample case expressions were examined with classifying judgment with respect to gender and marital status. The findings of rank-order judgment scaling revealed that all participants get angry most when they are treated unfairly and they get angry least when they are criticized. It was also found that females got angry more at being neglected, and males got angry more at arrogance and mistrust. It was concluded that married people got angry more at being neglected; unmarrieds got angry more at mistrust. The findings of classifying judgment scaling showed that all participants get angry the most when unnecessary and offending comments are made about their families. They get angry the least when their partners are fan of any subject. It also has been seen that married participants chose ‘Ignoring the subjects that I care about’ the most and those who are unmarried chose ‘Making unnecessary and offending comments about my family’ the most.

## 1. INTRODUCTION

When the other drivers do not obey the traffic rules while we obey the rules; when drivers go on picking up passengers on a fully loaded bus in public transportation; when our children do not listen to us or our boss mobbing us; when our parents do not allow us to do something or our partner does not pack his/her socks; when our teacher gives us a low mark or the person that we love does not love us, when we can't express ourselves sufficiently or when we experience the worst things all the time or while watching the evening news, we respond with one of our basic feelings: anger.

Anger is “a natural reaction to unsatisfied wishes, undesirable results and unmet expectations”. Anger, which works as a self-protection mechanism as long as the degree of this reaction is favorable, becomes dangerous when it turns into a deep hatred and aggression. Domestic violence, abuse, harassment, terrorism and murder etc. can be shown as examples of

CONTACT: Merve YILDIRIM SEHERYELI ✉ [yldrm.mrv.7806@gmail.com](mailto:yldrm.mrv.7806@gmail.com) 📧 Department of Educational Sciences, Faculty of Education, Hasan Kalyoncu University, Gaziantep, Turkey

ISSN-e: 2148-7456 /© IJATE 2020



situations that anger turns into danger. Therefore, to accept without denial, express in a controlled manner without suppressing it, understanding the reasons and restricting them can help the feeling of anger become favorable and effective before it turns into destructive behaviors (Soykan, 2013).

Many studies indicate that anger is not planned and it generally occurs as a result of basic painful feelings such as offence, resentment, rejection, fear, anxiety, frustration, being treated unfairly, criticism and humiliation (Balkaya, 2001; Balkaya & Şahin, 2003; Satici, 2014). It becomes easier to deal with this feeling when the reasons of anger are understood and situations that cause anger are noticed. Soykan (2013) explained why we should deal with anger as follows:

- Anger causes a lot of social and individual problems such as verbal, physical violence, abuse etc. and it gives rise to serious problems in interpersonal relations in work and family life.
- As a result of not being able to overcome anger, it leads to mental problems like avoiding social life, addiction to smoking/drugs, eating disorders and depression.
- Anger that is not expressed in appropriate ways triggers physical problems as cardiovascular disease, immune and excretory system discomfort.

In her thesis study, Balkaya (2001) developed Multidimensional Anger Scale and included the dimension of anger eliciting situations in addition to the dimensions which are symptoms of anger, anger reactions, anger related cognitions and interpersonal anger. However, she tried to reveal the differences/resemblances between anger and furiousness.

In their study, Balkaya & Şahin (2003) conducted a scale development study that discusses anger as a multidimensional issue. They included the dimension of anger eliciting situations together with the dimensions which are symptoms of anger, anger reactions, anger related cognitions, and interpersonal anger. Yet, these situations remained restricted to not being taken seriously, being treated unfairly and being criticized.

When the literature was examined, it was found that Erdoğan (2018), Kırdök (2017), Uğurcan (2018), and Yapabaş (2018) investigated the relations between anger level, anger management style and different variables (stress, depression, alexithymia, eating behaviors, codependency, early adaptation schemas) in their thesis studies.

Although anger mostly seems to be a feeling towards people who we do not like, in their study Kassinove & Suckodolsky (1995) has found that people get angry with the people they like most or the people they know, then they get angry with the people they do not know, and they get angry the least with the people that they dislike.

The reason why people get angry with the people they like most or the people they know is due to individual differences such as culture, education and perspective, between the people they communicate with most. Kaynak (2014) states that because couples are in constant interaction, their conflict areas increase, so anger is frequently experienced. However, he stated that the reactions of males and females to anger are also different. In most of the studies mentioned above, it has been pointed out that anger differentiates according to gender.

In interpersonal and romantic relationships or in marriage, while anger should be perceived as an individual difference, it is regarded as a war that must be won. Hence, many couples show behaviors like ignoring conflicts, denial, avoiding facing with each other, as they do not know how to cope with anger (Özmen Süataç, 2010). As a result of this, marital satisfaction and harmony and interpersonal communication get weaker, so it causes couples to give up on each other (Erok, 2013; Togay, 2016).

Psychological features, like feelings that cannot be observed directly and cannot be represented with physical magnitude, are tried to be defined as the way people perceive them. Psychophysics is the science that reveals the relationship between the measured (physical dimension) and perceived (psychological dimension) magnitude of these stimulants.

In this psychological dimension, there are no defined units or scales of the variables. Therefore, attempts to estimate the relations with least error led to scaling methods (Turgut & Baykul, 1992). Two different methods are used in scaling. In judgment method observers or experts scale stimulants in one dimension by identifying the location of each stimulant according to the other stimulant, whereas in response method the people that give response, not as experts but as subjects who give their own judgments, determine the location of stimulant according to its own location in scaling dimension (Anıl & Güler, 2006).

There are different types of scaling methods such as pair wise comparison (Güler & Anıl, 2009; Güvendir & Özer-Özkan, 2013; Yılmaz Koğar & Demircioğlu, 2016), ranking (Özkan & Arslantaş, 2013; Bozgeyikli, Toprak & Derin, 2016; Yaşar, 2016), classifying (Demirus & Gelbal, 2020; Güvendir & Özer-Özkan, 2013; Sayın & Gelbal, 2014), absolute judgment method (Tezbaşaran, 2017), summated rating scale and multidimensional scaling (Bülbül & Köse, 2010; Tüzüntürk, 2009). Although the studies conducted show that only one scaling method has been used, Acar Güvendir & Özer Özkan (2013), Albayrak Sarı & Gelbal (2015) have used pair wise comparison and rank-order judgment scaling methods together in their studies. In this study, both rank-order judgment and classifying judgment scaling will be used.

In order to carry out measurement process, a well-defined structure and operational definition of this structure are needed (Crocker & Algina, 2006). However, a need for bridge between psychological and physical space arise when it comes to the measurement of complex structures as feelings.

In this study, it was aimed to scale situations eliciting anger in individuals who are in a relationship. In accordance with this aim, rank-order judgment and classifying judgment scaling methods were used in order to investigate how mood-states and sample cases causing anger were ranked and to study the difference in this ranking. Research questions of the study are as follows:

1. How are the scale values and ranking of mood-states causing anger in males and females?
2. How are the scale values and ranking of mood-states causing anger in married participants and unmarrieds?
3. How are the scale values and ranking of sample cases causing anger in males and females?
4. How are the scale values and ranking of sample cases causing anger in married participants and unmarrieds?

By means of responses given to these research questions, it was intended to determine the reasons that individuals in a relationship (married or unmarried) feel angry most or least with a method different from the literature. It is considered that understanding the reasons behind anger will decrease the conflicts, and couples' desire to understand each other and maintain a relationship will contribute to relationship satisfaction.

## 2. METHOD

In this section, data analysis is explained by giving information about research model, study group, and data collection tools.

### 2.1. Research Model

In this study, it was aimed to scale responses to the situations eliciting anger in a relationship, given by the individuals who are married, engaged or have a romantic relationship, with rank-order and classifying judgment methods. This study was conducted to describe current situation

without changing or influencing facts or without generalizing it to the population. Hence it is a single survey study. In this research model, variables such as the case in question, subject, individual etc. are described separately (Karasar, 2016).

## 2.2. Study Group

Study group was determined using one of the non-random sampling methods, convenience sampling method which aims to save time, money and labor force, on a volunteer basis (Büyüköztürk et al., 2017). Analyses were carried out based on the responses of 255 people, as 4 of 259 stated that they had no relationship before. Information related to the study group is given in Table 1.

**Table 1.** Demographics of Study Group

Gender	Relationship Status									
	Married		Engaged		In a romantic relationship		No relationship at present		Total	
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	Frequency	%
Male	49	51,04	2	2,08	19	19,79	26	27,08	96	100,00
Female	87	54,72	3	1,89	36	22,64	33	20,75	159	100,00
Total	136	53,33	5	1,96	55	21,57	59	23,14	255	100,00

Table 1 shows that 96 (37.65%) out of 255 people in the study group are males and 159 (62.35%) out of 255 people are females. 59 (23.14%) people have stated that they have no relationship at present though they have had one before. 136 (53.33%) people have said that they are married, 5 (1.96%) people said they are engaged and 55 (21.57%) have said that they are in a romantic relationship.

## 2.3. Data Collection Tools

Two different forms developed by the researchers were used to collect data. 17 people were given the instruction “you are supposed to specify at which situations you get angry considering your relationship”, and 81 items were developed in total. When these items were examined, it was discovered that some items were the same or they were similar to each other. Number of items was reduced to 57 by the researchers. Then expert opinions were consulted to academicians one of whom was from psychological counseling and guidance department and the other was from measurement and evaluation department, and 23 sample case items and 6 mood-state items were decided to use. Items were revised in accordance with experts’ suggestions of revision about short and clear expression. Hence, mood-state form consisting of 6 items (Form A) and sample case form consisting of 23 items (Form B) were created. Afterwards, Principal Component Analysis (PCA) was performed to determine the dimensionality of the forms. The analysis of Form A consisting of 6 stimulants was carried out by using polycoric correlation in FACTOR 10.9.02 program and analysis of Form B consisting of 23 stimulants using Pearson correlation in SPSS 25 program. The results of PCA for both scales determined to be unidimensional are given in Appendix A and B. Unidimensional structures were confirmed by Parallel analysis results calculated by FACTOR 10.9.02 program and Monte Carlo PCA Program by Marley W. Watkins.

## 2.4. Data Analysis

In the first step in which Form A was given to 255 observers and they were asked to rank between 1 and 6 according to their priorities, scale values ( $S_j$ ) were obtained by using ratios that were calculated according to observers’ rank-order judgments to 6 stimulants (scale items). Rank-order judgment is a scaling type whose validity is quite high because it enables to make the biggest discrimination between stimulants (Turgut & Baykul, 1992; Anıl & İnal, 2017).

Frequency matrix was created by using the sequence numbers (1= the least anger eliciting mood-state,...6=the most anger eliciting mood-state, reserve coding was performed during data analysis) that were given by the participants to anger mood-state items in Form A. Afterwards, pair comparison of each stimulant was made with other stimulants except for the stimulant itself by using the formula below, and frequency matrix was created for each stimulant.

$$n(S_{ji} > S_{ki}) = f_{ji} \cdot (f_{k<i} + \frac{1}{2} \cdot f_{ki})$$

j, k: stimulants' numbers

i: value given in ranking

$f_{ji}$ : the number of  $r_i$  sequence value given to  $U_j$  stimulant

$f_{ki}$ : the number of  $r_i$  sequence value given to  $U_k$  stimulant

To create proportion matrix (P), an upper triangular matrix is created by dividing column sums of each stimulant to  $N^2$ ; a lower triangular matrix is obtained by subtracting these values from 1. The Z standard values of the values of P matrix's each element is obtained and by means of them Z unit normal deviations matrix is created. Mean column values of this matrix give  $S_j$  scale values. Sc scale values are calculated by shifting  $S_j$  values in a way that the smallest  $S_j$  value is 0 (Albayrak Sarı & Gelbal, 2015). This process was carried out by Microsoft Excel 2013 program.

