IJATE

Volume 5 Issue 4 2018



# International Journal of Assessment Tools in Education

International Journal of
Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

e-ISSN 2148-7456

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

Volume 5 Issue 4 2018

**Dr. Izzet KARA**
Editor in Chief
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

Publisher : İzzet KARA
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.gov.tr/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

# International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed online journal. IJATE accepts original theoretical and empirical English-language manuscripts in psycho-educational assessment. Theoretical articles addressing new developments in measurement and innovative applications are welcome. IJATE publishes articles appropriate for audience of educational measurement specialists and practitioners.

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)

• TR Index (ULAKBIM),

• DOAJ,

• Google Scholar,

• Index Copernicus International

• Index of Turkish Education,

• Open Access Journals,

• Akademik Dizin,

• Academic Keys,

• CiteFactor (ASJ),

• SIS (Scientific Index Service) Database,

• SCIPIO (Scientific Publishing & Information Online),

• MIAR 2015 (Information Matrix for Analysis of the Journals),

• I2OR Indexing Services,

• JournalTOCs

• Sosyal Bilimler Atıf Dizini (SOBIAD)

• International Innovative Journal Impact Factor (IIJIF)

## Table of Contents

***Research Article***

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 593–610

DOI: 10.21449/ijate.390603

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# A Model Proposal for Solving Problems Encountered in Performance Tasks

**Melek Gülşah Şahin** [1], **Nagihan Boztunç Öztürk** [2,*]

[1] Gazi University, Gazi Education Faculty, Department of Measurement and Evaluation in Education, Turkey
[2] Hacettepe University, Lifelong Learning Center, Turkey

**ARTICLE HISTORY**

*Received: 28 March 2018*
*Revised: 29 May 2018*
*Accepted: 04 August 2018*

**KEYWORDS**

Performance tasks,
Measurement and evaluation,
Classroom assessment,
High school education

**Abstract:** This study aims to establish a model proposal for solving problems in the use of performance tasks at the high school level. The study which adopted qualitative research method was planned in two stages. First, the problems faced by high school teachers ($n$ = 40) using performance tasks were determined, and for the second stage the opinions of academicians ($n$ = 8) related to the solution of the determined problems were consulted. Three stakeholders were identified in solving the problems identified during the first step of the study; those being the university, the Ministry of National Education (MoNE), and the teachers. Although the perceptions of the teachers' use of performance tasks in general were positive, it was seen that there problems were experienced in their application. Emphasis was placed on the importance of communication between the university, the MoNE and the teachers in order to overcome these problems. For the effective function of performance tasks, points such as diversity in classroom practices, attention to the characteristics of students and environmental conditions, and the use of rubrics were emphasised.

## 1. INTRODUCTION

The concept of skill seems to have already taken over the 21st century. In order to be considered ready for the 21st century, humanity has been conducting research, changing educational systems, introducing new job descriptions, and seeking out workers with 21st century skills (critical thinking, problem solving, creativity, innovation, collaboration, etc.). In the global sense, it has passed from a time where knowledge was the first thing that came to mind when success was mentioned, to a time when skills are the primary consideration. The concept of skill is now a global concept, yet the problems of actually training future employees for unspecified occupations remains a generic task for the future.

Skillsets have become the worldwide employment passports of the 21st century. If proper investment is not undertaken in skills, people suffer on the edges of society, technological

CONTACT: Nagihan BOZTUNÇ ÖZTÜRK ✉ nagihanboztunc@hacettepe.edu.tr Hacettepe University, Lifelong Learning Center, Ankara, Turkey

**ISSN-e: 2148-7456 /© IJATE 2018**

development does not turn into economic advancement, and international competitiveness may not be achieved in a global society built on knowledge. The nature of employment has altered considerably over the past decade, and the demand for ability has been increasing in OECD countries and beyond. Job shifts have been observed from the agricultural lands to the factories and professional workplaces in OECD countries from the 1960s through till recent measures were taken in 2009 (OECD, 2012). Information evaluation and analysis skills, and using these skills practically in the real world, have become of significant importance for students because they are the required specifications for the professions and for the leaders of higher education (O'Sullivan & Dallas, 2010). The need to have students with these skills has led to the need for educational training programmes to be updated. There is a worldwide common interest to change the educational vision by advocating views that are referred to as 'deeper learning' and '21st century skills' (Pellegrino, 2017). Educational researchers have been discussing how 21st century skills such as critical thinking, problem solving, collaborative working, and communication can be taught and how it can be evaluated.

In terms of its evaluation, when 21st century skills are considered, it is emphasised that formative and summative assessments should be planned in order to complement each other (OECD, 2017; P21Skills, 2007). It is important to establish an evaluation balance between high quality standardised tests and effective formative assessment. Formative assessment of 21st century skills in an effective way should carry a number of characteristics. Firstly, it should concentrate on 21st century skills and ideas, then it should make the thinking visible by exposing the conceptual strategy types that a student exploits in their problem solving. In addition, it must be structured in a way that the teachers can determine the background knowledge used by students to solve problems in the real world. It also should be genuine and based on performance; demanding that students apply their 21st century skills. It should derive data that can be used to directly inform teaching practices, and it should have a goal to form capacity – for both teachers and students. In addition, it must form part of a comprehensive assessment continuum and display a sense of learning as complex, integrated, and as shown in actual performance (P21Skills, 2007). The performance tasks, which are among the formative assessment tools, are activities aimed at developing and measuring students' high-level cognitive skills, and presenting problem situations that students may encounter in real life (Kutlu, Doğan, & Karakaya, 2009). Looking at the characteristics of formative assessment of 21st century skills, it is clearly seen that performance tasks should be employed; and McTighe (2016) stated that two or more 21st century skills can be involved in performance tasks.

In Turkey, performance tasks have been implemented in primary school grades 1-5 from the 2005-2006 academic year and implemented gradually in other grades from 2006-2007 (Gelbal & Kelecioğlu, 2007). Performance tasks were added to high school curricula under an education regulation dated 07.09.2013 (Milli Eğitim Bakanlığı [Ministry of National Education], 2013).

Problems have been encountered over time in the accomplishment of performance task evaluation such as tasks not having the ability to measure high-level skills, the ability to evaluate students in the process, lack of encouragement of using skills such as responsibility and communication skills and so forth. Benefits provided by performance tasks have been replaced by problems such as tasks being performed more by the families of students, that students copy information directly from the Internet, that evaluation takes time, and that students only perceive it as another grade to be achieved. In 2008, the MoNE introduced some

measures regarding the more effective use of performance tasks in educational processes. However, since these measures did not produce the desired results, the performance tasks were abolished in elementary schools under a regulation change in 2014 (Türk Eğitim Derneği [TED], 2014).

When the academic studies related to performance tasks are examined; it is seen that although the perceptions about performance tasks are generally positive, some problems in classroom practices (insufficiency in lesson hours, parental support, resources, appropriate performance tasks for students' level, overcrowded classrooms, shortage in objective assessment, time taken for evaluation etc.) have been encountered (Ak, 2013; Arı, 2010; Arslan, 2013; Aydin, Yörek, & Uğulu, 2015; Bal, 2012; Benli, 2010; Dilekmen & Aydoğdu, 2015, Göçgen, 2016; Metin, 2013; Palaz, Kılcan, Akbaba, & Çepni, 2015; Sarı, 2014; Seyhan, 2017; Tanrıverdi, 2012). However, there are also studies to be found in the literature showing that when performance tasks are used effectively, students are positively influenced in terms of their attitudes towards the classroom, creativity, learning, tendency to think critically, their development of collaborative skills, self-esteem, and their problem-solving skills (Aslan Yolcu, 2013; Bayrakçı, 2010; Ekici, 2016; Furat, 2009; Gürel, 2012; Kutlu et al., 2009; Öztürk, 2010; Sözer, 2015).

The problems encountered were seemingly not down to the features of the performance tasks, but mainly due to the assessment tool not being applied correctly (Türk Eğitim Derneği [TED], 2014). The removal from the curricula of the performance tasks, which have an important place in the evaluation of 21st century skills, and the lack of any substitute method creates a deficiency in the Turkish educational system. For this reason, it is important to present solutions for the effective use of performance tasks. When the published literature is examined, it can be seen that the majority of studies have been carried out at the primary education level. Performance tasks have not yet been removed from high school education institutions; this being the case, determining problems related to performance tasks and presenting solution proposals are important in terms of informing the appropriate decision-making units.

The current study aims to present a model proposal to address problems encountered in the implementation of performance tasks at the high school level. For this purpose, answers are sought to the following research questions.

1. What problems do high school teachers encounter with regard to performance tasks?
2. What kind of a model can be proposed to increase the reliability, validity, and usefulness of performance tasks?

## 2. METHOD

A qualitative research methodology has been adopted in this study. The qualitative approach is a method that can reveal the opinions and perspectives of individuals about the subject being investigated, and to reveal sub-factors forming these aspects (Yıldırım & Şimşek, 2008). The study has been conducted in three stages. In the first stage of the research, high school teachers working in Ankara, Turkey, were asked to identify problems experienced with performance tasks. In the second stage, in line with the teachers' opinions obtained in the first stage, faculty members were requested to propose solutions. In the third stage, a model will be proposed in light of the outcome of the first two stages of the study, and from examining the published literature.

## 2.1. The Participants

The participant group of the first stage of the research constitutes 40 teachers working in state high schools in Ankara, Turkey. In the study, maximum diversity sampling, one of the purposeful sampling methods was selected. The frequency distribution according to the demographic characteristics of the participant teachers is presented in Table 1.

**Table 1.** The demographic characteristics of the teachers

| Variables | Category | f | % | Variables | Category | f | % |
|---|---|---|---|---|---|---|---|
| Gender | Female | 25 | 62.5 | Major | Mathematics | 7 | 17.5 |
| | Male | 15 | 37.5 | | Turkish Philology | 6 | 15.0 |
| Age | 24-30 | 8 | 20.0 | | English | 4 | 10.0 |
| | 31-35 | 8 | 20.0 | | Physics | 5 | 12.5 |
| | 36-40 | 9 | 22.5 | | Chemistry | 4 | 10.0 |
| | 41-45 | 5 | 12.5 | | Biology | 3 | 7.5 |
| | 46-50 | 5 | 12.5 | | History | 3 | 7.5 |
| | 51-55 | 5 | 12.5 | | Philosophy | 2 | 5.0 |
| In-Service Training on Measurement and Evaluation | Yes | 14 | 35.0 | | Geography | 2 | 5.0 |
| | No | 26 | 65.0 | | Others* | 4 | 10.0 |
| Professional Seniority (year) | 0-5 | 11 | 27.5 | Average Number of Students | 20 and below | 2 | 5.0 |
| | 6-10 | 4 | 10.0 | | 21-30 | 18 | 45.0 |
| | 11-15 | 6 | 15.0 | | 31 and over | 20 | 50.0 |
| | 16-20 | 6 | 15.0 | | | | |
| | 21 and over | 12 | 30.0 | | | | |

***Other majors are music, physical education and French language

Table 1 shows that 25 (62.5%) of the teachers are female and 15 (37.5%) are male. According to the age variable, nine (22.5%) of the teachers are in the 36-40 age bracket. As to their education, 52.5% of the teachers graduated from an education faculty. When the distribution according to the branch variable is examined, seven (18%) of the teachers are from the mathematics department. In terms of their seniority, 11 of the teachers (27.5%) have 0-5 years' experience. The distribution of the average number of students in the teachers' classes shows an average of 31-40 students in the classes of the 20 (50%) of the teachers. Also, 26 (65%) of the teachers stated that they did not receive any in-service teacher training about measurement and evaluation.

In the second stage of the research, eight faculty members from an educational faculty at a university were interviewed. The reason for the selection of experts from four basic areas in educational sciences was that when the limitations and the literature (Arı, 2010; Aydın et al., 2015; Benli, 2010; Çiftçi, 2010; Metin, 2013; Sarı, 2014) were examined, it was thought that interdisciplinary solution proposals would be more appropriate and meaningful. Before being selected for interview, it was checked that each of the experts had studied the subject, that they followed up-to-date information, and that their solution proposals would be about specified limitations from their own fields. Information about the coding of the experts in the study is presented in Table 2.

**Table 2.** The information of the coding about the experts

| Code | Department | Professional Seniority (year) |
|---|---|---|
| E1-E2-E3 | Educational Measurement and Evaluation | 16-17-10 |
| E4-E5 | Curriculum and Instruction | 18-8 |
| E6-E7 | Educational Administration | 10-12 |
| E8 | Guidance and Psychological Counseling | 12 |

## 2.2. Data Collection and Analysis

The research was conducted in three stages. In the first stage of the research, the Teacher Opinion Form developed by the researchers was used to collect the data. Two measurement and evaluation experts reviewed the form, and it was finalised after the necessary corrections were applied. The form consists of two parts. Demographic information was sought in the first part, whilst in the second part, open-ended questions about opinions as to the problems experienced with regard to performance tasks were asked. The form was personally handed to the participant teachers by the researchers and then the questionnaire administered. Researchers collected the forms by hand, after waiting for the teachers to complete the form. Content analysis was applied to the data obtained from the form. Content analysis makes it possible to examine verbal, written and other materials objectively and systematically (Tavşancıl & Aslan, 2001). According to Yıldırım and Şimşek (2008), the main goal of content analysis is to reach the concepts and associations that can explain and thereby describe the data. For this purpose, first of all, conceptualisation of the data and then necessary logical arrangements must be made to create themes. The NVivo analysis software program was used in the analysis of the data.

In the first stage of the study the researchers transferred the data to an electronic media in order to establish the validity of the study. The data were examined and discussed by the researchers and then categorised. The researchers consulted with another expert in the field of measurement and evaluation in case of any problems in the data coding and naming, thus a consensus was achieved. In addition, the data based on the formation of each category are given clearly in the Findings section. After the categories were identified, two measurement and evaluation experts' opinions were asked about the convenience of the categories. Then the feedback regarding the naming of the categories and data classification were evaluated one by one by the researchers working together, and their final versions were decided. Moreover, in order to assure the validity of the study, direct citation (Yıldırım & Şimşek, 2008) has been included, with the researchers' own cues included as quotations.

In the second stage of the study, the aim was to take the views of academic members from a university faculty about the problems determined based on the opinions of the teachers and to elicit their suggestions for practical solutions to the reported problems. For this purpose, a semi-structured Expert Interview Form was created by the researchers. The expert interview form consists of two parts. Questions relating to demographic information about the experts formed the first part of the form. In the second part, codes obtained from the first stage of the study were used in accordance with field expert guidance to form questions relating to solution proposals to the problems reported, and to their solutions' limitations. At this stage, researchers should note that the study should be considered as being related to performance tasks. Prior to receiving expert opinion, the expert interview form was emailed to the experts and they were informed about the purpose, scope and the content of the study. The opinions of the experts were taken in face-to-face meetings by the researchers that lasted for 25-30 minutes. The experts' opinions were first transcribed then returned for their approval; after which, they were included in the study. Content analysis was also used in the analysis of the data obtained at this stage. The same data analytical path of the first stage was also followed for this second stage.

In the third phase of the study, a model was proposed by the researchers based on the previous two stages and from the literature. Opinions of two measurement and evaluation experts were gathered in order to assess the clarity of the proposed model and its compatibility with the information obtained from the other two stages. The model was finalised to include any proposed amendments.

Yıldırım (2010) mentioned a number of measures that could be applied by the researchers so as to improve the quality of the qualitative research. For this purpose, detailed explanations and

descriptions are made and the research carried out within certain limits. In addition, co-audits were carried out both in the creation of the themes and coding; and also, when considered necessary, external expert opinion was sought. Additionally, the data collection and analysis process is clearly presented in the study. Direct quotations of the teachers and from the expert opinions are included as direct evidence of the reliability of the data obtained in the study.

## 3. FINDINGS

In this part of the study, findings related to each sub-problem were presented.

### 3.1. First Sub-Problem

In this sub-problem, the teachers' opinions about the problems experienced with performance tasks are stated. A total of seven categories were derived from the teachers' views, which are; Preparation-Evaluation Difficulty, Facility Limitation, Subject Limitation, Students' Perspective, Students' Access to Effortless Knowledge (Copying), Time Limitation, and Being Compulsory. Examples and interpretations for each category are presented as follows.

a. Preparation-Evaluation Difficulty (f = 8): Teachers stated that there was difficulty in preparing the performance tasks appropriately for the age of the students and purpose of the task. They also mentioned that the evaluation was time consuming and difficult, and touched upon its limitations. For example, while T5 referred to tasks not prepared according to student age by saying 'It was unable to be prepared properly for the age of the students'; T27 stated that it could not be prepared according to its purpose by saying, 'It is difficult to prepare an appropriate performance task for its purpose'. Besides, T26 stated that 'evaluating is subjective', whilst T26 referred to the difficulty of evaluation as 'not seeing the point that the student does not understand'.

b. Facility Limitations (f = 5): As for the limitations of performance tasks, the teachers stated that classes were overcrowded, that there were no studying opportunities outside of the home and library, and that some students did not have the same opportunities. For example, T19 mentioned class overcrowdedness as, 'individual works cannot be accomplished, only group work can be done because of the overcrowded classes', and T40 referred to the facilities that some families provided as being limited by saying, 'some cannot have the same opportunities'.

c. Subject Limitation (f = 5): Teachers mentioned limited subject options for some performance tasks. For example, T13 stated that, 'We are obliged to give homework on certain topics'. T25 expressed that 'they are narrow-scoped' and T34 stated, 'It can be a problem when only limited to one subject'.

d. Students' Perspectives (f = 6): Teachers mentioned the limitations of performance tasks due to their students' perspectives, noting that students were reluctant to prepare performance tasks, that they did not take them seriously, that they just saw it as another grade and that they did not like preparing them. For example, T28 stated that students might have problems due to their grade-oriented thinking, expressing that 'students may not take it seriously when its grade weight is low'. While T21 mentioned that the creativity of students could be thwarted since it was regarded as just a grade point, indicating 'and the student sees it only as a grade raiser', and T39 emphasised that the students did not want to spend time with such work, saying 'it is perceived as a waste of time for the student'.

e. Students' Access to the Effortless Knowledge (Copying) (f = 4): Teachers indicated that pupils received parental help in preparing their performance tasks and copied from the Internet. T9 referred to the limitation of receiving family members' help, expressing that 'It is predominantly prepared by the parents when they are prepared at home', whilst T16 mentioned

that students made no effort, saying 'Students have to copy and they copy easily from the Internet due to the fact that similar tasks have been performed before'.

f. Time Limitation (f = 9): Teachers stated that limited classroom hours and performance task preparation time as a limitation. For example, T3 mentioned the time limit and that the tasks were not ideally prepared, stating 'There is a time limit effect as performance tasks cannot be prepared as desired'. While T8 pointed out imprecise preparation of homework with 'Cursory homework can occur when the time is limited for the student', T18 stated that students are unable to reveal their creativity, saying 'The limitations in terms of subject and time can prevent students' creativity and contribution'.

g. Being Compulsory (f = 3): Teachers stated that making performance tasks compulsory for each course was a limitation. They pointed out that the students' motivation decreased when it became compulsory. T25 said 'It is not expected that the student is willing, it can be considered as an obligation, not voluntary', while T21 expressed 'the efficiency decreases due to the fact that it becomes compulsory for all the branch teachers'.

### 3.2. Second Sub-Problem

In this second sub-problem, a model proposal was made based on the findings of the first sub-problem and on expert opinion. The conceptual figure of the proposed model is presented in Figure 1.

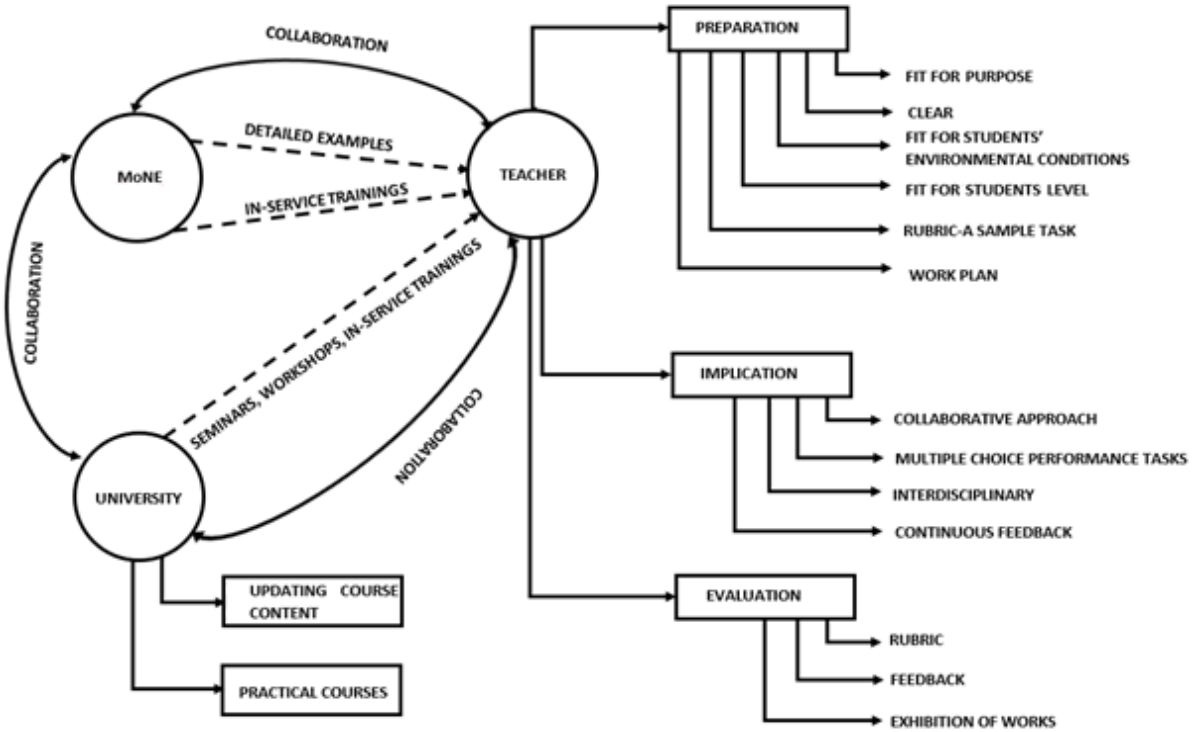



**Figure 1.** Conceptual figure of proposed model

As can be seen in Figure 1, the University, the MoNE, and Teachers are the three stakeholders involved in the model presented to overcome the limitations of performance tasks and projects. The relations between the relevant stakeholders, their duties, and the expert opinions used to form the model are detailed in the following subsections.

### *3.2.1. University*

In the content of the university stakeholder, there are proposals such as keeping the existing values up-to-date, proposing practical courses for the curriculum, cooperation with both the MoNE and teachers, and seminars, workshops and in-service training for teachers. The explanations for each proposal are presented as follows.

*a) Including practical courses in curriculum and updating course content:* In university faculties of education, it may be proposed that prospective teachers are offered lessons which require knowledge and skills expected of them in the future, and in which they have the opportunity to practice measurement and evaluation methods. In addition, teachers may be asked to apply measurement tools which they will practice as of the beginning of their teaching profession. Sample views are presented as follows.

***E1:*** *It seems that the prospective teachers have little time for the performance evaluation and rubrics in the scope of the measurement and evaluation course they have taken during their education. .... In this case, it would be better to devote more time to these issues and to carry out practice-directed studies. So the course has to be directed towards the field. For this, a course under the name of Measurement and Evaluation Practices can also be offered.*

***E2:*** *Although their performance tasks are very good, because of the teacher candidates' lack of knowledge and skills on teaching methods, they cannot be implemented. In addition, as an elective course at the undergraduate level, it is necessary to offer a "Measurement and Evaluation Practices in Schools" course.*

***E4:*** *The limited availability of opportunities that teachers mentioned is a limitation that prospective teachers can overcome with the Instructional Material Design lessons they have taken during their undergraduate education....Even, if all the universities and schools worked in cooperation, it would be possible to train teacher candidates better by updating the content of relevant course to meet the field needs.*

***E6:*** *In teacher training programmes, in cooperation with the HEC (Higher Education Council), it could be possible for the topics of performance and project tasks to take on an increased role in various courses of the curriculum.*

*b) Collaboration with the MoNE*: Universities may be recommended to cooperate with the MoNE particularly about the determination of needs, structuring in-service teacher training programmes and the preparation of programme contents. Sample views are presented as follows.

***E3:*** *It is important that universities work in coordination with the MoNE, especially in transferring information about the measurement of high-level skills and the resolution of the problems encountered.*

***E4:*** *If universities and schools work in cooperation, the relevant course content can be updated to meet the needs of the field in order that teacher candidates can be better educated.*

***E6:*** *Universities should cooperate with the MoNE in the presentation of solution proposals, and especially in determining teachers' needs. Particularly in the preparation of the programme contents, the views of relevant academic faculty members are important.*

*c) Providing Seminars, Workshops and In-service Trainings for Teachers:* It may be suggested for universities that faculty members who work in related fields to arrange seminars, workshops and in-service teacher training to address the practical knowledge deficiency of teachers and to resolve the problems they encounter. Some sample views are presented as follows.

***E5:*** *Teachers can be informed with different and correct examples with in-service teacher training programmes which can be offered by academic faculty members to teachers especially to determine and measure high-level mental skills.*

***E6:*** *Conferences, seminars, and workshops can be organised in cooperation with universities and schools to resolve the current problems of teachers in preparing and evaluating projects and performance tasks.*

### *3.2.2. The Ministry of National Education*

a) *In-service Teacher Training Program:* The Ministry of National Education could organise and enhance programmes related to the regulation of in-service teacher training. The following opinions were noted in connection with this.

***E2:*** *Teachers will benefit more if the materials provided to them during training are visual materials. The materials have to be interesting, visual and fit for the real purpose.*

***E4:*** *I think that teachers should be offered in-service teacher training if they feel inadequate about assessing performance tasks and projects; and that this training should definitely be practical in nature. For example, on the difficulty of evaluation, teachers could be trained on how to evaluate the relevant measurement tools through in-service training.*

***E6:*** *Firstly, the Ministry of National Education In-service Teacher Training department could increase in-service teacher training on performance tasks and the preparation and evaluation of projects.*

***E8:*** *They could also provide counselling services to teachers on the effective communication with their students. By this means, they could raise awareness in providing constructive feedback, rewarding students, and not judging or blaming students when they fail.*

*b) Presenting detailed examples to the teachers:* Teachers may be provided examples in the curriculum or separately in the material pool. Thus, the teachers will be informed about the appropriate measurement tools in the evaluation of high-level skills and about the measurement tools appropriate for the students' level. Example views are presented as follows.

***E2:*** *The teachers need to know what the performance task is. If there is a lack of basic information; even in the virtual environment, detailed informative examples about the performance task types could be shared with the teachers.*

***E5:*** *For students of different levels, detailed information could be included in the teacher's manual, and grading keys can be provided with the sample performance tasks and projects.*

***E6:*** *The MoNE Regulation of High School Education Institutions may include more descriptive information and examples of project and performance tasks, i.e. the stages of preparation and evaluation may be further elaborated.*

### *3.2.3. Teachers*

Suggestions were provided for teachers under the titles of Preparation, Implementation and Evaluation, which are detailed as follows.

#### *3.2.3.1. Preparation*

*a)* It should be clear: It may be suggested that teachers determine the purpose clearly in the preparation phase of both measurement tools. The purpose should not be directed to grading, but rather, if the intended purpose is to develop high-level mental skills of the students, the students will only demand help from outside when they are in difficulty. Sample views on this are presented as follows.

***E1:*** *If we look at the stages where the relevant measurement tools are implemented, the first thing that comes to mind is the need to provide clear goals. For this, students need to understand exactly what is expected of them and to direct their work accordingly. Besides, unnecessary performance tasks that are not within the scope of the determined purposes should be avoided....For both measurement instruments, the guidelines should be given as clear as possible. The rubrics must also be shared with the students in advance.*

***E4:*** *Families' and teachers' opinions also affect students seeing these tasks as simply grades. By emphasising their importance in real life, telling the students that these tasks are not just grades, but explaining what good it could do, would better motivate the students.*

***E7:*** *Since the goal is to improve the sense of responsibility in students and to make the student produce better, the current limited external support could serve much more for this purpose. When the responsibility is assigned to someone else (e.g., mother, father, sibling, friend), it is unlikely that the target behaviour will develop in the student. In order to avoid this, teachers should openly negotiate the situation with all relevant parties, notably the students. And they also should clearly state the expectations as a benefit for the students.*

***E8:*** *It is very important for students' identity development that they feel the sentiment "I did this with my own work and I presented a solution to this problem". The students in this developmental period should not be distracted by what they think is a waste of time. It has to be more realistic.*

*b) Fit for purpose and at students' level:* The students will be responsible for learning if it is suitable for the purpose of the task and aimed at their level. In addition, tasks that are not fit for purpose or at the students' level should be avoided, as it may demotivate them. Example opinions on this are as follows.