In the second step, 23 items in form B were given to 255 observers and they were asked to classify the items between 1 and 4 according to their anger level. There are some assumptions in the law of classifying judgment since they are asked to explain which sequence classes the stimulants belong to.

- 1- The structure can be divided into limited number of classes.
- 2- The boundary of any class is not an unmarried point, but a distribution that is called the distribution of boundary judgments.
- 3- When the observer chooses a class to put the stimulant in, the value of the stimulant is below the boundary value of that class (Turgut & Baykul, 1992). General formula of the law of rank-order judgment is as follows:

$$t_g - S_j = z_{jg} \cdot \sqrt{\sigma_j^2 + \sigma_g^2 - 2 \cdot r_{jg} \cdot \sigma_j \cdot \sigma_g}$$

$t_g$ : mean value of g boundary point

$\sigma_g$ : Standard deviation of observers' judgment belongs to g boundary

$\sigma_j$ : Standard deviation of observers' judgments belongs to  $U_j$  stimulant

$r_{jg}$ : The correlation between the perceived values of g boundary and  $U_j$  stimulant

$z_{jg}$ : Unit normal deviation of the ratio of number of placing g boundary that belongs to  $U_j$  stimulant into a lower boundary.

A frequency matrix is created by using the anger levels (1=little,...4=very) determined by the participants when they are exposed to anger-eliciting sample cases in form B, and then cumulative frequency matrix is obtained based on the columns. By dividing the elements of this matrix by the number of people, cumulative ratio matrix is obtained; by calculating z standard value of each element Z unit normal deviations matrix is obtained, and row means of this matrix are calculated. Through the graphics of lines  $y = mx+n$  drawn by using successive rows of z matrix, m and n values and  $a_j$  and  $s_j$  values are found out. By using these values scale values are calculated through the means of  $t_g$  boundary values, and then Sc scale values are calculated by shifting  $S_j$  values in a way that the smallest  $S_j$  value is 0. In this method,  $a_j$  and  $S_j$  values

are obtained by means of graphics (Anıl & İnal, 2017). This process was carried out by Microsoft Excel 2013 program as well.

### 3. FINDINGS

In this section, findings and comments are given related to scaling of data separately with the laws of rank-order and classifying judgment. Data was collected by using forms A and B, which were developed for the purpose of determining the situations causing anger in a relationship.

#### 3.1. Form A: Rank-Order Judgment Scaling

In this part, findings related to first and second research problems are given.

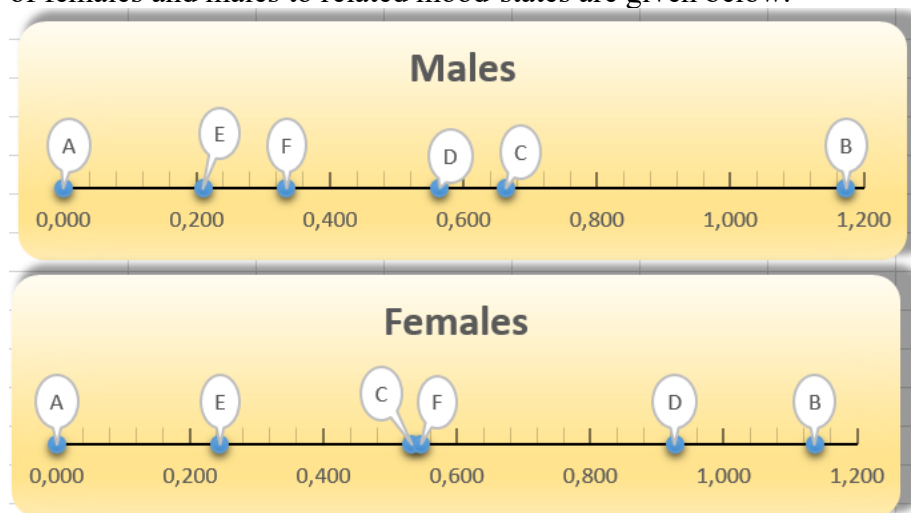
##### 3.1.1. Findings related to first research problem

In this section mood-states causing anger in 255 people’s (observers) relationships are scaled with rank-order judgment and scaling was performed for all participants and then for female and male participants separately.

**Table 2.** Ranking of Mood-States Causing Anger According to Females, Males and All Participants

Stimulants	All participants (N=255)		Males (N=96)		Females (N=159)	
	Stimulant ranks	Scale value	Stimulant ranks	Scale value	Stimulant ranks	Scale value
A Unfair treatment	1	0.000	1	0.000	1	0.000
B Being criticized	6	1.148	6	1.173	6	1.137
C Being neglected	4	0.580	5	0.663	4	0.532
D Arrogance	5	0.787	4	0.564	5	0.927
E Humiliation	2	0.232	2	0.210	2	0.245
F Mistrust	3	0.466	3	0.334	3	0.546

Table 2 shows that females, males and all participants get angry most at being treated unfairly and they get angry least at being criticized. Furthermore, it is seen that rankings of anger related to mistrust and humiliation are the same for females, males and all participants. Anger levels of females and males to related mood-states are given below.



**Figure 1.** Anger levels of females and males to 6 mood-states

Figure 1 demonstrates that females get angry more than males about worthlessness, and males get angry more than females about arrogance. When the mood-states causing anger were ranked between males and females, the most differentiating mood-state was arrogance.

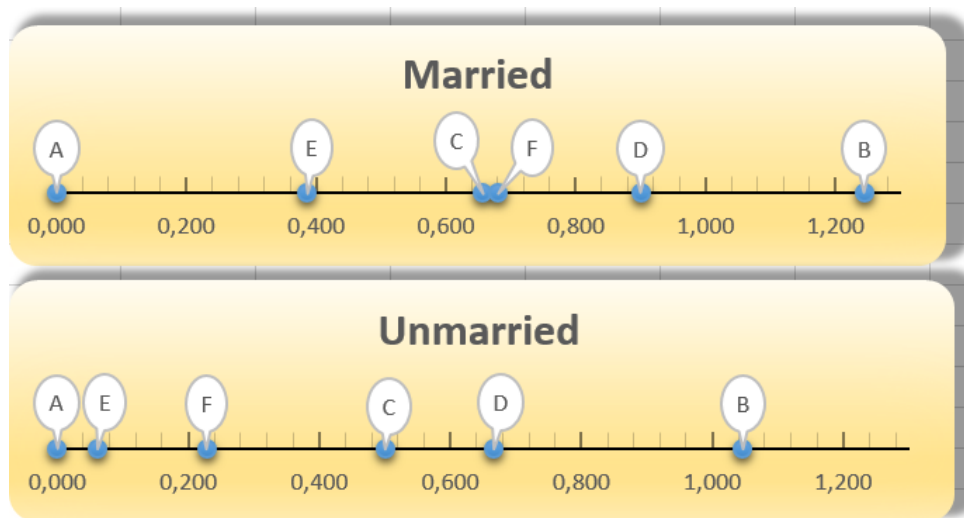
### 3.1.2. Findings related to second research problem

In addition to this analysis, participants' anger-eliciting mood-states were examined in terms of their marital status. Table 3 shows that married and unmarried participants get angry most at being treated unfairly, and they get angry least at being criticized. Furthermore, it is seen that the order of anger of married and unmarried people is the same in 'arrogance' and 'humiliation'. Anger levels of married and unmarried participants in related mood-states are shown below.

**Table 3.** Ranking of Mood-States Causing Anger According to Participants' Marital Status

	Stimulants	Married (N=136)		Unmarried (N=119)	
		Stimulant ranks	Scale value	Stimulant ranks	Scale value
A	Unfair treatment	1	0.000	1	0.000
B	Being criticized	6	1.244	6	1.047
C	Being neglected	3	0.657	4	0.500
D	Arrogance	5	0.900	5	0.666
E	Humiliation	2	0.386	2	0.062
F	Mistrust	4	0.681	3	0.228

Figure 2 shows that unmarried participants get angry more at mistrust whereas married participants get angry more at being neglected. It was found that when the mood-states causing anger were ranked between married and unmarried, the most differentiating mood-state was mistrust. Unmarried participants stated that they got angry at mistrust more.



**Figure 2.** Anger levels of married and unmarried participants to 6 mood-states

## 3.2. Form B: Classifying Judgment Scaling

In this part, findings related to third and fourth research problems are given.

### 3.2.1. Findings related to third research problem

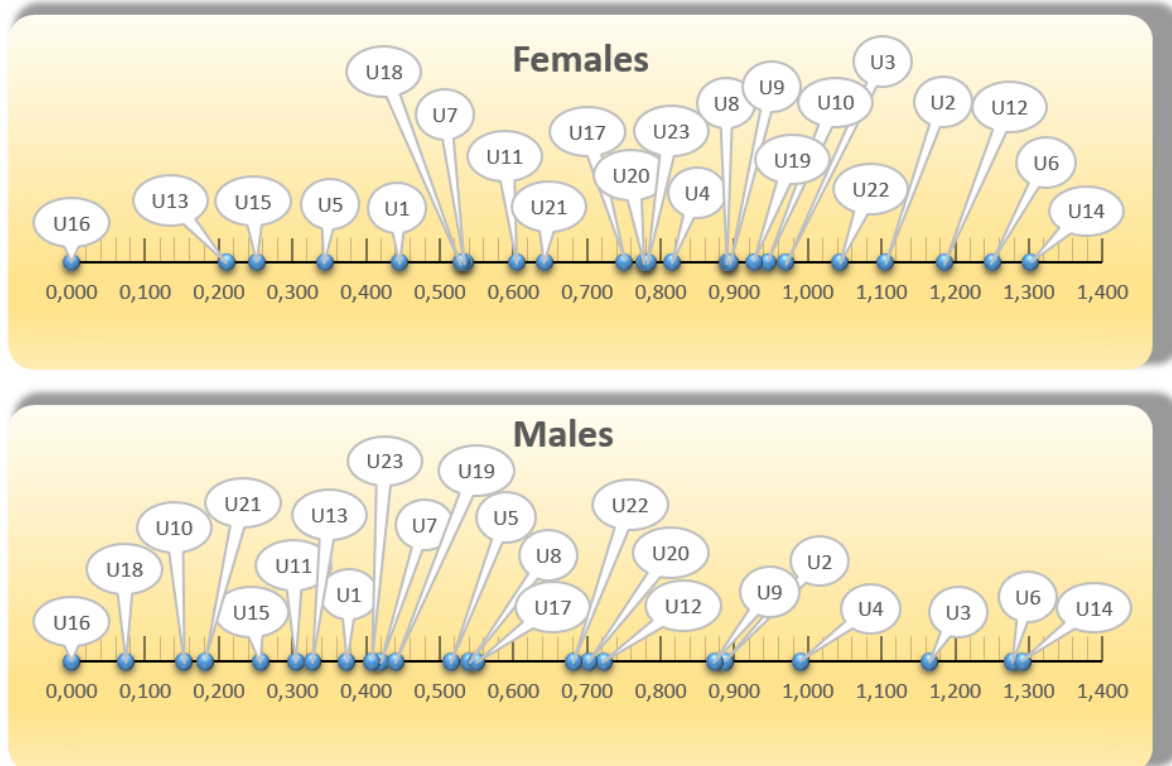
This section includes sample cases that are scaled with classifying judgment and cause anger in 255 people's (observers) relationships. Scaling was performed for all participants and then for female and male participants separately.