***E1:*** *Except for the intended purpose, unnecessary performance tasks should be avoided.*

***E2:*** *Teachers must first be given a limited performance task during their practical training. Skills acquisition for the relevant tasks should be provided; then an extended performance task must be provided. Tasks should be arranged proportionally and according to the class flow.*

***E3:*** *Tasks that are not in accordance with the level of the students' readiness will cause behaviours of seeking external help.*

***E4:*** *It is very important to know the level of students' readiness when performance tasks are given. The fact that a performance task is appropriate for the class level does not mean that it can be done. If the students' readiness level is not suited to the task, firstly the deficiencies should be addressed and then the performance task should be started. In this way, students can tackle the performance task on their own and with less support of their families.*

***E7:*** *It is necessary to take into account the students' cognitive and emotional characteristics, the appropriateness of the students' readiness and the availability of opportunities.*

***E8:*** *While assigning performance tasks, importance should be given to the students' developmental characteristics. In addition to the cognitive and emotional characteristics, it is necessary to take into account the turbulence in the psychology of this group of adolescents.*

c) Task strength should be emphasized: Teachers should emphasise clearly what high-level skills the students are expected to develop. This approach motivated the students. Some of their sample views are as follows.

***E5:*** *With interdisciplinary tasks assigned to students, the skills of the students will develop through this task which can remove subject options limitation from this study. At the same time, the motivation of learners who know their skills in other subjects will be strengthened, will also increase….It should be emphasised that task are directed towards the developmental areas of the students in this process.*

***E8:*** *The tasks that the students will perform should be part of a pragmatic structure and whatever is to be done should contribute to the students getting into university. If a student devotes her time to study, the task needs to be structured so that she can realise emotional satisfaction, and so she can see concrete contributions to her life in order to plan her life. The performance tasks and projects can make students feel useful in terms of their own identity development. Teachers need to emphasise the unique strengths of each work.*

d) The rubric and a sample task should be presented: It is advisable to present the rubric and a sample task report to the students. It is important in terms of time management and motivation. Example opinions are below.

***E1:*** *.... The rubrics must also be shared with the students beforehand. In this way, the students will understand the process better and will not need the support of others by easily guiding themselves in*

*their works ... ... With the desired works, sample works can be shared with the students. It can also be a guide for students.*

***E3:*** *... the student must be informed about on what criteria they will be assessed. This is important for the students in structuring their tasks. They try to strengthen the criteria that they will be evaluated with rather than dealing with the questions of whether this or that will be assessed... Previously prepared tasks can be presented to the students, which will be a roadmap to them ...*

e) A work plan should be established: Establishing a work plan is crucial in terms of time management, especially where time is limited. It will support more effective work. Sample views on this were as follows.

***E1:*** *Especially about the project, in order to make time management more efficient, the teacher should create a schedule at the beginning of the task and pay attention to the student's compliance with the schedule.*

***E3:*** *A work plan must be established and this plan must be adhered to because it is important for student guidance.*

***E7:*** *The main things that should be considered in the performance tasks and the projects are specifying when the task will be fulfilled and whose support should be sought in the fulfilment of the task.*

f) It should conform to the students' environmental conditions: If the task expected from the students are given in accordance with the environment and facilities of the students, they will prepare it themselves and they will not need any external material support. Example views are as follows.

***E2:*** *I do not think there would be a material-based problem in the performance task; but if this kind of problem occurs, I think the main problem would be that the performance task is not suited to the environment.*

***E5:*** *It is important to remember that creativity is important, as everyone has to prepare tasks with their own means. Tasks using the material facilities do not make much sense; on the contrary, it should also be emphasised that it is more important for a student to prepare his work by using his creativity within the bounds of possibility... and that the expected outputs of the performance and project tasks should never be materially enforced, and it should be expected that each student should use easily accessible objects.*

### *3.2.3.2. Implication*

a) Collaborative Approach: By adopting a collaborative approach, the students' motivation will increase and their senses of responsibility develop. Sample views on this are presented as follows.

***E4:*** *Performance tasks can not only be done individually, but also as group work. By this means, the students can learn to collaborate as well as self-evaluate. Along with cognitive development, social skills can also be developed.*

***E7:*** *It will be useful to follow a process or approach that will transform learning not into competition, but into cooperation. For this, performance tasks or projects can be given as group-works, and not as individual works.*

***E8:*** *Students at the high school level are unfortunately not at a stage whereby they just sit and do their homework. They have difficulties in puberty; at this stage, friends are more important than family, teachers, managers. That's why it can be suggested that the assigned tasks should be realistic and that group work should be emphasised.*

b) Continuous feedback should be given: Teachers' continuous feedback is important in terms of effective teacher guidance. Thus, if there is a problem in a student's work, it will be immediately noticed and the necessary corrective action can be taken. Sample views regarding this are as follows.

***E1:*** *The evaluation of the product should not be left to the end, and intermediate checks are required. In short, piecemeal evaluation of the project will provide a more effective time management.*

***E4:*** *The people who know the students best are their teachers. They know what the students are able to do or what not to do. By deciding whether the task will be a group work or an individual work, they can make very good things happen. If students effortlessly find ready information (so they can copy it) and this is realised by the teacher and tolerated, then undesirable behaviours will continue to develop in the student. For this reason, teachers should be good evaluators and should be able to warn students without offending them regarding this practise.*

***E8:*** *Everything that the students do is to be assessed and absolute feedback must be given. Feedback is an important tool for the students.*

c) Student should be offered multiple choice and interdisciplinary tasks should be prepared: Offering students multiple choices for their performance tasks and including interdisciplinary knowledge can enhance their motivation. Thus, every student will be able to make choices in the direction of their knowledge, skills and abilities and will thereby feel less restricted in their obligation. Example views on this are as follows.

***E2:*** *If students are given choices over their performance tasks in accordance with the class level, the students can choose performance tasks based on their own interests. Therefore, the students can be better provided with the opportunity to do performance tasks themselves.*

***E3:*** *Performance tasks and the diversity of projects can be enhanced. Students can choose a task appropriate to their own resources....in interdisciplinary tasks, coordinated work by the teachers is crucial in determining the students' readiness level because it does not serve the purpose of interdisciplinary study when you do not know the time spared for the relevant topic in other courses and the student's learning.*

***E5:*** *Considering the multiple intelligence approach, especially in students' performance tasks, may increase the students' motivation. Otherwise, if individuals with different abilities are expected to give outputs for one or two skills, they will not be able to realise their abilities and their motivation will reduce, or they will even seek help from outside.*

#### *3.2.3.3. Evaluation*

a) Exhibition of the works: Exhibition of students' works can be seen as an important step in terms of student motivation. If the works are exhibited, the students can assume responsibility for their work. Sample views with regard to this are presented as follows.

***E7:*** *It should be stated that the purpose of these tasks is not to evaluate the students, but to encourage them to produce innovative products and ideas with the knowledge that they have acquired. For this, physical, technical and financial support should be provided to the students in order to fulfil such activities at the school level. Products produced by students can be introduced to the school environment and local community in order for these activities to reach the masses.*

***E8:*** *Sharing the works done of both successful and unsuccessful students may increase their motivation. If the tasks and projects assigned involve an element of social responsibility, it can result in increased motivation of students in search of identity.*

## 4. DISCUSSION

This study puts forward a model which aims to examine performance tasks being implemented at Turkish high school education institutions through analysing problems as perceived by the teachers and suggesting a solution to those problems.

In the first part of the study, it was seen that the problems encountered in the implementation of performance tasks generally matched the problems encountered with performance tasks in primary education institutions. The problems determined in this study are also parallel to the published literature (Ak, 2013; Arı, 2010; Arslan, 2013; Aydın et al., 2015; Benli, 2010; Dilekmen & Aydoğdu, 2015; Göçgen, 2016; Karakuş & Karakoç Öztürk, 2016; Metin, 2013;

Sarı, 2014). These problems were difficulties experienced in time management, students creating plagiarising products by copy-pasting from the Internet, considering performance tasks as grade-oriented, overcrowded classes, and problems encountered in the preparation and assessment of performance tasks. When the results of the study are examined, it appears that the problems encountered in the implementation of performance tasks did not differ at the primary or high school education level.

In order to be able to overcome the problems faced, academicians were consulted in the fields of Measurement and Evaluation in Education, Curriculum and Instruction, Educational Administration and Planning, and Psychological Counselling and Guidance. Three main stakeholders were involved in the model created in accordance with their opinions, which are the University, the MoNE and Teachers.

There has to be constant collaboration between these three stakeholders. University's should be aware of problems in implementations and should be able to present solutions to them; the MoNE should provide the necessary conditions to support teachers; and teachers should share the difficulties, problems and disabilities they encounter with the other stakeholders.

The university stakeholder focuses on the fully-equipped graduation of teacher candidates who have not yet fulfilled their teaching profession. For this reason, teacher candidates should be highly mastered in their undergraduate educational sciences courses. In particular, it could be understood from the interviews that the content of the Measurement and Assessment course in Education should be updated in such a way to be of more use when teachers start to practice their profession. Another suggestion was that the subjects of development, implementation, and evaluation of measurement instruments which cannot be covered in detail on the course could be given under a different course within the curriculum. If prospective teachers undertake more practical work, they will be more aware of the essential features of performance tasks while practising their teaching profession. Besides, it should also not be forgotten that it does not mean that prospective teachers will implement performance tasks more effectively if they are successful only in the Measurement and Evaluation in Education course. With the Curriculum and Instructional training course, they can use performance tasks as a learning activity. Performance assessment can be seen as part of classroom instruction (Marzano, Pickering, & McTighe, 1993) and it is an important tool for learning and teaching (Darling-Hammond, 1994; Messick, 1994). Implementation in the classroom with a classroom management course can make it easier to implement these tasks through facilitating its implementation. Also, with the Psychological Counselling and Guidance course, performance tasks can be created by paying attention not only to the cognitive level of the students, but also to the students' developmental characteristics. For the teachers engaged in the teaching profession; seminars, workshops and in-service teacher training are recommended by the universities. Therefore, both the fulfilment of lifelong learning principle and the teachers' lack of knowledge should be considered.

For the MoNE as a stakeholder, it is suggested to create a pool of performance tasks that teachers can use. Teachers can perform the implementations more effectively if they have well-prepared performance tasks to hand and the rubrics with which to grade them. It is also necessary to provide environments where the teachers can communicate with each other and share information on a platform on which detailed examples are presented. In this respect, teachers from the same branches, but in different regions and who have different experiences can benefit from each other's experiences and learn from each other as colleagues. In addition, in-service teacher training provided by the MoNE should be determined according to the teachers' needs in order to be more effective. If necessary, training should be done within an online environment to enable active participation of each teacher. In the scope of in-service teacher training; it is suggested that teachers should be encouraged to participate in activities such as congresses, symposiums, training courses relevant to their branch and their personal

development activities. In this way, it can be ensured that teachers actualise themselves professionally.

The teacher as stakeholder is the proposed model's focus. Based on the findings of the teachers' opinions, it can be concluded that teachers do not see themselves as adequate in the preparation, implementation and evaluation of performance tasks. In line with this result, it is necessary for the teachers to completely know the features, types, implementation diversity and scoring types of the performance tasks. In the proposed model, the topics that teachers should pay attention to are grouped under the stages of preparation, implementation and evaluation.

In the preparation phase; the purpose of performance tasks should be presented clearly to the students (Herman, Aschbacher, & Winters, 1992; Perlman, 2003). Thus, the students will better understand why and what they are expected to do. In the literature, it is seen that in studies of teachers and students at the primary education level, the student's inability to understand the purpose of the task caused problems during the preparation of the performance task (Akdağ & Çoklar, 2009; Arı, 2010; Hacısalihoğlu, 2013). Similarly to the findings of the current study, Aydın et al. (2015) found that the performance tasks were not suitable for all students.

The appropriateness of the task to the environment and to the student's level will encourage the students to undertake performance tasks on their own. Taking into consideration the developmental characteristics of students, especially at the high school education level which corresponds to the adolescence period, tasting the sense of accomplishing something will encourage the student to do the performance tasks. This can also help to reduce the parental attitudes of helping with assignments (Güvey, 2009; Orhan, 2007; Şeker, 2009, Tüysüz, Karakuyu, & Tatar, 2010). Other points to be aware of during the preparation phase of the performance tasks are the rubrics and sharing the sample performance tasks with the students. Rubrics can also be used effectively for the feedback of performance tasks (Cargas, Williams, & Rosenberg, 2017; Chappuis, Stiggins, Chappuis, & Arter, 2014, McTighe, 2016; McTighe & Ferrara, 1998). One of the important problems faced by teachers in the implementation of the performance tasks is lack of time. It is recommended to prepare work plans so that they can solve this problem.

In the implication phase; the basis of the proposals is the diversification of the implementation of performance tasks. Some suggestions were provided to the problems encountered. These problems are that the classrooms are crowded, that the student's performance tasks are only seen an opportunity for grading, that it takes too much time to assess the tasks as teachers have too many performance tasks to supervise, and that the Internet can be used as a source. When the experts' opinions are considered in general; the importance is seen of teacher consultation in the implementation of performance tasks. When overcrowded classes and the characteristics of the students' developmental period are considered, it can be suggested that performance tasks should be organised within the context of a collaborative approach. If the purpose is to increase student motivation, the performance tasks that give students the opportunity to choose the performance task can be prepared. Instead of assigning individual performance tasks for each course, the workload can be reduced by preparing interdisciplinary performance tasks.