**Table 4.** Classification of Sample Cases Causing Anger According to Females, Males and All Participants

Stimulants	All participants (N=255)		Males (N=96)		Females (N=159)	
	Sti. ranks	Scale value	Sti. ranks	Scale value	Sti. ranks	Scale value
1 Not behaving according to etiquette (go on a visit empty-handed, to talk about politics everywhere, oratory etc.)	18	0.418	16	0.374	19	0.446
2 Ignoring the issues that I care about	4	1.012	5	0.888	4	1.105
3 Making the things that s/he does not want me to do	3	1.037	3	1.165	6	0.970
4 Making decisions without consulting me	8	0.879	4	0.991	11	0.815
5 Being extremely jealous of me	19	0.407	12	0.516	20	0.343
6 Not trusting me	2	1.258	2	1.278	2	1.250
7 Making huge amount of expenses without my knowledge	15	0.494	14	0.419	17	0.534
8 Trying to impose his/her ideas on me	9	0.750	11	0.540	10	0.890
9 Not being able to talk about any issue without a fight	7	0.884	6	0.873	9	0.894
10 Being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)	14	0.647	21	0.153	7	0.946
11 Behaving in an extremely calm and slow manner even in emergency work	16	0.487	18	0.305	16	0.604
12 Underestimating what I have done	5	1.006	7	0.722	3	1.185
13 Seeking for praise and tolerance all the time	21	0.253	17	0.327	22	0.210
14 Making unnecessary and offending comments about my family	1	1.295	1	1.291	1	1.303
15 Mess (socks, clothes etc.)	22	0.252	19	0.256	21	0.251
16 Being a fan of any subject (Team, political party etc.)	23	0.000	23	0.000	23	0.000
17 Having a harsh speaking style	12	0.679	10	0.549	14	0.751
18 Showing me as the bad cop when setting rules for the children	20	0.367	22	0.074	18	0.529
19 Not keeping his/her words on time	11	0.744	13	0.441	8	0.926
20 Being disrespectful towards my spare time	10	0.747	8	0.703	13	0.778
21 Spending too much time with technological devices or social environment	17	0.482	20	0.182	15	0.642
22 Failing to fulfill his/her responsibilities	6	0.908	9	0.682	5	1.044
23 Sharing the tasks unfairly	13	0.649	15	0.408	12	0.782

Table 4 shows that males, females and all participants get angry most when unnecessary and offending comments are made about their families. They get angry least about the situation that their partners are fan of any subject (team, political party etc.). Second situation they get angry most is mistrust. Anger levels of females and males to related sample cases are shown below. Figure 3 illustrates that females get angry more than males at the sample cases “trying to impose his/her ideas on me”, “being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)”, “behaving in an extremely calm and slow manner even in emergency work”, “underestimating what I have done”, “not keeping his/her words on time”, “spending too much time with technological devices or social environment”, “failing to fulfill his/her responsibilities”, “sharing the tasks unfairly”. It can be stated that males get angry more than females at the sample cases “not behaving according to etiquette (go on a visit empty-handed, to talk about politics everywhere, oratory etc.)”, “making the things that s/he does not want me to do”, “making decisions without consulting me”, “making extreme jealousy”,

“making huge amount of expenses without my knowledge”, “not being able to talk about any issue without a fight”, “seeking for praise and tolerance all the time”, “mess (socks, clothes etc.)”, “having a harsh speaking style”, “being disrespectful towards my spare time”. It was concluded that when the sample cases causing anger were ranked between males and females, the most differentiating sample case was “being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)”. Females stated that they got angry at this sample case more.



**Figure 3.** Anger levels of females and males to 23 sample cases

### 3.2.2. Findings related to third research problem

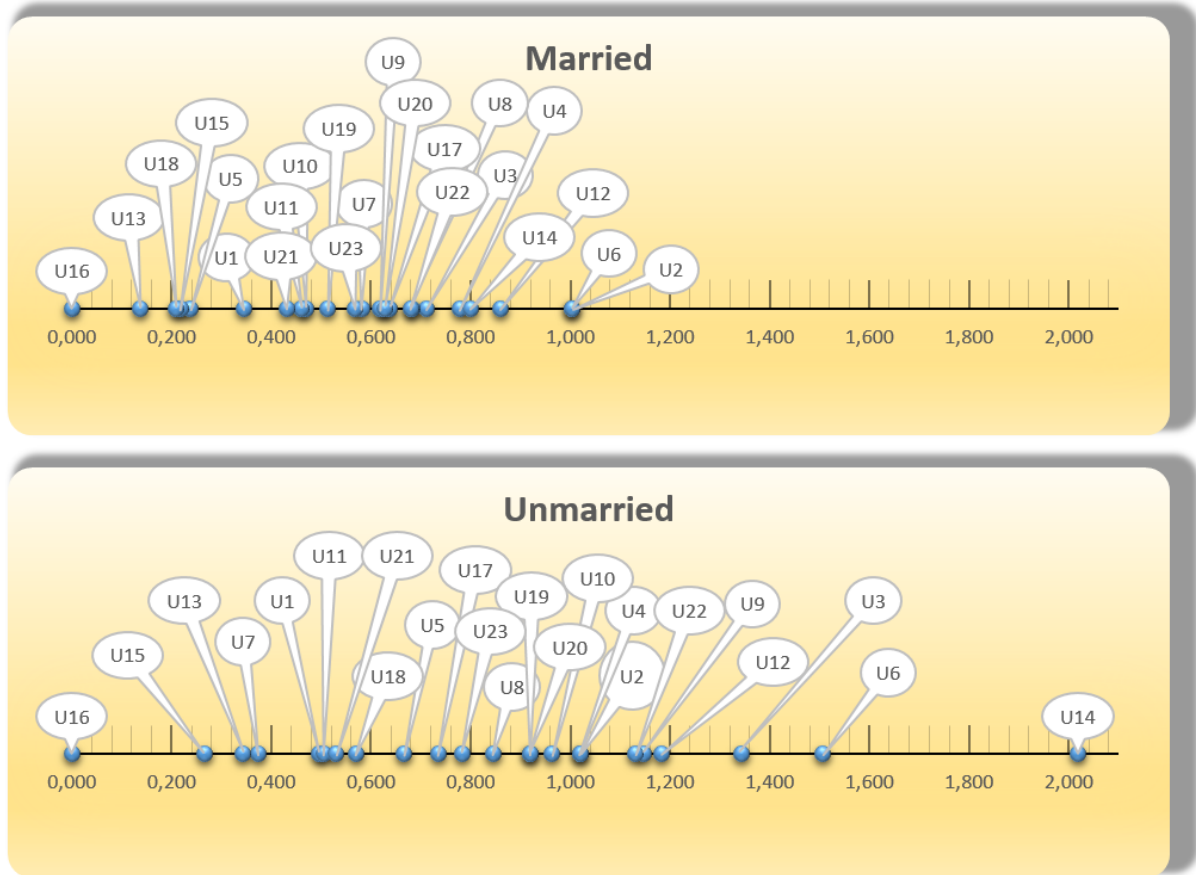
In addition to this analysis, participants’ anger-eliciting sample cases were examined in terms of their marital status. Table 5 reveals that married participants get angry most at the sample case “ignoring the issues that I care about” and unmarried participants get angry most at “making unnecessary and offending comments about my family”. It was also found that both groups got angry least at the sample case “being a fan of any subject (Team, political party etc.)”. It appears that rankings of anger related to sample cases “not trusting me”, “spending too much time with technological devices or social environment”, “being disrespectful towards my spare time” and “sharing the tasks unfairly” are the same for married and unmarried participants. Anger levels of married and unmarried participants to related sample cases are shown below.



**Table 5.** Classification of Sample Cases Causing Anger According to Participants’ Marital Status

Stimulants	Married (N=136)		Unmarried (N=119)	
	Sti. ranks	Scale value	Sti. ranks	Scale value
1 Not behaving according to etiquette (go on a visit empty-handed, to talk about politics everywhere, oratory etc.)	18	0.344	19	0.497
2 Ignoring the issues that I care about	1	1.004	7	1.021
3 Making the things that s/he does not want me to do	6	0.711	3	1.344
4 Making decisions without consulting me	5	0.777	8	1.020
5 Making extreme jealousy	19	0.238	15	0.667
6 Not trusting me	2	1.003	2	1.506
7 Making huge amount of expenses without my knowledge	12	0.581	20	0.375
8 Trying to impose his/her ideas on me	7	0.683	12	0.845
9 Not being able to talk about any issue without a fight	11	0.618	5	1.146
10 Being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)	15	0.469	9	0.963
11 Behaving in an extremely calm and slow manner even in emergency work	16	0.461	18	0.505
12 Underestimating what I have done	3	0.859	4	1.182
13 Seeking for praise and tolerance all the time	22	0.137	21	0.343
14 Making unnecessary and offending comments about my family	4	0.800	1	2.018
15 Mess (socks, clothes etc.)	20	0.218	22	0.265
16 Being a fan of any subject (Team, political party etc.)	23	0.000	23	0.000
17 Having a harsh speaking style	9	0.638	14	0.735
18 Showing me as the bad cop when setting rules for the children	21	0.210	16	0.570
19 Not keeping his/her words on time	14	0.514	11	0.919
20 Being disrespectful towards my spare time	10	0.628	10	0.920
21 Spending too much time with technological devices or social environment	17	0.431	17	0.530
22 Failing to fulfill his/her responsibilities	8	0.679	6	1.131
23 Sharing the tasks unfairly	13	0.568	13	0.782

Figure 4 shows that while married participants get angry at the sample cases “not behaving according to etiquette (go on a visit empty-handed, to talk about politics everywhere, oratory etc.)”, “ignoring the issues that I care about”, “making decisions without consulting me”, “making huge amount of expenses without my knowledge”, “trying to impose his/her ideas on me”, “mess (socks, clothes etc.)”, “underestimating what I have done” and “having a harsh speaking style” more than unmarried participants; unmarried participants get angry at the sample cases “making the things that s/he does not want me to do”, “making extreme jealousy”, “not being able to talk about any issue without a fight”, “being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)”, “seeking for praise and tolerance all the time”, “making unnecessary and offending comments about my family”, “not keeping his/her words on time” and “failing to fulfill his/her responsibilities” more than married participants. It was found that when the sample cases causing anger were ranked between married and unmarried participants, the most differentiating sample case was “making huge amount of expenses without my knowledge”. Married participants stated that they got angry at this sample case more.



**Figure 4.** Anger levels of married and unmarried participants to 23 sample cases

#### 4. DISCUSSION and CONCLUSION

In this study, the law of classifying judgment scaling was used for 6 items, and the law of rank-order judgment scaling was used for 23 items. Although the fact that judgments can be fully differentiated from each other in pair wise comparisons increases the consistency, it would be more useful to rank stimulants instead of this comparison as the number of items increases. As the number of items increases ranking would be more difficult and classification becomes more appropriate. Furthermore, differentiation of scale values with respect to which scaling method judge's decisions are obtained makes the method to be used important (Turgut & Baykul, 1992).

Apart from the result of the study that Balkaya (2001) conducted to reveal the differences/resemblances between anger and furiousness, this study aims to identify the reasons of anger which is one of the basic feelings in relationships. First, participants were asked to rank anger-eliciting mood-states between 1 and 6 from the most anger-eliciting mood-state to the least anger-eliciting one, and it was seen that regardless of their gender and marital status participants got angry most at unfair treatment, and they got angry least at being criticized. This may be associated with the fact that majority of the study group is composed of individuals whose education level is high and who are active working people; because it is likely that the people who are in communication with more than one person perceive criticism constructively.

Although there is no differentiation in the first two mood-states in terms of gender, the fact that females get angry at being neglected more than males, and males get angry at insensitivity, arrogance and mistrust more than females indicates that females get angry at general situations like not meeting their expectations, and males get angry at general situations like not being taken seriously. The fact that males get angry considerably at arrogance compared with females supports this result. In scale development study of Balkaya & Şahin (2003), situations causing anger are restricted to not being taken seriously, being treated unfairly and being criticized.

Likewise, even rankings of anger related to unfair treatment, humiliation, arrogance, being criticized are the same with regard to the marital status, it has been stated that married participants get angry more at being neglected while unmarried participants get angry more at mistrust. It is considered that there are differences due to the time spent together and interaction. These results correspond to results obtained by Kaynak (2014) even though the methods used are different.

Secondly, participants were asked to classify their anger at anger-eliciting sample cases between 1 and 4 from less to more. It was concluded that regardless of their gender and marital status all participants got angry least at the sample case "being a fan of any subject (Team, political party etc.)" and except for married participants they got angry most at sample case "making unnecessary and offending comments about my family". It was seen that married participants got angry most at the sample case "ignoring the issues that I care about". It is likely that this situation differs in married participants because of the expectation that in order to enjoy the time spent together, individual tastes should also be close to each other.

When sample cases are examined, the fact that females get angry more than males at the sample cases "trying to impose his/her ideas on me", "being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)", "underestimating what I have done" and "sharing the tasks unfairly" can be seen as a differentiation of anger levels of these cases with gender roles together with culture. The fact that females and males give the most differentiating reaction to sample case "being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry, etc.)" supports this judgment. It can be reported that it is also similar to Özmen Süataç (2010)'s idea that anger and reactions to anger should be treated as individual differences.

Similarly, even if rankings of anger related to sample cases "not trusting me", "spending too much time with technological devices or social environment", "being disrespectful towards my spare time" and "sharing the tasks unfairly" are the same in terms of the marital status; married participants stated that they got angry more at sample cases "not behaving according to etiquette (go on a visit empty-handed, to talk about politics everywhere, oratory etc.)", "ignoring the issues that I care about", "making decisions without consulting me", "making huge amount of expenses without my knowledge", "trying to impose his/her ideas on me", "underestimating what I have done", "mess (socks, clothes etc.)" and "having a harsh speaking style". Unmarried participants explained that they got angry more at sample cases "making the things that s/he does not want me to do", "making extreme jealousy", "not being able to talk about any issue without a fight", "being extremely connected to the gender roles (women do the cleaning, men earn money, men do not cry etc.)", "seeking for praise and tolerance all the time", "making unnecessary and offending comments about my family", "not keeping his/her words on time" and "failing to fulfill his/her responsibilities". It is considered that there are differences due to the time spent together and interaction as well. Similar to the studies of Erok (2013) and Togay (2016), the fact that the most differentiating sample case is "making huge amount of expenses without my knowledge" is considered to be associated with common investments such as common budget or being aware of the expenses in marriage.

Considering the results of this study, in order to protect family unity which is the most basic unit of the society and establish it in an appropriate way, it is recommended to conduct mixed researches to investigate the possible causes of anger mentioned above. In addition, it is suggested that crime prevention studies should be performed by developing awareness programs aimed at understanding the reasons behind anger. Comparisons of different scaling methods calculated on the same stimulants are also recommended.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

## ORCID

Merve YILDIRIM SEHERYELİ  <https://orcid.org/0000-0002-1106-5358>

Duygu ANIL  <https://orcid.org/0000-0002-1745-4071>

## 5. REFERENCES

- Acar Güvendir, M., & Özer Özkan, Y. (2013). A comparison of two scaling methods: Pair wise comparison and rank-order judgments scaling. *Journal of Educational Sciences Research*, 3(1), 105-119.
- Albayrak Sarı, A., & Gelbal, S. (2015). A Comparison of Scaling Procedures Based on Pair-Wise Comparison and Rank-Order Judgments Scaling. *Journal of Measurement and Evaluation in Education and Psychology*, 6(1), 126-141.
- Anıl, D., & Güler, N. (2006). An example of the scaling study by pair-wise comparison method. *Hacettepe University Journal of Education*, 30, 30-36.
- Anıl, D., & İnal, H. (2017). *Psikofizikte ölçekleme uygulamaları*. Ankara: Pegem Akademi.
- Balkaya, F. (2001). *The Development of multidimensional anger inventory and effect on some symptom groups*. Unpublished MA Thesis. Ankara University, Graduate School of Social Sciences, Ankara.
- Balkaya, F., & Şahin, N. H. (2003). Multidimensional Anger Scale. *Turkish Journal of Psychiatry*, 3, 192-202.
- Bozgeyikli, H., Toprak, E., & Derin, S. (2016). Teacher candidates' career values perceptions by rank order judgments scaling. *HAK-İŞ International Journal of Labor and Society*, 5(11), 204-225.
- Bülbül, S., & Köse, A. (2010). Examining between regional internal migration movements in Turkey with multidimensional scaling. *Istanbul University Journal of the School of Business Administration*, 39(1).
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2017). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.
- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. USA: Cengage Learning.
- Demirus, K. B., & Gelbal, S. (2020) Scaling of Students' Instructional Techniques Preferences used in Science Lessons. *Kastamonu Education Journal*, 28(1), 154-170.
- Erdoğan, M. (2018). *Relationship between early maladaptive childhood schemas with trait anger and anger expressions*. Unpublished MA Thesis. Ankara University, Graduate School of Health Sciences, Ankara.
- Erok, M. (2013). *Interpersonal cognitive distortions, relationship beliefs, interpersonal anger, interpersonal relationship, problem solving and marital conflict*. Master Thesis. Maltepe University, Graduate School of Social Sciences, İstanbul.
- Güler, N., & Anıl, D. (2009). Scaling through pair-wise comparison method in required characteristics of students applying for post graduate programs. *International Journal of Human Sciences*, 6(1), 627-639.
- Güvendir, M., & Özkan, Y. (2013). A comparison of two scaling methods: Pair wise comparison and rank-order judgments scaling. *Journal of Educational Sciences Research*, 3(1), 105-119.
- Karasar, N. (2016). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayıncılık.
- Kassinove, H., & Suckodolsky, D. (1995). *Anger disorders: basic science and practice issues*.

- Issues in Comprehensive Pediatric Nursing*, 18, 173-205.
- Kırkök, F. (2017). *The relationship between the levels of depression, alexithymia and anger of the individuals who have divorced and non-divorced parents*. Unpublished MA Thesis. Üsküdar University, Graduate School of Social Sciences, İstanbul.
- Özkan, M., & Arslantaş, H. İ. (2013). A study of scaling with ranking judgment method on characteristic of effective teacher. *Trakya University, Journal of Social Sciences*, 15(1), 311-330.
- Özmen Süataç, A. (2010). *Marital adjustment research in terms of interpersonal style and anger*. Master Thesis. Ege University, Graduate School of Health Sciences, İzmir.
- Satıcı, S. A. (2014). Anger Rumination Scale: Psychometric properties of the Turkish version. *Anatolian Journal of Psychiatry*, 15, 328-334.
- Sayın, A., & Gelbal, S. (2014). Başarıyı Etkileyen Faktörlerin Önem Derecelerinin Ardışık Aralıklar Yöntemiyle Ölçeklenmesi. *Amasya Üniversitesi, Eğitim Fakültesi Dergisi*, 3(1), 1-26.
- Soykan, Ç. (2013). Öfke ve öfke yönetimi. *Kriz dergisi*, 11(2), 19-27.
- Tezbaşaran, E. (2017). Scaling types of test with the method of absolute judgement and assessing them with The students' views. *HAYEF Journal of Education*, 14(1), 143-162.
- Togay, A. (2016). *Relations between self concealment, anger expression style and authenticity in married persons*. Unpublished MA Thesis. Hacettepe University, Graduate School of Educational Sciences, Ankara.
- Turgut, M. F., & Baykul, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.
- Tüzüntürk, S. (2009). Çok boyutlu ölçekleme analizi: suç istatistikleri üzerine bir uygulama. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 28(2), 71-91.
- Uğurcan, B. (2018). *Stress, depression, anxiety and alexithymia in psoriasis*. Master Thesis. İstanbul Bilim University, Graduate School of Social Sciences, İstanbul.
- Yapabaş, S. İ. (2018). *The relationship between eating behaviours, anger and codependency on overweight and obese individuals*. Unpublished MA Thesis. Işık University, Graduate School of Social Sciences, İstanbul.
- Yaşar, M. (2016). Scaling the features considered to have affected the academic success of teacher candidates on the basis of rank-order judgement scaling technique. *Pamukkale University Journal of Education*, 40, 274-288.
- Yılmaz Koğar, E., & Demircioğlu, E. (2016). A Scaling study about 6.-7.-8. grades mathematics learning domain. *Başkent University Journal of Education*, 3(1), 44-52.

## 6. APPENDIX

### 6. 1. Appendix A: Form A

**Table A1.** *KMO and Bartlett's statistic results*

KMO	0.639
Bartlett's statistic	326.1
p	0.00

**Table A2.** *Eigenvalues*

Variables	Eigenvalues	Proportion of Variance	Cumulative Proportion of Variance
1	2.391*	39.90	39.90
2	1.026	17.11	
3	0.912	15.23	
4	0.706	11.80	
5	0.676	11.32	
6	0.288	4.81	

\*Advised number of dimation according to Parallel Analysis test results.

**Table A3.** *Factor Loading and Reliability Coefficient*

Variables	F1
1	0.633
2	0.564
3	0.626
4	0.597
5	0.684
6	0.676
Cronbach Alpha	0.698
Omega	0.820

## 6. 2. Appendix B: Form B

**Table B1.** *KMO and Bartlett's statistic results*

KMO	0.887
Bartlett's statistic	2076.22
p	0.00

**Table B2.** *Eigenvalues*

Variables	Eigenvalues	Proportion of Variance	Cumulative Proportion of Variance
1	7.578*	32.946	32.946
2	1.473	6.403	
3	1.349	5.865	
4	1.148	4.993	
5	1.062	4.617	
6	1.003	4.361	
7	0.961	4.178	
8	0.867	3.769	
9	0.782	3.400	
...			

\*Advised number of dimation according to Parallel Analysis test results.