In the evaluation phase; it is recommended to provide rubrics from which to grade the students' tasks, provide feedback to the students, and to exhibit the students' works. In the assessment stage, the experts' opinions suggested displaying the students' performance tasks within the context of an activity, which may positively affect the students' attitudes towards the courses and their personality development. McTighe and Ferrara (1994) also stated that the presentation of performance tasks to family and friends positively influenced student motivation. School administrators and parents need to take an active part in this process in order to undertake such activities. In addition, the students must be given feedback at this stage of the performance

tasks. The rubrics can be used to structure the feedback. Clear and explicit scoring of the works provides an opportunity for the students to see their own strengths and weaknesses.

## ORCID

Melek Gülşah Şahin orcid.org/0000-0001-5139-9777
Nagihan Boztunç Öztürk orcid.org/0000-0002-2777-5311

## 5. REFERENCES

Ak, S. (2013). *The evaluation of performance tasks given at primary education according to students', teachers', and parents' opinions* (Unpublished Master's thesis). Çağ University, Mersin.

Akdağ, H., & Çoklar, A.N. (2009). İlköğretim 6. ve 7. sınıf öğrencilerinin sosyal bilgiler dersi proje ve performans görevlerini hazırlarken yararlandıkları kaynaklar, internetin yeri ve karşıklaştıkları güçlükler [The resources 6th and 7th grade students use while preparing projects and performance tasks for social studies course, use of the internet as a resource, and the difficulties that they experience]. *Adıyaman University Journal of Social Sciences, 2* (2), 1-16.

Arı, A. (2010). Öğretmenlere göre proje ve performans görevlerinin uygulanmasında karşılaşılan sorunlar [Difficulties which teachers face in implementing projects and performance tasks]. *Electronic Journal of Social Sciences, 9(*34), 32-55.

Arslan, A. (2013). Türkçe öğretmenlerinin proje ve performans ödevleri ile ilgili görüşlerinin değerlendirilmesi [Assessment of views of turkish teachers concerning projects and performance assignments]. *Journal of Language and Literature Education, 2*(6), 11-22.

Aslan Yolcu, F. (2013). *İlköğretim düzeyinde performans görevi ve proje uygulamaları sürecinde disiplinler arası yaklaşımın etkililiği üzerine bir çalışma* [*Effectiveness of the ınterdisciplinary approach on process of performance task and project implementation at primary level*] (Unpublished Doctoral dissertation). Hacettepe University, Ankara.

Aydın, H., Yörek, N., & Uğulu, İ. (2015). Lise öğrencilerinin performans ödevlerine karşı tutumlarının bazı değişkenler açısından incelenmesi [An investigation of some variables related students' attitudes towards performance assignments]. *Mustafa Kemal University Journal of Graduate School of Social Sciences, 12*(31), 211-229.

Bal, A. P. (2012). İlköğretim öğrencilerinin matematik dersi performans görevi hazırlama sürecine ilişkin görüşleri ve yaşadıkları sorunlar [Primary school students' views and challenges on performance task preparation process in mathematics course]. *Pegem Journal of Education and Instruction, 2*(1), 11-23.

Bayrakçı, Ö. (2010). *İlköğretim sosyal bilgiler dersindeki araştırma ödevlerinin (etkinlik, performans ve proje) öğrencilerin sosyalleşmesine katkısı* [*The contribution of the research tasks (projects, performances and activities) in social sciences to the socialization of the students*] (Unpublished Master's thesis). Sakarya University, Sakarya.

Benli, N. (2010). *İlköğretim I. kademede verilen performans görevlerinin öğretmen ve veli görüşleri açısından değerlendirilmesi* [*The evaluation of performance tasks' given in elementary education according to teachers' and parents' views*] (Unpublished Master's thesis). Mustafa Kemal University, Hatay.

Cargas, S., Williams, S., & Rosenberg, M. (2017). An approach to teaching critical thinking across disciplines using performance tasks with a common rubric. *Thinking Skills and Creativity, 26*, 24-37.

Chappuis, J., Stiggins, R., Chappuis, S., & Arter, J. (2014). *Classroom assessment for student learning: Doing it right - Using it well*. Second Edition. Pearson

Çiftçi, S. (2010). İlköğretim birinci kademe 4. ve 5. sınıf öğretmenlerinin performans görevlerine ilişkin görüşleri [The opinions of the teachers in upper primary classes concerning the student performance tasks]. *Elementary Education Online*, *9*(3), 934-951.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5-30.

Dilekmen, M., & Aydoğdu, F. (2015). Sınıf öğretmenlerinin performans görevlerine ilişkin algılarının değerlendirilmesi [The evaluation of primary- school teachers' perceptions about performance tasks]. *Erzincan Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *8*(1), 51-62.

Ekici, E. (2016). *Görsel sanatlar eğitiminin öğrencilerin performans görevlerine katkılarına ilişkin öğrenci görüşleri [About the contribution of visual art education to the performance duties of students]* (Unpublished Master's thesis). Gazi University, Ankara.

Furat, E. (2009). *Performans görevlerinin öğrencielrin fen ve teknoloji dersine ilişkin tutumlarına ve eleştirel düşünme eğilimlerine etkisi* [*Influence of the performance assignments on attitudes of the students towards science and technology courses and their tendencies to critical thinking*] (Unpublished Master's thesis). Ankara University, Ankara.

Gelbal, S., & Kelecioğlu, H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar [Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront]. *H. U. Journal of Education*, *33*, 135-145.

Göçgen, S. (2016). *Ortaokul 7. sınıf öğrencilerinin sosyal bilgiler dersinde performans görevi hazırlarken karşılaştıkları sorunların incelenmesi* [*Examination of the problems 7th grade secondary school students encounter when they are preparing their performance mission*] (Unpublished Master's thesis). Atatürk University, Erzurum.

Gürel, N. (2012). *Performans görevlerinin öğrencilerin matematik başarısına ve tutumuna etkisi* [*The effects of performance task on students' math achievement and attitude towards mathematics*] (Unpublished Master's thesis). Mehmet Akif Ersoy University, Burdur.

Güvey, E. (2009). *İlköğretim 1-5. sınıf öğretim programlarında yer alan proje ve performans görevlerine ilişkin öğretmen ve veli görüşleri* [*Opinions of teachers and parents related to projects and performance tasks which take part in primary school 1-5 grade education programs*] (Unpublished Master's thesis). Eskişehir Osmangazi University, Eskişehir.

Hacısalihoğlu, M. (2013). *6. sınıf öğrencilerinin proje ve performans görevleriyle ilgili görüşleri ve karşılaştıkları güçlükler* [*6th grade students' opinions and encountered difficulties in project and performance tasks*] (Unpublished Master's thesis). Karadeniz Technical University, Trabzon.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Karakuş, M., & Karakoç Öztürk, B. (2016). İlköğretim Türkçe öğretmenlerinin performans değerlendirmeye yönelik görüşlerinin incelenmesi [A study of elementary school Turkish language teachers' views on performance evaluation]. *Journal of Çukurova University Institute of Social Sciences*, *25*(1), 283-302.

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2009). *Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme.* Ankara: Pegem Akademi.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes; performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.

McTighe, J. (2016). *Designing authentic performance tasks and rubrics.* Retrieved 16 May, 2018 from http://schd.ws/hosted_files/stlinstl2017/23/McTighe%20Handout%20%232%20MISD%206.6.17.pdf.

McTighe, J., & Ferrara, S. (1994). *Performance-based assessment in the classroom.* Retrieved 13 October, 2017 from https://www.jaymctighe.com/wordpress/wp-content/uploads/2011/04/Performance-Based-Assessment-in-the-Classroom.pdf.

McTighe, J., & Ferrara, S. (1998). *Assessing Learning in the Classroom: Student Assessment Series*. Washington, DC: National Education Association.

Messick, S. (1994). The interplay of evidence and consequences in validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Metin, M. (2013). Öğretmenlerin performans görevlerini hazırlarken ve uygularken karşılaştığı sorunlar [Teachers' difficulties in preparation and implementation of performance task]. *Educational Sciences: Theory & Practice, 13*(3), 1645-1673.

Milli Eğitim Bakanlığı. (2013). *Milli eğitim bakanlığı ortaöğretim kurumları yönetmeliği.* Retrieved 05 December, 2017 from http://www.resmigazete.gov.tr/eskiler/2013/09/20130907-4.htm.

OECD. (2012). B*etter skills, better jobs, better lives: a strategic approach to skills policies.* Paris: OECD Publishing.

OECD. (2017). *Pedagogical knowledge and the changing nature of the teaching profession*. Paris: OECD Publishing.

Orhan, A.T. (2007). *Fen eğitiminde alternatif ölçme ve değerlendirme yöntemlerinin ilköğretim öğretmen adayı, öğretmen ve öğrenci boyutu dikkate alınarak incelenmesi [Exploring the alternative assessment strategies by considering primary school pre-service teachers and students dimensions]*. (Unpublished Doctoral dissertation). Gazi University, Ankara.

O'Sullivan, M. K., & Dallas, K. B. (2010). A Collaborative approach to implementing 21st Century skills in a high school senior research class . *Education Libraries, 33*(1), 3-9.

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's thesis). Karadeniz Technical University, Trabzon.

P21Skills. (2007). *21st Century skills assessment: A partnership for 21st century skills e-paper*. Retrieved 12 October, 2017, from http://www.p21.org/storage/documents/21st_Century_Skills_Assessment_e-paper.pdf.

Palaz, T., Kılcan, B., Akbaba, B., & Çepni, O. (2015). Sosyal bilgiler dersinde verilen performans görevlerinin muhatapları tarafından değerlendirilmesi [The evaluation of performance homeworks assigned in social studies courses by respondents]. *International Journal of Turkish Education Sciences, 5*, 248-264.

Pellegrino, J. W. (2017). Teaching, learning and assessing 21st century skills. In S. G. (Ed.), *Pedagogical knowledge and the changing nature of the teaching profession* (pp. 223-251). Paris: OECD Publishing.

Perlman, C. C. (2003). Performance Assessment: Designing Appropriate Performance Tasks and Scoring Rubrics. In Wall, J.E., & Walz. G.R. (Ed.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators* (pp. 497-506). Greensboro, NC: CAPS Press.

Sarı, F. (2014). *İlköğretim birinci kademe performans görevlerinin öğretmen ve veli görüşleri açısından değerlendirilmesi [An evaluation of elementary school performance tasks from*

*teachers and parents' point of views]* (Unpublished Master's thesis). Ondokuz Mayıs University, Samsun.

Seyhan, A. (2017). Ortaöğretim öğrencilerinin performans görevlerine ilişkin görüşleri [View of the secondary school student related to performance task]. *Electronic Journal of Social Sciences, 16*(60), 60-75.

Sözer, E. (2015). *Çok seçenekli performans görevlerinin öğrencilerin akademik başarı, özgüven ve derse karşı tutumlarına etkisi* [*Influence of the multiple choice performance tasks on students' academic achievement, selfconfidence and attitudes of the students towards*] (Unpublished Master's thesis). Gazi University, Ankara.

Şeker, M. (2009). *İlköğretim 5. sınıf öğrencilerinin performans görevindeki başarıları ile ailelerinin eğitim-öğretim çalışmalarına katılım düzeyleri arasındaki ilişkinin belirlenmesi* [*An investigation on the relation between the success of fifth class' students' on performance works and the rate of their family support on educational -teaching works]*. (Unpublished Master's thesis). Çukurova University, Adana.

Tanrıverdi, T. (2012). *İlköğretim dördüncü ve beşinci sınıf öğretmen ve öğrencilerinin sosyal bilgiler dersinde performans görevine yönelik görüşleri* [*Fourth and fifth grade primary school teachers' and students' views about performance tasks in social studies course*]. (Unpublished Master's thesis). Çukurova University, Adana.

Tavşancıl, E., & Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri*. İstanbul: Epsilon Yayınevi.

Türk Eğitim Derneği [TED]. (2014). *TEDMEM 2014 Eğitim Değerlendirme Raporu*. Retrieved 10 December, 2017 from https://tedmem.org/yayin/2014-egitim-degerlendirme-raporu.

Tüysüz, C., Karakuyu, Y., & Tatar, E. (2010). Fen ve teknoloji dersindeki performans görevlerine yönelik veli tutumlarının belirlenmesi [Opinions of students' parents about performance task in science and technology class]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 4(1), 108-122

Yıldırım, A., & Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri.* Ankara: Seçkin Yayınları.

Yıldırım, K. (2010). Nitel araştırmalarda niteliği arttırma [Raising the quality in qualitative research]. *Elementary Education Online, 9*(1), 79-92.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 611–630

DOI: 10.21449/ijate.455175

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# A Mixture Partial Credit Analysis of Math Anxiety

**İbrahim Burak Ölmez** [*,1] **Allan S. Cohen** [1]

[1] University of Georgia, Department of Educational Psychology, Athens, USA

**ARTICLE HISTORY**

*Received: 25 June 2018*
*Revised: 11 August 2018*
*Accepted: 25 August 2018*

**KEYWORDS**

Math anxiety,
Middle grades students,
Mixture partial credit model

**Abstract:** The purpose of this study was to investigate a new methodology for detection of differences in middle grades students' math anxiety. A mixture partial credit model analysis was used to detect distinct latent classes based on homogeneities in response patterns. The analysis detected two latent classes. Students in Class 1 had less anxiety about *apprehension of math lessons* and *use of mathematics in daily life*, and more *self-efficacy for mathematics* than students in Class 2. Students in both classes were similar in terms of *test and evaluation anxiety*. Moreover, students in Class 1 were found to be more successful in mathematics, mostly like mathematics and mathematics teachers, and have better educated mothers than students in Class 2. Manifest variables of gender, attending private or public schools, and education levels of fathers did not differ among the latent classes. Characterizing differences between members of each latent class extends recent advances in measuring math anxiety.

## 1. INTRODUCTION

Identifying affective characteristics such as anger, anxiety, and depression that students experience in school settings and focusing on these characteristics in order to improve students' learning are significant challenges for educators. Math anxiety, as one such characteristic, can be defined as "feelings of tension and anxiety that interfere with the manipulation of numbers and the solving of mathematical problems in a wide variety of ordinary life and academic situations" (Richardson & Suinn, 1972, p. 551). Indicators of math anxiety include physical sensations of discomfort and distress during mathematics test-taking situations, feeling pressure to have the correct answers, fears of making mistakes and not understanding the given word problems especially in front of peers in a classroom (Luo, Wang, & Luo, 2009), and physiological reactions such as sweaty palms, being sick, and vomiting (Harper & Daane, 1998).

Math anxiety has been shown to cause low academic performance (Ashcraft, 2002), reduced cognitive information-processing (Young, Wu, & Menon, 2012), low working memories and

CONTACT: İbrahim Burak Ölmez ✉ i.burakolmez@hotmail.com University of Georgia, Department of Educational Psychology, Athens, USA

ISSN-e: 2148-7456 /© IJATE 2018

spatial abilities (Novak & Tassell, 2017), and low perceptions of one's own mathematics abilities (Hembree, 1990). Low math abilities and low working memory, as well as non-supportive teachers can also be considered as important risk factors for the existence of math anxiety (Ashcraft, Krause, & Hopko, 2007). Math anxiety can lead to avoidance of taking math classes throughout the high school and college years and avoidance of selecting career paths involving mathematics (Ashcraft & Moore, 2009).

Identifying students with math anxiety in the middle grades is critical for dealing with math anxiety as early as possible because math anxiety is known to peak during the secondary grades (Hembree, 1990). Hill et al. (2016) considers the development of math anxiety as closely related to increasing educational demands when moving from lower grades to higher grades such as middle grades to secondary grades. Studies with prospective and in-service mathematics teachers have also demonstrated that math anxiety is even common in this group through their experiences with math anxiety as K-12 students, and this situation negatively influences how they teach mathematics to their students (Bekdemir, 2010; Gresham, 2017; Stoehr, 2017).

For detecting students' math anxiety, previous research has employed several methods. These include use of exploratory and confirmatory factor analysis to detect the dimensions of math anxiety (e.g., Baloğlu & Zelhart, 2007; Hopko, 2003; Kazelskis, 1998) and use of structural equation modeling to explain the relationship between math anxiety and several variables such as mathematics achievement (e.g., Harari, Vukovic, & Bailey, 2013; Krinzinger, Kaufmann, & Willmes, 2009; Meece, Wigfield, & Eccles, 1990). One common feature of these studies is that their analyses are at the total score level due to their use of the total scores on a math anxiety scale. One concern with use of total scores is that it may miss important information available in the patterns of item level responses. Another common feature of these studies is that all examinees in a given sample belong to a single population. In this study, we examine item level patterns of math anxiety with an eye to detecting differences that may exist between latent classes in the population.

Given the potential deleterious effects of math anxiety on achievement and career choice, better methods for accurately measuring math anxiety are important in order to be able to ameliorate its effects. One such method is the use of mixture item response theory (IRT) models (Mislevy & Verhelst, 1990; Rost, 1990). These models may be appropriate when distinct latent classes are suspected. In the context of the present study, a latent class indicates a statistically determined grouping of students with homogeneous response patterns. That is, distinct latent classes are reflected in the patterns of responses to items on the instrument (Bolt, Cohen, & Wollack, 2001). The classes are labeled as latent because they are not directly observable. Item parameters are allowed to differ between latent classes in a mixture IRT model reflecting the differences in response propensities between the latent classes (Izsák, Jacobson, de Araujo, & Orrill, 2012). Research on the use of mixture IRT models has shown that they can be useful in detecting latent classes of individuals that differ along one or more dimensions. Previous research, for example, has found mixture IRT models to be useful in understanding differences in teachers' mathematical reasoning (Izsák, Orrill, Cohen & Brown, 2010), differences in mathematical knowledge (Izsák et al., 2012), differences in response to test time limits (Bolt, Cohen & Wollack, 2002), differences in strategy use in solving problems (Bolt, Cohen & Wollack, 2001) and differences in personality traits such as depression (Hong & Min, 2007) and creativity (Sen, 2016). In addition, Cho, Bottge, Cohen and Kim (2011) have shown this method to provide instructionally useful information for individual latent classes. In this study, we used mixture IRT methodology to explore item level differences in patterns of math anxiety based on middle grades students' responses to a math anxiety scale.

The purpose of this study was to investigate the use of mixture IRT methodology for detecting distinct latent classes of math anxiety among middle grades students that differ in their item level patterns of math anxiety. The following research questions were addressed:

1. Are there distinct latent classes of middle grades students that differ in their math anxiety?

2. What does the existence of these latent classes imply about the different response patterns of math anxiety that exist in this population?

3. Do latent classes differ with respect to several manifest variables such as mathematics achievement, gender, liking mathematics, liking mathematics teachers, attending private or public school, education levels of mothers and fathers?

### 1.1. The Rasch Model

The Rasch Model is a probabilistic model which is used to express item difficulties and examinee abilities on the same scale. The probability of an examinee correctly answering an item is a function of the difference between the examinee's latent ability and the item difficulty (Rasch, 1960/1980). The probability of obtaining the correct answer with respect to each item is given as follows:

$$P(x=1|\theta_j, \beta_i) = \frac{\exp(\theta_j-\beta_i)}{1+\exp(\theta_j-\beta_i)} \quad (1)$$

where $\theta_j$ is the latent ability parameter of an examinee $j$ and $\beta_i$ is the difficulty of item $i$. The probability of answering a given item correctly is expected to be relatively higher for an examinee with higher ability compared to an examinee with lower ability.

### 1.2. The Mixture Rasch Model

The mixture Rasch model (MRM; Rost, 1990) is a combination of a latent class model (Lazarsfeld & Henry, 1968) and a Rasch model. In contrast to the standard Rasch model, which assumes that the same Rasch model applies to all examinees in the population, the MRM assumes that distinct latent classes exist in the population and that a different Rasch model applies to each. Hence, the MRM allows different Rasch models to apply to different latent classes in the population.

In the MRM, the relative difficulty of the items is determined by a class membership parameter. The number of items which the examinee is expected to answer correctly is influenced by a continuous latent ability specific to the latent class. For each item, the MRM estimates a separate item difficulty for each latent class and for each examinee, a probability of being a member of a particular latent class.

The dichotomous form of the mixture Rasch model is employed when an item can be scored in two categories, such as agree or disagree. This form can be expressed as follows:

$$P(x_{ij}=1|g, \theta_{jg},) = \frac{\exp(\theta_{jg}-\beta_{ig})}{1+\exp(\theta_{jg}-\beta_{ig})} \quad (2)$$

where $P$ is the probability of a correct response in the mixture Rasch model, $g$ is an index for the latent class ($g$ = 1, 2, …, $G$), $\theta_{jg}$ is the latent ability of an examinee $j$ within class $g$, and $\beta_{ig}$ is the difficulty parameter of item $i$ for class $g$. When there is only one latent class, the mixture Rasch model is the same as the Rasch model in equation (1).

The polytomous form of the mixture Rasch model is used when items are scored with more than two categories. This type of items can be used when an answer is given partial credit rather than full credit or when an answer is in one of several categories such as strongly agree, agree, neutral, disagree, or strongly disagree. This form of the Rasch model is called a partial credit model (PCM; Masters, 1982). The probability of an answer for the mixture form of this model, the mixture partial credit model (MixPCM), can be written as follows:

$$P(x_{ij}=k|\theta_{jg}) = \frac{\exp[\sum_{r=1}^{k}(\theta_{jg}-\delta_{irg})]}{\sum_{t=0}^{m_i}[\exp\sum_{r=1}^{t}(\theta_{jg}-\delta_{irg})]} \quad (3)$$

where $P$ is the probability that examinee $j$ gives a response in category $k$ of item $i$, $\theta_{jg}$ is the latent trait of an examinee $j$ in latent class $g$, and $\delta_{irg}$ is a threshold parameter indicating the intersection of adjacent category response curves.

As can be seen in equation (3), the relationship between the probability of selecting a response in a given category and the latent trait is allowed to vary between latent classes. The differences in response patterns to each item of a questionnaire reflect homogeneities in characteristics of members of each latent class. In the MixPCM, the relative difficulty of a particular response category among the ordered categories is determined by a class membership parameter and the number of items answered. In this way, it is possible that the MixPCM could assign two examinees with similar test scores to different latent classes as a result of the differences in their response patterns.

## 2. METHOD

### 2.1. Participants

The sample consisted of 244 Turkish 6th and 7th grade students attending public and private schools in southwestern Turkey (Table 1). Parental consent was obtained through signed letters prior to the study.

**Table 1.** Gender and grade levels of the participants

| Gender | N | % | Grade Level | N | % | School Type | N | % |
|---|---|---|---|---|---|---|---|---|
| Male | 128 | 52.5 | 6th Grade | 120 | 49.2 | Public | 152 | 37.7 |
| Female | 116 | 47.5 | 7th Grade | 124 | 50.8 | Private | 92 | 62.3 |
| Total | 244 | | | | | | | |

### 2.2. Instruments

The *Math Anxiety Scale* (MANX; Erol, 1989) is a 45-item scale written in Turkish. Each item has four options, scored from 1 (“never”) to 4 (“always”). Scores can range from 45 to 180 points. Higher scores demonstrate a higher level of math anxiety.

Erol (1989) reported an internal consistency reliability estimate for the MANX of .91 on a sample of 380 high school students. The internal consistency reliability estimate in this study was .90. This was consistent with previous results on the MANX of .92 on a sample of 754 middle school and high school students (Erktin, Dönmez & Özel, 2006). Erktin et al. detected four factors that explained 40% of the variance. These factors were *test and evaluation anxiety*, *apprehension of math lessons*, *use of mathematics in daily life*, and *self-efficacy for mathematics*. The English translation of the MANX and its underlying factors are presented in Appendix A.

Demographic information was also obtained on a questionnaire attached to the MANX about students' mathematics grades at the end of the previous semester (i.e., grades ranging from 1 to 5), their gender (i.e., "male" or "female"), whether or not they liked mathematics (i.e., "Yes" or "No"), whether or not they liked their mathematics teacher (i.e., "Yes" or "No"), the type of school they attended (i.e., "Public" or "Private"), and their parents' education levels (i.e., "illiterate", "primary school", "secondary school", or "college"). Students were able to complete the MANX and the demographic information questionnaire in 30 minutes.

### 2.3. Data Analysis

Before implementing the MixPCM analysis, the dimensionality of the data was checked for ensuring the unidimensionality assumption by using exploratory factor analysis. The data were analyzed using the MixPCM as implemented in the computer program WINMIRA (von Davier, 2001). An exploratory MixPCM analysis was used to determine the number of latent classes in the data. This was done by fitting different MixPCM models with different numbers of latent classes to the data. In this way, the MPCM was estimated with one, two, three, and four latent classes. Three information indices were compared to select the best fitting model: Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), and the consistent AIC (CAIC; Bozdogan, 1987). Each information criterion index is defined as follows:

$$\text{AIC} = -2 \log L + 2p$$

$$\text{BIC} = -2 \log L + p (\log N)$$

$$\text{CAIC} = -2 \log L + p (\log N + 1)$$

where $L$ is the value at the maximum of the likelihood, $p$ is the number of estimated parameters, and $N$ is the sample size. AIC, BIC, and CAIC all include penalty functions to modify the $-2 \times$ log likelihood for either the number of parameters or the sample size or both. Because AIC has been found to be less accurate due to its sensitivity to sample size (Baghaei & Carstensen, 2013; Li, Cohen, Kim, & Cho, 2009), the model with the smallest BIC values were selected as the best fitting model. Next, the characteristics of each latent class were analyzed by examining differences in item thresholds between latent classes. In addition, differences in manifest variables between latent classes were evaluated using independent sample t-tests and chi-square tests between the latent classes.

## 3. FINDINGS

### 3.1. Unidimensionality for the Scale

An exploratory factor analysis using maximum likelihood estimation as implemented in the SPSS 16.0 software (SPSS Inc., 2007) indicated eigenvalues of the first three factors as 14.1, 2.6, and 2.5. The total variance explained by the first factor was 31.4%. Even though three factors were larger than 1.0, using Reckase's (1979) criterion, the assumption of unidimensionality also could be inferred.

### 3.2. Model Selection

Values for the three information indices are given in Table 2. Minimum values for AIC, BIC, and CAIC of 12883.82, 13705.72, and 13978.72, respectively, all suggested a two-class solution in the data. Based on results in Table 2, the two-class MixPCM model was determined to be the best fit.

**Table 2.** Model fit indices of the Mixture Rasch model

| Model | AIC | BIC | CAIC |
|---|---|---|---|
| One class | 13757.02 | 14166.47 | 14302.47 |
| Two classes | **12883.82** | **13705.72** | **13978.72** |
| Three classes | 13091.45 | 14325.81 | 14735.81 |
| Four classes | 13335.07 | 14981.88 | 15528.88 |

*Note.* AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = Consistent Akaike information criterion; the smallest information criterion index is bold.

After determining the best fitting model (see Table 2), mean assignment probabilities were calculated for each latent class (see Table 3). Students were classified into the latent class for which they had the highest mean assignment probability. Low off-diagonal values in comparison to the high diagonal values suggest that the two-class solution had good mean assignment probabilities, with Class 1 consisting of 126 students (51.5%) and Class 2 consisting of 118 students (48.5%).

**Table 3.** Mean assignment probabilities for the two-class solution

| | Proportion | Mean Assignment Probability | |
|---|---|---|---|
| Latent Class | in Latent Class | Class 1 | Class 2 |
| Class 1 | 51.5 | 0.999 | 0.001 |
| Class 2 | 48.5 | 0.005 | 0.995 |

Item thresholds indicate the point on the latent continuum between adjacent score categories and indicate the relative ease of endorsing each of the four categories by members of each of the two latent classes. As thresholds decrease, the likelihood of endorsement of particular response category increases. Item thresholds for each latent class are plotted in Figure 1 and Figure 2. Thresholds lower on the scale (e.g., -3, -2) indicate that examinees had a greater propensity to endorse that response category. Similarly, thresholds higher on the scale (e.g., 2, 3) indicate that examinees had a greater propensity to endorse a higher response category. Thresholds may differ by latent class. This means that, the relative propensity for endorsing a category of an item is specific to each latent class. Because the MANX has four response categories ranging from "never" to "always" for each item, there are three possible thresholds that can be used to interpret the math anxiety level as follows:

Categories: "never" "sometimes" "usually" "always"

Scores: (1) (2) (3) (4)

Thresholds: |--------T1--------|--------T2--------|--------T3--------|

For example, if an examinee's trait level is smaller than the first threshold, then the response is expected to be "never." If an examinee's trait level is smaller than the second threshold but larger than the first threshold, then the response is expected to be "sometimes."