**Table B3.** *Factor Loading and Reliability Coefficient*

Variable	F1	Variable	F1
1	0.743	13	0.542
2	0.728	14	0.530
3	0.723	15	0.522
4	0.701	16	0.486
5	0.680	17	0.473
6	0.677	18	0.416
7	0.648	19	0.400
8	0.636	20	0.376
9	0.623	21	0.426
10	0.603	22	0.458
11	0.557	23	0.450
12	0.547		
Cronbach Alpha	0.904		

## A Study on Developing Scale for Teacher Perceptions towards Spelling Rules

Ali Turkel <sup>1,\*</sup>

<sup>1</sup> Dokuz Eylul University, Buca Education Faculty, Izmir, Turkey

### ARTICLE HISTORY

Received: 01 October 2019

Revised: 14 February 2020

Accepted: 07 March 2020

### KEYWORDS

Spelling mistakes,  
Scale development,  
Teacher perceptions,

**Abstract:** The purpose of this study is to develop a scale to determine Primary Education and Turkish Language teachers' perceptions regarding the frequency of spelling mistakes which their students make. Therefore, this experimental form made to serve this goal was presented to field specialists in terms of consultation; and each item in the form was regulated to ensure content validity rates in accordance with the feedback provided by the specialists. Items with the validity rate below 0.80 were omitted from the form. The trial form consisting of 34 items was administered to 232 Primary Education and Turkish Language teachers who teach at schools under the jurisdiction of Ministry of National Education (MoNE) via e-mail, and the gathered data were analyzed. With the help of Exploratory Factor Analyses (EFA), a four-dimensioned construct with 19 items including frequently made mistakes regarding acronyms, spelling of conjunctions and suffixes, spelling of capital letters, spelling of compound words, and the spelling of the words affected by word formation processes. In the analyses, the relations between the sub-scales of the original scale were taken into consideration, revealing that factors had positive and significant relationships and sub-dimensions were the constituents of the general structure named spelling mistakes, which made up the upper structure. Goodness of fit indices (GFI) of the model was detected to be quite high. Confirmatory Factor Analysis (CFA) administered to the second research group justified the EFA results. The internal consistency coefficient, calculated as .91, for the entire scale was found to be quite reliable.

## 1. INTRODUCTION

Writing, as acknowledged, is one of the four basic skills to be developed in the teaching of the first language and includes various dimensions. Both Turkish Language and Primary Education teachers aim at developing these various sub-dimensions in harmony with a careful consideration for balance in the process of teaching. One of the sub-dimensions to be developed is the skill for implementing spelling rules. In the teaching curricula (Ministry of National Education [MoNE], 2006, 2008), spelling rules planned to be taught were presented separately according to grade levels. Starting from the first grade in primary school, teaching of spelling rules gradually moves from easy to hard. However, the literature focusing on the application of these rules reflects certain complaints concerning the high number of problems related to learning these rules.

---

CONTACT: Ali Türkel ✉ [ali.turkel@hotmail.com](mailto:ali.turkel@hotmail.com) 📧 Dokuz Eylul University, Buca Education Faculty, Turkish Language Teaching Dept. İzmir, Turkey

ISSN-e: 2148-7456 /© IJATE 2020



According to Bayat (2013), language of writing (written language) is the language of education and enlightenment. In his study where he focused on pre-service teacher's writing process, Bayat (2013) advocates that the written language of pre-service teachers can provide clues about their teaching competencies. In a general perspective, writing skills of the students can generate outputs associated with their general thinking skills.

According to Aksoy (1985, 1990) language of writing is of three stages in terms of its narration features as follows: accurate writing, good writing and calligraphy. In *accurate writing*, which is the first prerequisite and step of writing, it is fundamental to express the aimed meaning in accordance with complete and precise language rules. Indeed, writing accurately is important for people to communicate correctly, to fulfill the functions necessary for their lives, and essentially to express themselves correctly in every field. Thusly, spelling rules and punctuations are the constructs which set a series of rules for the sake of accuracy to prevent misunderstandings and form a shared written language among people. In the up-to-date Turkish Dictionary of Turkish Language Institution (2019), *writing* is defined as “*Transcription of a language into written form by following certain rules, spelling*” while *spelling rules* is depicted as “*The rules determining the ways the words in a language are written*”. The ambiguity that emerges when spelling rules are not abided prevents the messages from being transferred as intended (Kıbrıs, 2010, p.128). Aksoy (1985,1990), who defined spelling mistake as “the spellings which do not follow the rules in spelling dictionary and the spelling in the word index of this dictionary”, stated that there are various spelling mistakes committed.

In Turkey, rules concerning the spelling are set with efforts of committees elected by Turkish Language Institution (TLI), and these committees hold the right to make changes and updates to the rules from time to time. Even though these changes and updates can possibly cause confusions, it seems only natural for this institution to make changes in its views at some points in time. Whereas European countries like Germany, Portugal, Greece, and France possess such a dictionary as in Turkey, some countries like England and Italy do not have similar spelling dictionaries. It is acceptable that Turkey, who adopted the Latin alphabet not until 1928, experiences problems related to spelling more often than other Western countries do. Besides, it is known that Turkey faced debates regarding spelling even when she still used Arabic alphabet (Ünver, 2008; Kannas, 2019).

Considering the literature, the prominent reasons for students to make spelling mistakes are the lack of knowledge about the correct spelling of a word and forgetting the spelling of words due to learning them incorrectly or using them very rarely. These circumstances can be explained with lack of sufficient reading and writing. There can be uses defying the rules of grammar and Spelling Dictionary such as the misuse of suffixes (ağlıyan, yapmassa), failure to employ word formation processes (kitabı, ağacı), writing of “ki, de” conjunctions and “mi” adverb, using slang and local dialects, incorrect hyphenation of words at the end of the lines, writing the words as pronounced in spoken language (peki, abi, etc.), the use of intensive adjectives (bembeyaz, sapsarı), confusion of resembling words (eğer-eyer, öyle-öğle, saç-sac), the use of circumflexes and apostrophes, writing of Arabic and Persian noun phrases, writing of proper nouns, titles, numbers, acronyms, and foreign words (Aktaş & Gündüz, 2003, p.115-116).

In addition to the study conducted by Aktaş and Gündüz (2003) in which a classification was made regarding the topics where students committed mistakes, the literature includes studies that dealt with students' levels of accuracy in spelling. Among the studies, Bağcı (2011) revealed in his research concerning the 8<sup>th</sup> grades that, with a rate of 69%, topics where the students were least successful were the spellings of suffix “-ki”, conjunction “de”, suffix “-DA”, and softening of the consonents (Bağcı, 2011, p.702). In another research, Karagül (2010) attempted to determine the 6<sup>th</sup> through 8<sup>th</sup> grade students' levels of implementing spelling and punctuation rules suggested by Turkish class syllabus. As for the results of the study, the 6<sup>th</sup>

grade students generated a success rate of 70% in implementing spelling rules whereas the 7<sup>th</sup> graders showed a success rate below 70% concerning “spelling the words that can be confused with one another”, “spelling the numbers”, and “spelling the capital letters”. Additionally, considering the 8<sup>th</sup> graders, even though they were successful at spelling rules at 70% rate, they made frequent mistakes in the sense of spelling “ki conjunction” and “mi question tag”.

Besides the students’ justifications for committing spelling mistakes as determined by Aktaş (2003), a study carried out by Türkel, Yaman and Aksu (2017) accounts for the dimension of teachers. That is, the study by Türkel et al. (2017) focusing on Turkish language and primary education teachers’ perceptions regarding spelling rules and spelling mistakes of students, revealed that teachers reported problems caused by the spelling rules of TLI. Accordingly, it was found that teachers regarded the spelling rules established by TLI as inconsistent at 66% rate. In connection, the teachers explained this inconsistency with frequently changing rules, excessive number of exceptions, the lack of consensus on the explanations, and inclusion of memorized items. Furthermore, to teachers, the increase in students’ accurate spelling can be ensured by presenting reasonable and fixed rules, diminishing exceptions, encouraging reading books rather than imposing a rule-oriented nature, and avoiding frequent changes in rules in the spelling guide.

Bayat (2019), in the study that focused on text construction process, highlighted fundamental constituents of the process, and underlined the implementation of spelling rules as one of the three important factors that lead to exposing the quality of the text’s internal structure.

While listing the stages employed in the writing process, Flower and Hayes (1981) mentioned revision process which included evaluation and editing sub-skills. Editing contains internal elements as well as external elements such as spelling and punctuation. Similar thought resurfaces in the last step of Kellogg’s writing model, and Kellogg stated that this step comprises of reading and editing (Kellogg, 1996). The editing concept appears as the last step of Zamel’s (1983) Cognitive Writing Process. Cho (2003), Gebhard (1983), Sommers (1988), Alamargot and Chanquoy (2001), who dealt with writing process and attempted to determine how this process was handled in various writers, emphasized similar concepts by using certain terms.

To equip students with an effective written expression skill, it is vital to determine the mistakes committed in the context of writing. It is not other than teachers who would know about the frequency of mistakes committed by students in a sub-heading. Therefore, a scale to categorize teacher views related to students’ spelling mistakes would be quite beneficial. To help reduce spelling mistakes, making sense out of the mistake process by categorizing the mistakes with the help of a teacher perceptions scale and identifying the steps to be taken for the sake of a solution might provide a considerable support. Motivated by this, a scale towards collecting Primary Education and Turkish Language teachers’ perceptions was developed to determine the frequency of spelling mistakes committed by the students.

## 2. METHOD

In the methodology section of the research, the research model, participants, the development process of the data collection tool, production of the items, preparation of the trial form, ensuring the content validity, data collection process, and pre-test stage are included.

### 2.1. Research Model

The study takes on a survey research design which aims to determine the construct validity and internal consistency of the “Scale for Primary Education and Turkish Language Teacher Perceptions towards Spelling Rules” and to test the model. Survey research is used for the identification of a specific group’s features (Büyüköztürk et al., 2016) and the depiction of the group’s thoughts and perceptions about a subject area (Lodico, Spaulding, & Voegtler, 2006).

## **2.2. Research Group (Participants)**

The research universe of the study includes Primary Education and Turkish Language teachers who teach at primary and secondary level schools under the jurisdiction of MoNE in Turkey during 2016-2017 academic years. Two independent participant groups took place in the research. For the both stages of the research, random sampling out of the sampling models was used to conduct sampling process. In the process of forming participant groups, the groups were confined simultaneously and some prerequisites were set as follows: they taught written expression course, were voluntary, and worked in different cities and schools for the sake of maximum diversity. Background information forms were exploited in the acquisition of personal data.

The first participant group included 226 teachers. Factorial structure of the scale was established and a reliability study was conducted on the data gathered from this group. 21, 32% (n=48) of these teachers was Primary Education while 78,8% (n=178) was Turkish Language teachers. In addition, 43, 4% (n=98) of the participants was male whereas 56,6% (n=128) was female teachers.

On the data gathered from the second participant group, whether the factorial structure plotted from the scale was confirmed was investigated. This participant group included 236 teachers, 20.3% (n=48) of which was Primary Education and 79.7% (n=188) was Turkish Language teachers. Moreover, 41.1% (n=97) of the group was male while 58.9% (n=139) was female teachers.

## **2.3. Ensuring the Appropriateness of the Data and Development of the Data Collection Tool**

In scale development studies, it is recommended to take some common steps listed as follows: identifying the general need, establishing a theoretical structure, consulting with scholars, initializing the scale, piloting, administering EFA and CFA (Seçer, 2015). Based on this, in the development of the scale, initially, a literature review was conducted, and then a pool of items was constructed to be followed by the consultation with scholars. Following these steps, results of exploratory factor analysis and reliability computations were obtained to detect the reliability and validity of the research, and the model was tested via confirmatory factor analysis.