Figure 1 and Figure 2 present plots of the item thresholds for Class 1 and Class 2 (see also Appendix A for item threshold values). It is evident that students in Class 2 were less variable in making their endorsements than students in Class 2, with thresholds ranging from -7.41 to

9.00. Students in Class 1 had lower tendency to endorse items above threshold 1, but greater tendency to endorse items above threshold 3.

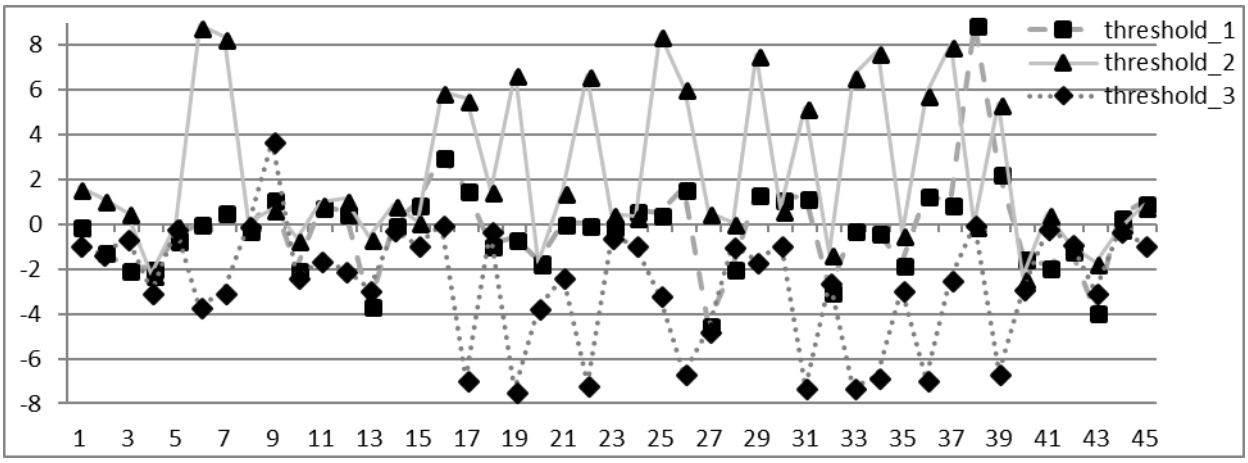



**Figure 1.** Item thresholds for Class 1

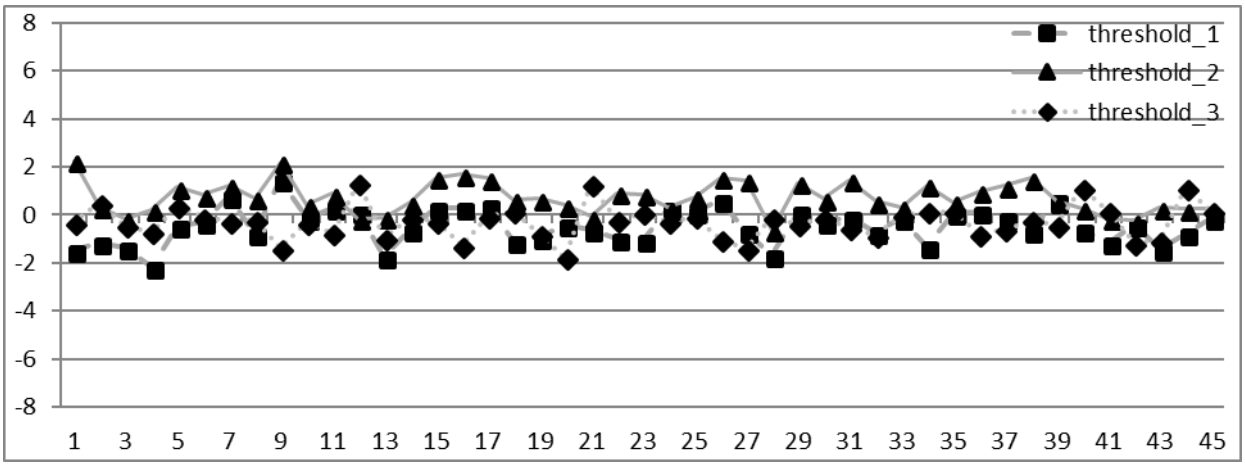



**Figure 2.** Item thresholds for Class 2

### 3.3. Analyses of Mean Item Thresholds and Item Response Distributions

The mean item threshold is the mean of all item thresholds for an item (Masters, 1982). Higher mean item thresholds indicate lower propensities of endorsement. In addition to the mean item thresholds, item response distributions were compared between the two latent classes to examine similarities and differences in item responses for each latent class. Analyses of the mean item thresholds and the item response distributions led to three main results: (1) Students in Class 1 were less anxious than students in Class 2 in terms of having anxiety about *apprehension of math lessons* and *use of mathematics in daily life*. Students in Class 1 also expressed feelings of more *self-efficacy for mathematics*, as evidenced by greater propensity to endorse items such as feeling comfortable asking the teacher questions in class. Students in both latent classes, however, had similar levels of *test and evaluation anxiety*.

Differences in the mean item thresholds of 1 logit or more were considered as reflecting differences between the two latent classes. The mean item thresholds are provided for the two latent classes in Appendix B. Items 4, 6, 7, 9, 10, 13, 16, 20, 25, 27, 29, 32, 35, 37, 38, 40, and 43 appeared to be have different response propensities for Class 1 and Class 2. Moreover, these response propensity differences in the mean item thresholds are plotted in Figure 3 along with descriptions of the content of each item in ascending order of the mean item thresholds.

| Logit | Class 1 | Class 2 |
|---|---|---|
| 3 | Bothered by math book (16); Uneasiness with calculations in daily life (38) | |
| 2 | Rejecting help (29)<br>Uneasiness with the next class (37)<br>Butterflies in one's stomach (7)<br>Feeling uneasy (25)<br>Panic (6); Confusing the change (9) | |
| 1 | | Confusing the change (9); Butterflies in one's stomach (7)<br>Rejecting help (29); Feeling uneasy (25)<br>Bothered by math book (16); Liking math problems (40); Liking math sections (35)<br>Uneasiness with calculations in daily life (38); Uneasiness because the next class is math (37); |
| 0 | | Panic (6)<br>Keeping the account (10)<br>Easy to ask teacher questions (27); Enjoying geometry (32)<br>First math class with hope (20)<br>Being relaxed while waiting for exam results (43)<br>Liking math homework (4); |
| -1 | Liking math sections (35); Keeping the account (10) | Liking numbers or graph (13) |
| -2 | Liking math problems (40); Enjoying geometry (32); First math class with hope (20)<br>Liking math homework (4); Liking numbers or graph (13)<br>Being relaxed while waiting for exam results (43)<br>Easy to ask teacher questions (27) | |
| -3 | | |

**Figure 3.** Ordering items by mean item thresholds for each class

*Note.* Items with difference of 1 logit or more are ordered. Parentheses refer to item numbers. Vertical axis is the logit scale.

Items 6, 7, 16 and 37 reflect anxiety in the form of *apprehension of math lessons*. Students in Class 1 were less likely than students in Class 2 to select a higher numbered category. This is represented in Figure 3 by showing the item text for these items listed under Class 1 and Class 2. For example, Item 16, "Bothered by Math book," is shown under the columns of Class 1 and Class 2 in Figure 3. The mean item threshold for Class 1 was 3.00 but for Class 2, it was .25 (see Appendix B). The proportions selecting the options of "never" and "always" in Class 1 were 98.7% and 0% respectively, in contrast to 53.3% and 15.1% in Class 2, respectively. Similarly, for Item 37, "Uneasiness because the next class is math," the mean item threshold for Class 1 was 2.19, and for Class 2, it was .15. These values indicate that students in Class 1 were less likely to endorse a higher numbered category for Item 37 than students in Class 2.

Item 6 asked students to indicate how much they "Panic" when a lot of mathematics problems are given as homework and Item 7 asked whether they get "Butterflies in one's stomach" when studying a hard mathematics topic. For Item 6 and Item 7, Class 1 mean item thresholds were 1.78 and 2, respectively, and for Class 2, the mean item thresholds were .14 and .58, respectively. This suggests that students in Class 1 mostly agreed with the option "never" (80.2% for Item 6 and 87.1% for Item 7) in comparison to students in Class 2 (38.6% for Item 6 and 64.2% for Item 7).

On the other hand, students in Class 1 more strongly endorsed items 4, 10, 13, 20, 32, 35, and 40. These contained positive statements related to mathematics lessons such as liking mathematics sections of the social classes, liking mathematics problems and homework, and liking numbers and geometry. For example, for Item 40, "Liking math problems," the mean item threshold for Class 1 was -2.21 and for Class 2, it was .25. The proportion selecting the option "always" was 70.3% in Class 1 as compared to 7.8% in Class 2.

Items 9, 29 and 38 focused on anxiety about the *use of mathematics in daily life*. Students in Class 1 were less likely to select a higher category for these items than for students in Class 2 (see Figure 3). As an example, a marked difference occurred in Item 38, "Uneasiness with calculations in daily life" with Class 1 mean item threshold was 3.00 and Class 2 mean item threshold was .19. For this item, all students in Class 1 selected the option "never," while 34% of the students in Class 2 made this choice.

Items 2, 3, 8, 11, 14, 18, 19, 21, 22, 24, 25, 28, 30, 33, 41, 42, and 44 asked students to rate their ideas about *test and evaluation anxiety*. Mean item thresholds as well as the distributions of responses were similar for both latent classes. The one exception to this trend occurred on Item 25, "Feeling uneasy." The item asked students to rate if they felt uneasy the week before a math exam. The mean item threshold for Class 1 was 1.95 and for Class 2, it was .27. For this item, 85.8% of the students in Class 1 selected the option "never," and 47.7% of the students in Class 2 selected this choice.

Finally, Items 27 and 43 involved *self-efficacy for mathematics*. Mean item thresholds and the distribution of responses were different across response categories for the two latent classes (see Figure 3). Item 27 asked whether students found it "Easy to ask teacher questions." The mean item threshold for Class 1 was -2.87 and for Class 2, it was item -.23. The "always" option was selected by 76.3% of the Class 1 students, but only 26.2% of the Class 2 students answered "always." For Item 43, "Being relaxed while waiting for exam results," the mean item threshold for Class 1 was -2.84 and for Class 2 it was -.73. For this item, the proportion selecting the option "always" in Class 1 was 72.8% compared to 39.5% in Class 2. Thus, Item 43 was easier Class 1 students to respond "always" and harder to respond "never."

### 3.4. The Relationships between Manifest Variables and Latent Class Membership

To obtain detailed information about the characteristics of each latent class, the relationships between the manifest variables and latent class membership were examined using independent sample t-tests and chi-square tests. Based on students' responses, the mean ability logit score in Class 1 (M = -1.34) was significantly lower than the mean score in Class 2 (M = -.27) (t (df = 148) = -12.94, $p$ < .01). This indicated that students in Class 1 were less anxious than those in Class 2. Regarding mathematics achievement, there was a significant difference between the two latent classes (t (df = 111) = 3.71, $p$ < .01), suggesting that the students in Class 1 were more successful in mathematics than students in Class 2. Furthermore, mother's education level was significantly higher for students in Class 1 than Class 2 (t (df = 136) = 2.36, $p$ <.05), but there was no significant difference between the latent classes with respect to fathers' education level (t (df= 136) = 1.07, $p$ = .29).

A chi-square test between the two latent classes for gender was not significant ( $\chi^2(1)$ = .98, $p$ = .32). On the other hand, the associations between students' liking mathematics and liking their mathematics teachers, and latent class membership were found as significant ($\chi^2(1)$ = 11.83, $p$ < .01 and $\chi^2(1)$ = 6.30, $p$ < .01, respectively). This indicated that students' being in Class 1 or Class 2 is related to their liking mathematics and liking their mathematics teachers. Finally, there was no association between students' attending private or public schools and latent class membership ($\chi^2(1)$ = .57, $p$ = .45).

## 4. DISCUSSION AND CONCLUSION

In the present study, a mixture partial credit model (MixPCM) was used to detect differences in math anxiety of middle grades students. With respect to the first research question regarding existence of latent classes, two latent classes were detected, indicating the presence of distinct latent classes exist of math anxiety. The classes were similar in size but had different patterns of math anxiety.

With respect to the second research question regarding characteristics of latent classes, Class 1 consisted of students who were less anxious about *apprehension of math lessons* and *use of mathematics in daily life*, and who had more *self-efficacy for mathematics* than students in Class 2. However, there did not exist any differences between Classes 1 and 2 in terms of *test and evaluation anxiety*.

With respect to the third research question regarding the effects of manifest variables on class membership, students in Class 1 also appeared to be less anxious than students in Class 2, as evidenced by their mean scores on the MANX. In addition, students in Class 1 were reported being more successful in mathematics, liked mathematics and mathematics teachers, and had better educated mothers in comparison to students in Class 2. No significant association was found between the two latent classes for either gender, attending private or public schools, or fathers' education level.

The results reported here on the relationships between math anxiety and the manifest variables were consistent with findings in the literature. As previous research indicates that math anxiety was negatively related to mathematics achievement (e.g., Hembree, 1990), students in Class 1, in the present study, reported being less anxious and more successful in mathematics. Moreover, previous research on the effects of positive attitudes and education levels of mothers on math anxiety has found that positive attitudes towards mathematics and education levels of mothers were negatively associated with math anxiety (e.g., Engelhard, 1990; Meece, Wigfield, & Eccles, 1990). Similarly, students in Class 1 reported being less anxious and being more likely to have positive attitudes such as enjoying mathematics and liking their mathematics teachers and to have mothers with higher education levels than students in Class 2. On the other hand,

there did not appear to be consensus from previous research regarding the effects of gender, type of school attended, or education levels of fathers on math anxiety (Alkan, 2018). Some studies reported significant effects of gender on math anxiety (e.g., Luo et al., 2009; Wigfield & Meece, 1988), but others found no associations (e.g., Birgin, Baloglu, Catlioglu, & Gurbuz, 2010). Incorporating the analysis of manifest variables into the MixPCM analysis and obtaining consistent results with previous research strengthen the validity of the interpretation about the characteristics of each latent class reported in this study.