Prior to statistical analyses, assumptions related to normality, missing values, and outliers were taken into consideration for the model construction. Regarding the normal distribution of the data, measures of central tendency were detected to be close to each other. In addition, values obtained from the division of Skewness and Kurtosis amounts with the standard error were computed between -1.96 and +1.96 range. Lastly, Kolmogorov-Smirnov and Shapiro-Wilk  $p$  values did not generate any statistical significance (Can, 2013). Furthermore, Barlett Sphericity test suggested that the sample provided normal distribution conditions (Şencan, 2005; Brace, Kemp, & Snelgar, 2006; Tavşancıl, 2018). Considering the issue with regards to missing values, there was no missing values in the data due to online data collection procedures. To identify the outliers in the data set, Mahalanobis distances were calculated. It is required to use the value of  $p < .001$  in Mahalanobis distances (Tabachnick & Fidell, 2007). Therefore, in this respect, six outliers were extracted from the sample of 226 participants. Normality values and outliers were screened to ensure the appropriateness of the data for the administration of Confirmatory Factor Analysis. For this purpose, data gathered from 236 participants were examined.

## **2.4. Writing the Items and Preparation of the Trial Form**

To develop the scale, the related literature was primarily reviewed, and in the determination of the items, 37 perceptions that served the purpose of the research were written down by taking the mistakes made by students in their written expressions into consideration. A special

sensitivity was shown towards explicitness and legibility of the items by selecting expression with a single judgement.

Following this stage, specialist feedback was requested from the scholars working in Turkish Language Teaching and Turkish Language and Literature departments. A 36-item trial form was produced after making necessary revisions related to language and expression of the items based on the evaluation of the scholars.

The trial form including 36 items used 5-item Likert Scale model which expected participants' perceptions about the mistakes their students committed by responding to options as follows: "Never" (1), "Rarely" (2), "Sometimes" (3), "Often" (4), and "Always" (5). Instructions were adhered to the scale's information section which described the purpose of the scale and the correct way for responding to items.

## **2.5. Content Validity**

To ensure content validity, the trial form of 36 items was presented to five scholars in the field for their perceptions. A three-level rating scale was used to receive scholars' feedback, requesting the scholars to choose among "appropriate", "partially appropriate", and "not appropriate" options for each item. The content validity of the items based on scholar perceptions was determined by referencing the content validity rate developed by Veneziano and Hopper (1997). According to the rate taken as reference, the main requirement was to take the minus one of the ratios of the number of scholars responding positively to the total number of the scholars. Thus, two items whose content validity rates were under .80 were omitted to establish the ultimate trial form.

## **2.6. Piloting Stage and the Collection of the Data**

So as to test the reliability and validity of the draft perception scale with 34 items prepared, 232 Primary Education and Turkish Language teachers (the first participant group) that taught at different schools tied to MoNE during 2016-2017 academic year were administered a pilot application.

The draft scale planned for the piloting stage was delivered to 232 teachers in total who taught at MoNE schools via e-mail during 2016-2017 academic year, and their responses were saved online for individual analysis. Furthermore, personal information of teachers was collected through background data collection forms.

## **3. RESULT / FINDINGS**

This section presents the evidence with regards to reliability and validity of the scale that was piloted.

### **3.1. Construct Validity**

Construct validity is about the extent of measurability of the scores in relation with the concept (construct) intended to be measured (Cohen & Swerdlik, 2009; Çokluk, Şekercioğlu, & Büyüköztürk, 2016). In relation, Factor Analysis (FA) seeks answer to the question "Do the scores obtained by this test measure the construct that the test assumes to be measuring?". If the point in investigating the construct validity is to reveal the factorial status of the scale, "exploratory factor analysis" techniques are used. On the other hand, if the objective is to confirm the predetermined factorial structure of the scale, "confirmatory factor analysis" techniques are preferred (Büyüköztürk et al., 2016). In this direction, both Exploratory and Confirmatory Factor Analysis techniques were used to provide evidence for the construct validity of the scale in the research.

### 3.1. Exploratory Factor Analysis

In the identification of the factors in the study, Principal Components Analysis (PCA) with which a perspective from the components of the theoretical structure to the fundamental dimensions was provided was used. PCA is required to be used especially when the researcher's aim is to develop a scale (Şencan, 2005). A rotation is possible during the factor extraction (Büyüköztürk, 2016). Factorial structures are evaluated more easily through rotation procedures (Akbulut, 2010). For this reason, out of vertical rotation methods, maximized variation (varimax) was used.

Additionally, correlation values are observed to evaluate the appropriateness for EFA, which requires inter-item correlation values between .10 and .90 (Özdamar, 2016) and majority of these values above .30 (Field, 2009). Following the analyses, it was revealed that majority of the correlation values were .30 and all the correlations between the items were statistically significant ( $p < .05$ ).

To determine the suitability of the data gathered, Kaiser-Meyer-Olkin (KMO) coefficient and Barlett Sphericity test were administered. The size of the sample can be deduced with the help of KMO value (Şencan, 2005; Büyüköztürk, 2016) while Barlett Sphericity test provides insight on the sample's conditions of normal distribution (Şencan, 2005; Brace, Kemp & Snelgar, 2006; Tavşancıl, 2018). As the result of the analysis, KMO coefficient was calculated as (.896). In addition to the KMO value that was considerably close to 1, and the result of Barlett's test of Sphericity calculated as (1976.141;  $p < .05$ ) revealed that the samples of the study were sufficient and appropriate for the analysis (Akgül & Çevik, 2003). Moreover, a scree plot was generated to determine the number of factors that could illustrate the interaction among items. Regarding the eigenvalues' contribution to the variance, scree plot is regarded as quite useful for reducing the number of factors on the account of visibility (Çokluk et al., 2018). The related information concerning the scree plot used for determining the number of factors in the developed scale is presented in Figure 1.

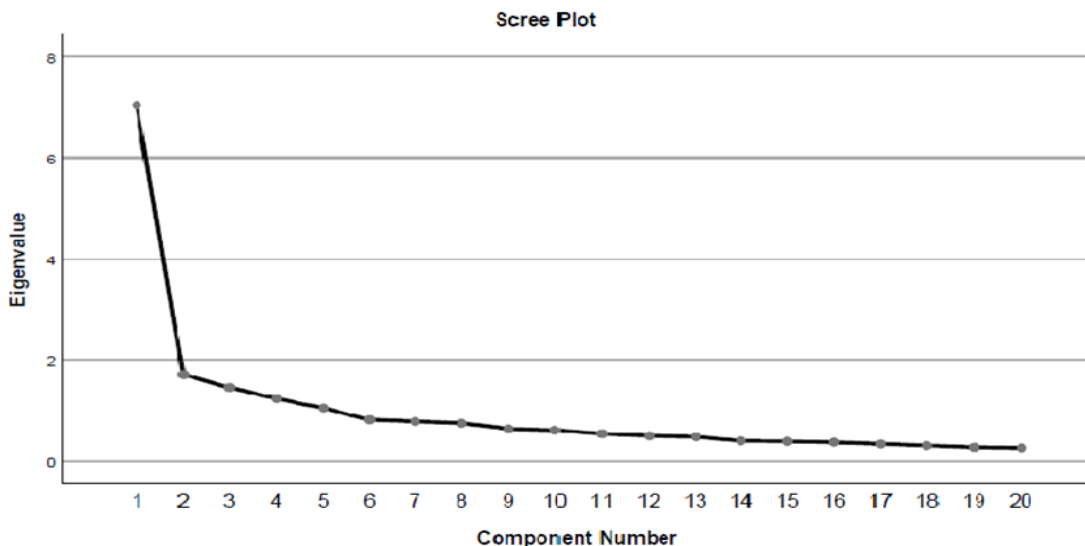


Figure 1. Screen Plot Graphic

Considering the graphic in Figure 1, it was detected that the break point (the point where the highly-accelerated rapid decrease stops) actualized after the fourth factor. It is explained that the breaking point indicates the factorizing owing to the examination of scree plot graphic. Moreover, it is recommended that, according to Kaiser Criterion, factors with eigenvalues above 1 be kept during factor extraction (Büyüköztürk, 2016). Thus, a structure with four factors was established ensuring that it was related to the theoretical construct. 15 items were

removed from the analysis due to either factor loading values being under 0.40 (Ferguson & Takane, 1989) or loadings for more than one factor with a loading value below .10 (Büyüköztürk, 2016). Table 1 presents the results of factor analysis and reliability analysis for the clusters downgraded to 19 items and loaded around 4 sub-dimensions.

**Table 1.** Factor Analysis Results after Varimax Rotation

Factors and Items	Explained Variance (%)	Eigenvalue	$\bar{X}$	SD	Item total r	Factor load
<b>Factor 1 (<math>\alpha=.83</math>)</b>						
Item 1	14.509	2.757	2.98	.802	.45	.73
Item 2			2.40	.890	.46	.74
Item 3			2.55	.884	.56	.79
Item 4			2.17	1.042	.60	.77
<b>Factor 2 (<math>\alpha=.83</math>)</b>						
Item 9	16.434	3.122	3.50	.818	.55	.70
Item 12			3.36	.790	.52	.74
Item 14			3.03	.821	.54	.69
Item 15			2.65	.877	.56	.72
Item 16			2.96	.847	.54	.72
<b>Factor 3 (<math>\alpha=.82</math>)</b>						
Item 7	15.919	3.025	2.34	1.008	.40	.64
Item 10			2.49	.818	.44	.56
Item 20			2.26	.997	.52	.74
Item 21			1.87	.849	.50	.73
Item 22			2.07	.879	.52	.73
<b>Factor 4 (<math>\alpha=.82</math>)</b>						
Item 28	15.784	2.999	2.31	.835	.45	.687
Item 29			2.71	.958	.41	.619
Item 30			2.44	.879	.53	.788
Item 31			2.87	.833	.48	.739
Item 32			2.33	.914	.51	.668
<b>Overall (<math>\alpha=.91</math>)</b>			62.646	11.894	2.59	10.449

According to Table 1, reliability values for Factor 1 through Factor 4 were calculated as .83, .83, .82, and .84 in respective orders whereas reliability coefficient of .91 was generated for the whole scale. As Bayram (2004) pinpointed, a Cronbach's Alpha value above .70 can be regarded as appropriate in terms of reliability. This outcome, therefore, indicates that the scale has a high reliability level.

Additionally, it was revealed that the finalized scale was comprised of 4 factors in total and these factors explained 62.6% of the variance. Taking item contents into consideration, the four dimensions were categorized as the most commonly made mistakes on 1- *spelling of acronyms, conjunctions, and suffixes*, 2- *spelling of capital letters*, 3- *spelling of compound letters*, and 4-

spelling of words that went through word formation processes. In social sciences, the ideal range for the explained variance is acknowledged as between 40-60% (Scherer, 1988).

Furthermore, it is suggested that correlations between sub-scale and the total scale scores should be reported (Pallant, 2011). It is observed that the correlation between sub-scales ranges from .45 to .59. Whereas, the correlation score between sub-scales and the total scale in the range of .75 and .83 indicates a statistically significant relationship. Findings regarding the correlations between sub-scales and the total score are described in [Table 2](#).

**Table 2.** Correlations between Sub-scales and Total Score

Sub-dimensions*	Factor 1	Factor 2	Factor 3	Factor 4	Total Score
Factor 1	1				
Factor 2	.45**	1			
Factor 3	.53**	.53**	1		
Factor 4	.59**	.54**	.56**	1	
Total Score	.75**	.79**	.83**	.81**	1

\*n=226; \*\*p<.001

As can be referenced in [Table 2](#), significant correlation in a positive direction is observed among sub-dimensions of the scale and between each factor and the whole scale. The obtained results can serve as evidence for the construct validity. Following these, factor-based discrimination procedures commenced. Item discrimination procedure involves the scores gained by the comparison of the scores of those from upper and lower quarters (27%) via independent sample t-test. The main point of this procedure is to display whether a response given to a specific item has changed between upper and lower groups, thusly, indicating the power of discrimination (Büyüköztürk, 2012). Thus, in this context, an independent samples t-test was used to determine whether there was a statistically significant difference between arithmetic means of upper and lower 27% groups, and the results of item-total scores were screened as shown in [Table 3](#).

**Table 3.** Total Score Lower-Upper 27% Findings

Group	n	$\bar{X}$	SD	df	t	p<	$\eta^2$
Lower 27%	61	39.2	4.99967		-		
		623		120	13.8	.05	.61*
Upper 27%	61	49.0	2.34497		19		
		328					

\*Large effect size (Büyüköztürk, 2016)

Upon implementing the independent samples t-test to determine statistical significance between lower and upper 27% groups separately designated for the discrimination of scale total scores, differences among all groups indicated statistical significance ( $p<.05$ ). It is suggested that impact scale be taken into consideration to underline the power of statistical significance (Akbulut, 2010). Regarding the scale of impact in terms of total scores, a wide impact scale is observed (Büyüköztürk, 2016).

Subsequently, to detect the discrimination impact of scale items, an item-based upper-lower 27% analysis was implemented. Results of the item analysis were shown in [Table 4](#).

**Table 4.** Item-Based Lower-Upper 27% Findings

Factor	Item	t	p<	$\eta^2$
Factor 1	Item 1	-2.665	.01	.056*
	Item 2	-2.393	.05	.046*
	Item 3	-2.916	.01	.066**
	Item 4	-2.693	.01	.057*
Factor 2	Item 9	-6.556	.001	.264***
	Item 12	-5.050	.001	.175***
	Item 14	-3.873	.001	.111**
	Item 15	-4.609	.001	.150***
	Item 16	-4.189	.001	.128**
Factor 3	Item 7	-2.702	.01	.057*
	Item 10	-5.329	.001	.191***
	Item 20	-3.260	.01	.081**
	Item 21	-2.835	.01	.063**
	Item 22	-3.793	.001	.107**
Factor 4	Item 28	-3.332	.01	.085**
	Item 29	-4.088	.001	.122**
	Item 30	-4.132	.001	.125**
	Item 31	-4.106	.001	.123**
	Item 32	-3.827	.001	.109**

\* Small effect size; \*\* Medium effect size; \*\*\*Large effect size (Büyüköztürk, 2016)

Following the independent samples t-test run to unearth statistical significance between lower and upper 27% groups that were individually assigned for the discrimination of all scale items, all group differences were found to be statistically significant ( $p<.05$ ). Considering the item-based analyses, apart from the significant difference between all items' upper and lower scores, a small impact for 4 items, a moderate impact for 11 items, and a wide impact for the remaining 4 items were discovered (Büyüköztürk, 2016).

### 3.2. Confirmatory Factor Analysis

The aim of the confirmatory factor analysis is to test the model generated as a result of the exploratory factor analysis (Seçer, 2015). Originally, in scale development studies, a confirmatory factor analysis of the model obtained after the exploratory factor analysis is necessary as an additional technique. In the analysis of the data obtained in the research, LISREL 8.7 software was used to test the model. Standardized results concerning the model are presented in Figure 2.

In circumstances where the objective is to increase the fit indices, some modifications can be made to serve this purpose (Seçer, 2015). These modifications need to be in accord with the theoretical framework (Sümer, 2000). In this sense, some modifications were made in Item 12, Item 19, Item 20, and Item 22 in accordance with the theoretical basis. Modifications for the items that were theoretically significant and were accumulated under the same factor were employed. During the implementation of these modifications, it is recommended that the decline in chi-square values be taken into consideration (Şencan, 2000). Therefore, the all modifications were performed with these delicate considerations.

To prove model appropriateness, fit values need to be examined (Seçer, 2015; Büyüköztürk, 2016). In the research, Chi-Square Goodness index, Goodness of Fit Index (GFI), Adjusted



Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), Normed Fit Index (NFI), Non-normed Fit Index (NNFI), Parsimony Goodness of Fit Index (PGFI), Root Mean Square Error of Approximation (RMSEA), Root Mean Square Residual (RMR), Incremental Fit Index (IFI), Relative Fit Indices (RFI), and Standardized Root Mean Square Residuals (SRMR) values were transferred. Fit values are presented in Table 5.

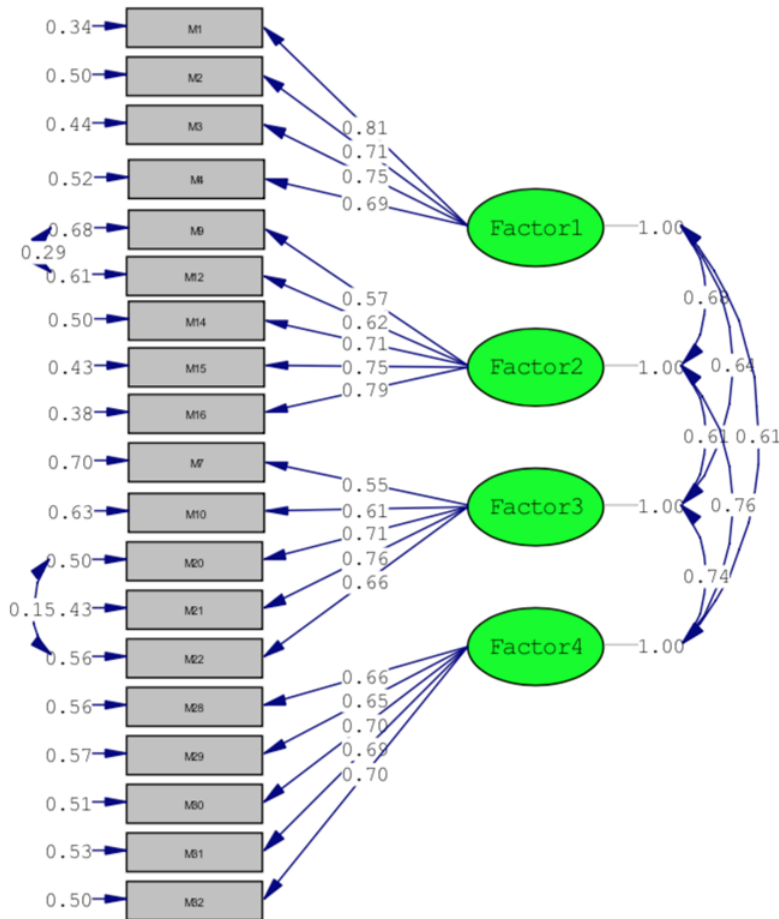


Figure 2. CFA with Standardized Results

Table 5. Fit values of the scale

Fit Index	Obtained Value	Reference Vlues	Source
Chi-Square	199.97	N/A	N/A
df	144	N/A	N/A
Chi-Square/df	1.39	≤ 2	(Tabachnick and Fidell, 2007)
GFI	0.92	≥0.90	(Sümer, 2000)
AGFI	0.89	≥0.90	(Sümer, 2000)
NFI	0.96	≥0.95	(Sümer, 2000)
NNFI	0.99	≥0.95	(Sümer, 2000)
CFI	0.99	≥0.95	(Sümer, 2000)
PGFI	0.70	≥0.50	(Mulaik et al., 1989)
RFI	0.96	≥0.90	(Marsh and Hau, 1996)
IFI	0.99	≥0.90	(Marsh and Hau, 1996)
RMSEA	0.041	≤ 0.05	(Sümer, 2000)
RMR	0.037	≤ 0.05	(Brown, 2006)
SRMR	0.046	≤ 0.05	(Brown, 2006)

Upon consideration of the table, it is explicit that fit indices for the CFA model overlaps with cutting points/reference values provided by the literature. Although AGFI value is barely above the limit, considering that fit indices should be dealt with as a whole (Jöreskog & Sörbom, 1993), it can be stated that the model showed an appropriate fit.

#### 4. DISCUSSION, CONCLUSION and RECOMMENDATIONS

As the product of the research, a perception scale with 4 sub-dimensions (spelling of acronyms, conjunctions, and suffixes; spelling of capital letters; spelling of compound words, spelling of words that went through word formation processes) and 20 items to be used to determine students' spelling mistakes was developed. Initially, the pool of 36 items was reduced to 34 items based on the feedback (content validity rate) from the scholars in the field to produce the pilot form. Subsequently, the research carried on with two participant groups, and the first group including 232 primary education and Turkish language teachers were administered the 34-item scale. The factor analysis revealing 4 sub-dimensions led to a re-run of the factor analysis which generated 19 items whose factor loading values were above 0.40. The final version of the scale which included 19 items and 4 sub-categories was observed to be explaining the 62,6% of the total variance, therefore, qualifying as ideal (Scherer, 1988).

On another note, exploratory and confirmatory factor analyses of the scale were run for the construct validity of the scale. In the sense of EFA, two outcomes were reached. First, KMO test result (0.896) revealed that the scale was sufficient for factor analysis in terms of sample size (Pallant, 2011). Second, the data was ensured to be appropriate based on both carrying the prerequisite of 10 observations per variable (Şencan, 2005) and Barlett Sphericity test result (1976.141;  $p < .05$ ) in the sense of normal distribution (Tavşancıl, 2018; Brace, Kemp & Snelgar, 2006). In addition, even though inter-item correlation values above .30 are recommended for the majority, the research included some correlations below .10. Upon the examination of the related literature, it can be deduced that sufficiency for the correlation matrix was obtained due to Barlett Sphericity test being statistically significant ( $p < .05$ ) (Ho, 2006). On the other hand, item factor loading values generated appropriate results (Ferguson & Takane, 1989), and the explained variance was acceptable for the field of social sciences (Scherer, 1988). Moreover, according to the scree plot graphic, factorization ended right after the breaking point (Büyüköztürk, 2016) and the eigen values for each factor were computed above 1 (Büyüköztürk, 2016) in accordance with Kaiser criterion. Thusly, it is possible to state that the criteria suggested by the literature concerning the EFA were met. Additionally, it is explicit that correlations between the scale's sub-factors were linear to one another and they indicated statistically significant relations. Correlation values can be deemed as appropriate.

It can be confirmed that the findings reached after the confirmation of the construct obtained as a consequence of EFA with CFA were aligned with the reference values and that majority of them were above acceptable values (Mulaik et al., 1989; Marsh & Hau, 1996; Sümer, 2000; Brown, 2006; Tabachnick & Fidell, 2007). Although the AGFI value was below the acceptable level, it can be evidenced that the value was around the threshold. It is suggested that fit indices should rather be assessed as a whole, not individually (Jöreskog & Sörbom, 1993). In this perspective, it can be deduced that fit indices in the study, when examined as a whole, meet the criteria dictated by the literature.

Cornbach Alpha reliability coefficient of the scale was calculated as .91. Four sub-dimensions of the scale, on the other hand, were computed as .83, .83, .82, and .82 respectively. As a result of this analysis, consistency of the items in the whole scale and the items in the corresponding factors was tested. The obtained coefficient can be between 0 and 1, and the closer it gets to 1, the more reliable it becomes as a value (Ural & Kılıç, 2006). According to these findings, it can be validated that the scale is reliable in terms of both the whole scale and its sub-dimension

levels (Özdamar, 2016). Regarding the lower-upper 27% analyses, it was observed that all items had statistically significant differences. In the sense of both item and total scores, mostly medium and large effect sizes were traced (Büyüköztürk, 2016). In addition, considering the items' total correlations; it was observed that all of them contributed to the scale with correlation values below .40. In this context, it is suggested that correlations above .30 can be accepted as functioning values (Büyüköztürk, 2016). Therefore, it is plausible to state that the scale meets the criteria observed by the literature in terms of reliability and validity.