The results of this study suggest that within a population of students there exist latent classes that differ in their math anxiety. Use of the MixPCM provided information at the item level that revealed potentially useful distinctions that may not be easily detectable at the total score level. As such, this methodology provides a useful tool for identifying latent classes of students with different patterns of math anxiety. In school settings, teachers can use results from the MixPCM to detect those students and provide appropriate interventions specific to the needs of students in each latent class. For example, they can provide a classroom environment supporting students in each latent class instead of highlighting their mistakes; focus on reducing some particular students' anxiety levels towards mathematics lessons by not calling on these students to solve a problem at the board; engage some students with more mathematics related activities in daily life by presenting simulated real-life situations and asking word problems in a real-life context; and help some students build self-confidence for mathematics through asking mathematical problems from simple to more complex.

The present study takes a step towards detecting math anxiety by examining the utility of a new methodology for detection of distinct latent classes based on different patterns of math anxiety. The results provide evidence that the MixPCM, when applied to a math anxiety scale, can provide fine-grained information about latent classes of middle grades students and their differences in math anxiety.

Future studies focusing on detecting and characterizing latent classes of students with respect to math anxiety in different populations would be helpful. Such studies will help provide a more complete understanding about middle grades students' math anxiety.

## ORCID

İbrahim Burak Ölmez https://orcid.org/0000-0002-4931-2174

Allan S. Cohen https://orcid.org/0000-0002-8776-9378

## 5. REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Alkan, V. (2018). A systematic review research: 'Mathematics anxiety' in Turkey. *International Journal of Assessment Tools in Education*, 5(3), 567–592.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science, 11*(2), 181–185.

Ashcraft, M. H., Krause, J. A., & Hopko, D. R. (2007). Is math anxiety a mathematical learning disability? In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 329–348). Baltimore: Brookes.

Ashcraft, M. H., & Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment, 27*(3), 197–205.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation, 18*(5), 1–12.

Baloğlu, M. & Zelhart, P. F. (2007). Psychometric properties of the revised mathematics anxiety rating scale. *The Psychological Record, 57*, 593–611.

Bekdemir, M. (2010). The pre-service teachers' mathematics anxiety related to depth of negative experiences in mathematics classroom while they were students. *Educational Studies in Mathematics, 75*, 311–328.

Birgin, O., Baloglu, M., Catlioglu, H., & Gurbuz, R. (2010). An investigation of mathematics anxiety among sixth through eighth grade students in Turkey. *Learning and Individual Differences, 20*, 654–658.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika, 52*, 345–370.

Cho, S.-J., Bottge, B. A., Cohen, A. S., & Kim, S.-H. (2011). Detecting cognitive change in the math skills of low-achieving adolescents. *The Journal of Special Education, 45*(2), 67–76.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133–148.

Engelhard, G. (1990). Math anxiety, mother's education, and the mathematics performance of adolescent boys and girls: Evidence from the U.S. and Thailand. *Journal of Psychology, 124*(3), 289–298.

Erktin, E., Dönmez, G., & Özel, S. (2006). Psychometric characteristics of the mathematics anxiety scale. *Education and Science, 31*(140), 26–33.

Erol, E. (1989). *Prevalence and correlates of math anxiety in Turkish high school students*. Unpublished master thesis, Bogazici University.

Gresham, G. (2017). Preservice to inservice: Does mathematics anxiety change with teaching experience. *Journal of Teacher Education, 69*(1), 90–107.

Harari, R. R., Vukovic, R. K., & Bailey, S. P. (2013). Mathematics anxiety in young children: An exploratory study. *The Journal of Experimental Education, 81*(4), 538–555.

Harper, N. W., & Daane, C. J. (1998). Causes and reduction of math anxiety in preservice elementary teachers. *Action in Teacher Education, 19*(4), 29–38.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*, 33–46.

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Szücs, D. (2016). Math anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences, 48*, 45–53.

Hong, S., & Min, S. (2007). Mixed Rasch modeling of the Self-Rating Depression Scale: Incorporating Latent Class and Rasch Rating Scale models. *Educational and Psychological Measurement, 67*(2), 280–299.

Hopko, D. R. (2003). Confirmatory factor analysis of the Math Anxiety Rating Scale-Revised. *Educational and Psychological Measurement, 63*(2), 336–351.

Izsák, A., Jacobson, E., de Araujo, Z., & Orrill, C. H. (2012). Measuring mathematical knowledge for teaching fractions with drawn quantities. *Journal for Research in Mathematics Education, 43*(4), 391–427.

Izsák, A., Orrill, C. H., Cohen, A. S., & Brown, R. E. (2010). Measuring middle grades teachers' understanding of rational numbers with the mixture Rasch model. *The Elementary School Journal, 110*, 279–300.

Kazelskis, R. (1998). Some dimensions of mathematics anxiety: A factor analysis across instruments. *Educational and Psychological Measurement, 58*, 623–633.

Krinzinger, H., Kaufmann, L., & Willmes, K. (2009). Math anxiety and math ability in early primary school years. *Journal of Psychoeducational Assessment, 27*(3), 206–225.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353–373.

Luo, X., Wang, F., & Luo, Z. (2009). Investigation and analysis of mathematics anxiety in middle school students. *Journal of Mathematics Education, 2*(2), 12–19.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*(1), 60–70.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.

Novak, E., & Tassell, J. L. (2017). Studying preservice teacher math anxiety and mathematics performance in geometry, word, and non-word problem solving. *Learning and Individual Differences, 54*, 20–29.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: The Danish Institute of Education Research. (Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago, IL: The University of Chicago Press)

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230.

Richardson F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology, 19*(6), 551–554.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Sen, S. (2016). Applying the mixed Rasch model to the Runco ideational behavior scale. *Creativity Research Journal, 28*(4), 426–434.

SPSS Inc. (2007). SPSS for Windows, Version 16.0. Chicago, SPSS Inc.

Stoehr, K. J. (2017). Mathematics anxiety: One size does not fit all. *Journal of Teacher Education, 68*(1), 69–84.

Von Davier, M. (2001). WINMIRA [Computer Software]. St. Paul, MN: Assessment Systems Corporation.

Wigfield, A. & Meece, J. L. (1988). Math anxiety in elementary and secondary school students. *Journal of Educational Psychology, 80*, 210–216.

Young, C. B., Wu, S. S., Menon, V. (2012). The neurodevelopmental basis of math anxiety. *Psychological Science, 23*(5), 492–501.

**Appendix A.** Item information of the MANX across three latent classes

| Items | Content | Underlying Factor | Class 1 threshold 1 | Class 1 threshold 2 | Class 1 threshold 3 | Class 2 threshold 1 | Class 2 threshold 2 | Class 2 threshold 3 |
|---|---|---|---|---|---|---|---|---|
| Item 1 | When my friend is asked to answer a question in Math class, I feel glad that I am not in his/her shoes. | Apprehension of Math Lessons | -0.02 | 1.67 | -0.91 | -1.46 | 2.25 | -0.33 |
| Item 2 | I panic when I start the Math part of a common test. | Test and Evaluation Anxiety | -1.14 | 1.14 | -1.31 | -1.15 | 0.32 | 0.49 |
| Item 3 | I am intimidated when asked to answer a question that I do not completely know the answer to. | Test and Evaluation Anxiety | -1.96 | 0.55 | -0.59 | -1.40 | -0.15 | -0.44 |
| Item 4 | I like doing Math homework.*** | Apprehension of Math Lessons | -1.92 | -2.21 | -3.00 | -2.19 | 0.21 | -0.73 |
| Item 5 | I dislike the formulas in the science classes. | Apprehension of Math Lessons | -0.66 | -0.04 | -0.17 | -0.46 | 1.15 | 0.37 |
| Item 6 | I panic when I am assigned homework which includes a lot of math problems. | Apprehension of Math Lessons | 0.11 | 8.89 | -3.67 | -0.30 | 0.82 | -0.10 |
| Item 7 | I feel butterflies in my stomach when I prepare to study a hard math topic. | Apprehension of Math Lessons | 0.64 | 8.36 | -3.00 | 0.75 | 1.25 | -0.26 |
| Item 8 | I become unable to think about anything one hour prior to math exam. | Test and Evaluation Anxiety | -0.19 | 0.06 | -0.06 | -0.81 | 0.71 | -0.22 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item 9 | I feel confused when I try to count the change I received from my purchase from the school cafeteria; most of the time I just get what is given to me without counting the change. | Use of Mathematics In Daily Life | 1.19 | 0.73 | 3.72 | 1.45 | 2.21 | -1.37 |
| Item 10 | I would like to keep the accounts for a school club or activity that I am participating in.*** | Apprehension of Math Lessons | -1.96 | -0.63 | -2.34 | -0.16 | 0.43 | -0.32 |
| Item 11 | I feel intimidated to check my math score when I am given my class report. | Test and Evaluation Anxiety | 0.82 | 0.97 | -1.60 | 0.31 | 0.88 | -0.78 |
| Item 12 | I feel reluctant to explain the problems even the ones I can solve. | Self-efficacy for Mathematics | 0.54 | 1.16 | -2.03 | 0.10 | -0.16 | 1.34 |
| Item 13 | I like any topic explained to me in numbers and graphs rather than verbal explanations.*** | Apprehension of Math Lessons | -3.54 | -0.58 | -2.895 | -1.73 | -0.11 | -0.97 |
| Item 14 | I feel terrible a day before the math exam. | Test and Evaluation Anxiety | 0.07 | 0.89 | -0.23 | -0.60 | 0.51 | -0.13 |
| Item 15 | Even if I think that a shop clerk gave me the wrong change, I say nothing because I cannot make calculations while somebody is watching me. | Use of Mathematics In Daily Life | 0.95 | 0.14 | -0.92 | 0.28 | 1.55 | -0.27 |
| Item 16 | The math book bothers me. | Apprehension of Math Lessons | 3.08 | 5.92 | 0.00 | 0.31 | 1.70 | -1.27 |
| Item 17 | I cannot even make an addition operation while somebody is watching. | Use of Mathematics In Daily Life | 1.60 | 5.61 | -6.89 | 0.39 | 1.52 | -0.05 |
| Item 18 | I become so nervous before the important math exams that I forget all I know. | Test and Evaluation Anxiety | -0.89 | 1.51 | -0.28 | -1.08 | 0.64 | 0.13 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item 19 | I feel afraid when the teacher gives a pop quiz on math. | Test and Evaluation Anxiety | -0.60 | 6.78 | -7.41 | -0.97 | 0.68 | -0.81 |
| Item 20 | I always enter the first math class of the year with hope.*** | Apprehension of Math Lessons | -1.62 | -1.68 | -3.68 | -0.43 | 0.41 | -1.79 |
| Item 21 | While studying for a math exam, I may not prepare enough at times because of worrying about the score I will get. | Test and Evaluation Anxiety | 0.11 | 1.49 | -2.36 | -0.60 | -0.07 | 1.26 |
| Item 22 | I feel an inability to succeed while going through the pages of the math book. | Test and Evaluation Anxiety | 0.04 | 6.69 | -7.14 | -1.01 | 0.93 | -0.22 |
| Item 23 | I cannot dare to ask the points I do not get in the math class. | Self-efficacy for Mathematics | -0.01 | 0.49 | -0.55 | -1.04 | 0.87 | 0.08 |
| Item 24 | I feel uneasy even when I calculate the GPA for my class report. | Test and Evaluation Anxiety | 0.67 | 0.38 | -0.92 | 0.27 | 0.29 | -0.29 |
| Item 25 | I feel uneasy a week before the math exam. | Test and Evaluation Anxiety | 0.52 | 8.48 | -3.15 | 0.14 | 0.76 | -0.08 |
| Item 26 | Even making calculation related to time gives me discomfort. | Use of Mathematics In Daily Life | 1.63 | 6.01 | -6.64 | 0.61 | 1.56 | -1.02 |
| Item 27 | I can easily ask something that I did not understand to the math teacher after the class.*** | Self-efficacy for Mathematics | -4.41 | 0.54 | -4.73 | -0.70 | 1.44 | -1.43 |
| Item 28 | I feel nervous and pessimistic while waiting for the announcement of the result for a math exam that I think I failed at. | Test and Evaluation Anxiety | -1.90 | 0.10 | -0.95 | -1.67 | -0.62 | -0.11 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item 29 | When I am asked to help a primary school student with his/her homework, I may refuse to help because I feel afraid that there may be some problems I couldn't solve. | Use of Mathematics In Daily Life | 1.41 | 7.59 | -1.63 | 0.12 | 1.34 | -0.40 |
| Item 30 | When I think of the math subjects I have to learn before graduating from high school, I doubt if I am ever going to finish school. | Test and Evaluation Anxiety | 1.17 | 0.71 | -0.88 | -0.33 | 0.68 | -0.11 |
| Item 31 | Dealing with numbers makes me annoyed. | Apprehension of Math Lessons | 1.22 | 5.25 | -7.26 | -0.11 | 1.44 | -0.54 |
| Item 32 | Geometry questions remind me of fun puzzles.*** | Apprehension of Math Lessons | -2.93 | -1.29 | -2.56 | -0.72 | 0.56 | -0.86 |
| Item 33 | I feel tense when my friend solves a problem and I cannot understand his/her solution. | Test and Evaluation Anxiety | -0.20 | 6.66 | -7.27 | -0.12 | 0.32 | -0.03 |
| Item 34 | I feel confused in math class. | Apprehension of Math Lessons | -0.30 | 7.75 | -6.77 | -1.31 | 1.25 | 0.13 |
| Item 35 | The most likable part of the social classes are the parts that consist of math, even if they are miniscule.*** | Apprehension of Math Lessons | -1.73 | -0.40 | -2.89 | 0.07 | 0.53 | 0.13 |
| Item 36 | I struggle with listening to the teacher in the math class. | Apprehension of Math Lessons | 1.36 | 5.84 | -6.90 | 0.14 | 0.98 | -0.79 |
| Item 37 | I feel uneasy when I know that the following lesson is math. | Apprehension of Math Lessons | 0.99 | 8.01 | -2.43 | -0.15 | 1.20 | -0.60 |
| Item 38 | I feel bothered by the necessity of making calculations by solving mathematical problems in my daily life even if they are simple. | Use of Mathematics In Daily Life | 9.00 | 0.00 | 0.00 | -0.70 | 1.52 | -0.25 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item 39 | I feel depressed by the math book. | Apprehension of Math Lessons | 2.35 | 5.40 | -6.62 | 0.61 | 0.56 | -0.43 |
| Item 40 | Opening any book on math and looking at one of its pages full of mathematical problems makes me happy.*** | Apprehension of Math Lessons | -1.43 | -2.35 | -2.86 | -0.64 | 0.27 | 1.13 |
| Item 41 | When I am given a problem to solve, I panic if I cannot remember the formula necessary for the solution. | Test and Evaluation Anxiety | -1.85 | 0.49 | -0.16 | -1.16 | -0.16 | 0.14 |
| Item 42 | Five minutes before the math exam, my heart starts starts beating fast. | Test and Evaluation Anxiety | -1.09 | -0.94 | -0.82 | -0.41 | -0.23 | -1.17 |
| Item 43 | When I think that I succeeded at a math exam, I feel relaxed and peaceful while waiting for the announcement of the results.*** | Self-efficacy for Mathematics | -3.86 | -1.70 | -2.99 | -1.44 | 0.29 | -1.05 |
| Item 44 | If the teacher asks me to solve a math problem that I have been working on for a while at the blackboard, I forget what I have done out of excitement. | Test and Evaluation Anxiety | 0.39 | -0.13 | -0.26 | -0.80 | 0.24 | 1.12 |
| Item 45 | If a friend asks me to solve a math problem that has been published in a magazine, I am afraid of being embarrassed by not being able to solve even the easiest problems. | Use of Mathematics In Daily Life | 0.10 | 0.85 | -0.88 | -0.12 | 0.24 | 0.14 |

*Note.* *** = Reverse-coded; MANX = Math Anxiety Scale (Erol 1989).