It is observed in the literature that scale development studies on written expression are mainly conducted on attitudes, anxiety, belief, and self-efficacy subjects. No scale development research towards spelling mistakes is present. Perceptions of teachers concerning various teaching circumstances can make great contributions in terms of functionality and quality of the education. Besides, identifying teacher perceptions regarding wrongdoings of knowledge and practical skill towards spelling rules which make up a great deal of the writing process can elevate the quality of development of writing skill. In this sense, this scale, which could be used to solve problems related to categorization of the frequency of spelling mistakes and reduction of students' spelling mistakes during the writing skill acquisition process, can be administered to ensure that teachers categorize spelling mistakes and related findings in accordance with the four sub-dimensions to channel their activities towards these mistake types. Based on the discussion provided on the matters, the following recommendations can be presented:

1. This scale can be administered to teachers working in different levels of educational contexts, and situation reports for spelling rules in accordance with various steps can be generated.
2. This study can be upgraded to be administered in higher education levels after an adaptation process.
3. A similar version of this study can be developed in a different field to detect the situation in punctuation marks.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### ORCID

Ali Türkel  <https://orcid.org/0000-0003-4743-8766>

### 5. REFERENCES

- Akbulut, Y. (2010). *SPSS applications in social sciences: frequently used statistical analyses and SPSS solutions with explanations*. İstanbul: İdeal Kültür Pub.
- Akgül, A., & Çevik, Ç. (2003). *İstatistiksel analiz teknikleri [Statistical analysis techniques]*. İstanbul: Emek Pub.
- Aksoy, Ö. A. (1985). *Yine dil yanlışları 270 sözün eleştirisi [Again language mistakes 270 criticism of the word]*. Ankara: Öğretmen Pub.
- Aksoy, Ö. A. (1990). *Dil yanlışları [Language mistakes]*. Ankara: Yalçın Emel Pub.
- Aktaş, Ş., & Gündüz, O. (2003). *Yazılı ve sözlü anlatım kompozisyon sanatı [Written and oral expression composition art]*. Ankara: Akçağ Pub.
- Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. London: Kluwer Academic Pub.
- Bağcı, H. (2011). Elementary 8th grade student's level of ability to apply spelling rules and punctuation marks. *Turkish Studies*, 6(1), 693-706.

- Bayat, N. (2013). Choice of vocabulary and syntax mistakes of novice teachers in writing. *Electronic Journal of Social Sciences*, 12(43), 116-144.
- Bayram, N. (2004). SPSS data analysis in social sciences, Bursa: Ezgi Pub.
- Brace, N., Kemp, R., & Snelgar, R. (2006). *SPSS for psychologists*. New York: Palgrave Macmillan Press.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.
- Büyüköztürk, Ş. (2016). *Handbook of data analysis for social sciences*. Ankara: Pegem Academy Pub.
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E, Karadeniz Ş., Demirel, F. (2016). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Ankara: Pegem Academy Pub.
- Cohen, R. J., & Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement*. Boston: McGraw-Hill.
- Cho, Y. (2003). Assessing writing: are we bound by only one method. *Assessing Writing*, 8(3), 165-191. [Doi:10.1016/S1075-2935\(03\)00018-7](https://doi.org/10.1016/S1075-2935(03)00018-7)
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2018). *Multiple variable statistics, SPSS and LISREL applications for social sciences*. Ankara: Pegem Academy Pub.
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365-387.
- Ferguson, F., & Takane, Y. (1989). *Statistical analysis in psychology and education*. McGraw Hill Book Company.
- Field, A. (2009). *Discovering statistics using SPSS*. London: Sage Publications.
- Gebrhard, R.C. (1983). Writing processes, revision and rhetorical problems: a note on three recent articles. *College Composition and Communication*, 34(3), 294-296.
- Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. London & New York: Chapman & Hall/CRC.
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the simplis command language*. Lincolnwood: Scientific Software International, Inc.
- Kansas, S. (2019). *Fransızca yazım kılavuzu [French spelling guide]*. Paris: Hatier Pub.
- Karagül, S. (2010). *Application level of the noted writing and its rules in Turkish lesson teaching programme of the elementary 6-8. Class students*. Unpublished MA Dissertation, Dokuz Eylül University, İzmir.
- Kellogg, R.T. (1996). A model working memory in writing. In C.M. Levy & S. Ransdell (Eds.). *The science of writing: theories, methods, individual differences and applications* (p.57-72). Mahwah, NJ: L. Erlbaum Associates.
- Kıbrıs, İ. (2010). *Türkçe 1: yazılı anlatım. [Turkish 1: written expression]*. Ankara: Kök Pub.
- Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2006). *Methods in educational research: from theory to practice*. San Francisco: Jossey-Bass.
- Marsh, H. W., & Hau, K. T. (1996). Assessing goodness of fit: is parsimony always desirable. *The Journal of Experimental Education*, 64(4), 364-390.
- MEB (2006). *İlköğretim Türkçe Dersi Öğretim Programı (6-8. Sınıflar) [Primary Education Turkish Course Curriculum (Grades 6-8)]*. Ankara: MEB Pub.
- MEB (2009). *İlköğretim Türkçe Dersi Öğretim Programı (1-5. Sınıflar) [Primary Education Turkish Course Curriculum (Grades 1-5)]*. Ankara: MEB Pub.
- MEB (2018). *İlköğretim Türkçe Dersi Öğretim Programı (6-8. Sınıflar) [Primary Education Turkish Course Curriculum (Grades 1-5)]*. Ankara: MEB Pub.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C. D. (1989). An evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin*, 105(3), 430-445.

- Özdamar, K. (2016). *Structural equation modelling for scale and test development*. Eskişehir: Nisan Pub.
- Pallant, J. (2011). *SPSS survival manual: a step by step guide to data analysis using SPSS*. Crow's Nest: Allen & Unwin Pub.
- Seçer, İ. (2015). *Process of psychological test development and adaptation: SPSS and LISREL applications*. Ankara: Anı Publishing.
- Şencan, H. (2005). *Validity and reliability in social and behavioral measurements*. Ankara: Seçkin Publishing.
- Sümer, N. (2000). Structural equation models: basic concepts and sample applications. *Turkish Psychology Articles*, 3(6), 49-74.
- Ural, A., & Kılıç, İ. (2006). *Bilimsel araştırma süreci ve SPSS ile veri analizi [Scientific research process and data analysis with SPSS]*. Ankara: Detay Pub.
- Ünver, İ. (2008). Suggestion on standart writing in transcription. *Turkish Studies*, 3(6),1-46.
- Tavşancıl, E. (2018). *Measuring the attitudes and data analysis via SPSS*. Ankara: Nobel Academy Pub.
- TDK Türkçe Sözlük (2019). *Turkish Dictionary of Turkish Language Institution*. [http://www.tdk.gov.tr/index.php?option=com\\_gts&view=gts](http://www.tdk.gov.tr/index.php?option=com_gts&view=gts) (Access: 05.03.2019).
- Türkel A., Yaman B., Aksu C. (2017). *Perceptions of classroom and Turkish teachers' towards spelling rules and students' writing mistakes*. 10.UTEOK, Okan Üniversitesi, İstanbul.
- Veneziano L., & Hooper J. (1997). A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior*, 21(1), 67-70.

## 6. APPENDIX

**Table A1.** The Scale of Perceptions of Teachers' on Spelling Mistakes (Original [Turkish] version).

Please indicate how often your students perform the actions and situations given below by marking the relevant figure as (X).

Maddeler	Her zaman	Bazen	Sıklıkla	Nadiren	Hiçbir zaman
1.Öğrencilerim, özel adların ilk harfini küçük yazarak yanlışlar yapar (atatürk, ayşe, Paris caddesi, türkçe vb.).	5	4	3	2	1
2.Öğrencilerim, cins adların ilk harfini büyük yazarak yanlışlar yapar (Kuş, Araba, Oyuncak vb.).	5	4	3	2	1
3.Öğrencilerim, tümce başlarındaki ilk sözcüğe küçük harfle başlayarak yanlışlar yapar (kedim arkamdan geldi vb.).	5	4	3	2	1
4.Öğrencilerim sözcük ortasında büyük harf kullanarak yanlışlar yapar (seFer, aDa vb.).	5	4	3	2	1
5.Öğrencilerim, yazılarında ağız özelliklerini kullanarak yanlışlar yapar (gelirem, meğersem, halbüsem, gidek vb.).	5	4	3	2	1
6.Öğrencilerim, bağlaç olan “de” ve ek olan “-DE”nin yazımında yanlışlar yapar (Ben de para yok; bende gelmek istiyorum, Ahmet de bizimle gelecek vb.).	5	4	3	2	1
7.Öğrencilerim, ünlülerle ilgili ses olaylarının gerçekleştiği (ünlü daralması, ünlü düşmesi, ünlü türemesi)ne uğrayan sözcüklerde yanlışlar yapar (bekliyor, ağızının, azcık, sapsağlam vb.).	5	4	3	2	1
8.Öğrencilerim, bağlaç olan “ki” yazımında yanlışlar yapar (Çalışki başarasın; eminimki, Senki en yakın arkadaşısın...).	5	4	3	2	1
9.Öğrencilerim, bitişik yazılması gereken ad ve sıfatları ayrı yazarak yanlışlar yapar (vatan sever, uyur gezer, dedi kodu, bir kaç, çok bilmiş vb.).	5	4	3	2	1
10.Öğrencilerim, bitişik yazılması gereken birleşik eylemleri ayrı yazarak yanlışlar yapar (gele bilirim, red ettim kaçtı verdi, şaşta kaldı vb.).	5	4	3	2	1
11.Öğrencilerim, ayrı yazılması gereken birleşik eylemleri bitişik yazarak yanlışlar yapar (farketti, terketti vb.).	5	4	3	2	1
12.Öğrencilerim, sözcükteki veya sözcüğün ekindeki “ğ” sesini yazmayarak yanlışlar yapar (fotoraf, öğretmen, baktında, oldunu, adamcaz vb.).	5	4	3	2	1
13.Öğrencilerim, bazı sözcüklere fazladan ünlü veya ünsüz harfi ekleyerek yanlışlar yapar (bağazi, hayyal, messela vb.).	5	4	3	2	1
14.Öğrencilerim, bazı sözcüklerdeki harfleri düşürerek yanlışlar yapar (galba, heralde, kavaltı vb.).	5	4	3	2	1
15.Öğrencilerim, ölçü birimlerinin yazımında yanlışlar yapar (sm, lt, vb.).	5	4	3	2	1
16.Öğrencilerim, yer-yön adlarının yazımında yanlışlar yapar (içerde, dışarda, burda, ilerde vb.).	5	4	3	2	1
17.Öğrencilerim, temel kısaltmaları küçük yazarak ve/veya harflerin aralarına nokta koyarak yazımında yanlışlar yapar (tbmm, T.D.K vb.).	5	4	3	2	1
18.Öğrencilerim, coğrafi adların yazımında ilk addan sonra gelen adlara küçük harfle başlayarak yanlışlar yapar (Asya yakası, İstanbul boğazı vb.).	5	4	3	2	1
19.Öğrencilerim, özel adlara ek getirilmesi ile ilgili yazım yanlışları yapar (Konağ’a, Burağ’a, Serab’a vb.).	5	4	3	2	1