**Appendix B.** Mean item thresholds of each item for Class 1 and Class 2

| Items | Class 1 Mean Item Thresholds | Class 2 Mean Item Thresholds | Difference |
|---|---|---|---|
| Item 1 | 0.246 | 0.155 | 0.091 |
| Item 2 | -0.436 | -0.114 | -0.322 |
| Item 3 | -0.669 | -0.659 | -0.010 |
| Item 4 | -2.375 | -0.900 | -1.474* |
| Item 5 | -0.285 | 0.352 | -0.637 |
| Item 6 | 1.776 | 0.141 | 1.635* |
| Item 7 | 1.999 | 0.579 | 1.420* |
| Item 8 | -0.060 | -0.106 | 0.046 |
| Item 9 | 1.882 | 0.763 | 1.119* |
| Item 10 | -1.645 | -0.017 | -1.628* |
| Item 11 | 0.063 | 0.134 | -0.071 |
| Item 12 | -0.106 | 0.427 | -0.534 |
| Item 13 | -2.339 | -0.934 | -1.405* |
| Item 14 | 0.243 | -0.072 | 0.315 |
| Item 15 | 0.057 | 0.520 | -0.462 |
| Item 16 | 3.000 | 0.247 | 2.753* |
| Item 17 | 0.106 | 0.617 | -0.510 |
| Item 18 | 0.117 | -0.104 | 0.221 |
| Item 19 | -0.410 | -0.363 | -0.047 |
| Item 20 | -2.327 | -0.605 | -1.721* |
| Item 21 | -0.252 | 0.196 | -0.446 |
| Item 22 | -0.137 | -0.103 | -0.034 |
| Item 23 | -0.022 | -0.030 | 0.008 |
| Item 24 | 0.043 | 0.089 | -0.046 |
| Item 25 | 1.950 | 0.271 | 1.679* |
| Item 26 | 0.364 | 0.382 | -0.018 |
| Item 27 | -2.866 | -0.229 | -2.637* |
| Item 28 | -0.915 | -0.802 | -0.113 |
| Item 29 | 2.457 | 0.352 | 2.105* |
| Item 30 | 0.332 | 0.080 | 0.252 |

| | | | |
|---|---|---|---|
| Item 31 | -0.263 | 0.261 | -0.524 |
| Item 32 | -2.258 | -0.341 | -1.917* |
| Item 33 | -0.273 | 0.053 | -0.327 |
| Item 34 | 0.225 | 0.022 | 0.203 |
| Item 35 | -1.675 | 0.241 | -1.916* |
| Item 36 | 0.099 | 0.112 | -0.013 |
| Item 37 | 2.188 | 0.149 | 2.039* |
| Item 38 | 3.000 | 0.193 | 2.807* |
| Item 39 | 0.379 | 0.246 | 0.133 |
| Item 40 | -2.211 | 0.252 | -2.463* |
| Item 41 | -0.508 | -0.394 | -0.114 |
| Item 42 | -0.948 | -0.601 | -0.347 |
| Item 43 | -2.847 | -0.733 | -2.114* |
| Item 44 | 0.001 | 0.186 | -0.185 |
| Item 45 | 0.324 | 0.086 | 0.238 |

*Note.* * : significance at 1 or greater.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 631–644

DOI: 10.21449/ijate.409110

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Turkish Adaptation of the Family Adaptability and Cohesion Scale IV

**Turgut Turkdogan** [*,1] **Erdinc Duru** [1] **Murat Balkis** [1]

[1] Pamukkale University, Department of Psychological Counseling and Guidance, 20020 Denizli, TURKEY

**ARTICLE HISTORY**

*Received: 31 March 2018*
*Revised: 07 August 2018*
*Accepted: 27 August 2018*

**KEYWORDS**

Circumplex model, Family adaptability, Family cohesion, family communication, Family functioning, Family satisfaction.

**Abstract:** The aim of this study was to examine the psychometric properties of the Turkish version of Family Adaptability and Cohesion Scale IV (FACES IV) Questionnaire Package, and also to achieve a valid and reliable assessment tool for the further investigations of the Circumplex Model of Marital and Family Systems in Turkish culture. A total of 1613 (65.4% female, 34.6% male) university students agreed to participate in the study. The construct validity of the FACES IV was examined with confirmatory factor analysis. Also, the convergent validity, criterion-related validity, internal reliability, and test-retest reliability analyses were examined within the scope of validity and reliability studies. Findings indicate that Turkish form of FACES IV can be used as a valid and reliable scale with sufficient psychometric properties to evaluate the family cohesion, family adaptability, family communication, family satisfaction, and the family functioning as a whole in Turkish culture.

## 1. INTRODUCTION

The Circumplex Model of Marital and Family Systems is one of the most famous theoretical models investigating family functioning around the world (Gladding, 2011; Hamilton & Carr, 2016; Kouneski, 2002; Olson, 2011), and the three prominent dimensions that have significant impacts on the family functioning are named as (a) *cohesion*, (b) *flexibility*, and (c) *communication* in the Circumplex Model (Hamilton & Carr, 2016; Kouneski, 2002; Olson, Russell, & Sprenkle, 1983, 1989; Olson, Sprenkle, & Russell, 1979).

The cohesion is a major dimension affecting family functioning in a healthy manner (Barber & Buehler, 1996; Beavers & Hampson, 2000; Doherty & Hovander, 1990; Epstein, Baldwin, & Bishop, 1983; Gladding, 2011), and it defines the emotional bonding that family members have to each other (Olson et al., 1979, 1989). On the other hand, the extreme values in the cohesion dimension that named as *disengaged* / *disconnected*, or *enmeshed* / *overly connected* are accepted as unhealthy dynamics for the family functioning (Gladding, 2011; Hamilton & Carr, 2016; Olson & Gorall, 2003; Olson et al., 1989).

CONTACT: Turgut Turkdogan ✉ tturkdogan20@gmail.com Pamukkale University, Department of Psychological Counseling and Guidance, 20020 Denizli / TURKEY

ISSN-e: 2148-7456 /© IJATE 2018

The flexibility is another major dimension affecting family functioning in a healthy manner (Beavers & Hampson, 2000; Doherty & Hovander, 1990; Epstein et al., 1983; Gladding, 2011), and it defines the quality and expression of leadership and organization, role relationships, and relationship rules and negotiations in the family system (Olson & Gorall, 2006; Olson, 2011). On the other hand, the extreme values in the flexibility dimension that named as *rigid / inflexible*, or *chaotic / overly flexible* are accepted as unhealthy dynamics for the family functioning (Gladding, 2011; Hamilton & Carr, 2016; Olson & Gorall, 2003; Olson et al., 1989).

The communication is the third dimension affecting family functioning in a healthy manner (Beavers & Hampson, 2000; Epstein et al., 1983; Hamilton & Carr, 2016), and it essentially defines the positive communication skills used in family system, such as listening emphatically, paying attention to the subject being discussed, and expressing personal feelings sincerely (Olson & Gorall, 2003; Olson, 2000, 2011).

The main hypothesis of the Circumplex Model is that balanced level of adaptability and flexibility is the main source of the difference between the healthy and unhealthy functioning families across the family life cycle (Gladding, 2011; Hamilton & Carr, 2016; Kouneski, 2002; Olson et al., 1979, 1983, 1989). Also, another important hypothesis of the Circumplex Model is that families with a balanced level of adaptability and flexibility will tend to display more positive communication skills compared to families with extreme types of adaptability and flexibility (Hamilton & Carr, 2016; Kouneski, 2002; Olson et al., 1989).

Although the literature contains many assessment tools evaluating family functioning through similar dimensions [*Family Structure Assessment Device* (Gulerce, 2007), *Beavers Systems Model Self-Report Family Inventory* (Beavers & Hampson, 1990), *Family Assessment Measure III* (Skinner, Steinhauer, & Sitarenios, 2000), *Family Environment Scale* (Moos & Moos, 2009), *Family Relations Scale* (Tolan, Gorman-Smith, Huesmann, & Zelli, 1997), *McMaster Family Assessment Device* (Epstein et al., 1983), and *Systemic Clinical Outcome and Routine Evaluation* (Stratton, Bland, Janes, & Lask, 2010)], almost all studies which investigate the cross-cultural validity of the Circumplex Model have been conducted with the *Family Adaptability and Cohesion Scale* (FACES). Specifically, the curvilinear method that includes the combined assessment of the balanced and unbalanced dimensions of the family functioning is can be regarded as the distinctive feature of FACES IV.

Within this scope, FACES IV is a newly revised self-report scale assessing the levels of cohesion and flexibility dimensions that affect family functioning in the Circumplex Model (Gorall, Tiesel, & Olson, 2006; Olson, 2011; Olson, Gorall, & Tiesel, 2007). FACES IV Questionnaire Package includes six scales assessing two balanced dimensions (cohesion and flexibility), and four new unbalanced dimensions (disengaged, enmeshed, rigid, chaotic). Additionally, the package includes two other scales that assess the quality of the communication skills among family members (family communication) and the quality of functioning of the family system (family satisfaction). Although the third version of the scale (FACES III) has been adapted to the Turkish culture (Okman-Fisek, 1990), the previous versions of the FACES (I, II, and III) are able to assess the relationship between two balanced dimensions (cohesion and flexibility) and family satisfaction in a linear way (Olson, 2000). On the other hand, FACES IV is also able to assess the relationship between other unbalanced and extreme dimensions (disengaged, enmeshed, rigid, and chaotic) and family satisfaction in a curvilinear way (Gorall et al., 2006; Olson et al., 2007; Olson, 2011).

In summary, the Circumplex Model studies are conducted in more than 1200 studies in many countries around the world, and also supportive findings have been reached for the major hypotheses of the model (Kouneski, 2002; Olson, 2011). Moreover, the cultural adaptation studies on the last version of the FACES have already been completed in many countries such as Greek

(Koutra, Triliva, Roumeliotaki, Lionis, & Vgontzas, 2012), Hungary (Mirnics, Vargha, Toth, & Bagdy, 2010), Israel (Pirutinsky & Kor, 2013), Italy (Baiocco, Cacioppo, Laghi, & Tafa, 2013; Loriedo, Di Nuovo, & Visani, 2013), Portuguese (Pereira & Teixeira, 2013), Slovakia (Sebokova, Jurisova, Popelkova, Uhlarikova, & Zatkova, 2016), Spain (Rivero, Martinez-Pampliega, & Olson, 2010), and Uruguay (Costa-Ball, Gonzalez-Tornaria, delArca, Masjuan, & Olson, 2013). Correspondingly, more information about the cross-cultural validity of the model is tried to be reached by means of the data obtained from these adaptation studies. Also, an important reference for the future studies emphasizes the need for new researches, especially in collectivist Asian cultures (Kouneski, 2002; Mirnics et al., 2010).

Therefore, the lack of an assessment tool for investigating the major hypotheses of such a famous family approach is remarkably referring to an important gap in Turkish literature, and Turkish adaptation of the FACES IV may offer positive contributions to the literature that investigate the family functioning among Asian cultures with respect to the Circumplex Model. Additionally, Turkish adaptation of the FACES IV may offer positive contributions to the intervention studies aiming to increase the family functioning based on the Circumplex Model among Turkish families. Consequently, the aim of this study was to examine the psychometric properties of the Turkish version of Family Adaptability and Cohesion Scale IV Questionnaire Package (FACES IV), and also to achieve a valid and reliable assessment tool for the future investigations of the Circumplex Model of Marital and Family Systems in Turkish culture.

## 2. METHOD

### 2.1. Sample

The research population consists of 53,063 university students who were studying at different faculties of Pamukkale University in the fall semester of 2016-2017 academic year. The data were collected from 1613 (65.4% female, 34.6% male) university students chosen with random cluster sampling method, and the sample consists of Faculty of Education, Faculty of Health Sciences, Faculty of Science and Letters, Vocational High School of Child Care and Youth Services, and Vocational High School of Technical Sciences. In this sense, a representative number of a sample has been reached for the population with a 2.5% margin of error and 95% confidence level (Krejcie & Morgan, 1970). The mean age was 20.79 (*SD* = 3.70). Also, a sub-sample of 97 students chosen with convenience sampling method participated in the study of convergent validity. Besides, a sub-sample of 129 students chosen with convenience sampling method participated in the study of test-retest reliability within a three-weeks interval.

### 2.2. Instruments

***Family Adaptability and Cohesion Scale IV Questionnaire Package (FACES IV):*** FACES IV is an individual self-report scale assessing the levels of cohesion and flexibility dimensions that affecting family functioning in the Circumplex Model (Gorall et al., 2006; Olson et al., 2007; Olson, 2011). The three sub-scales that comprise the family cohesion are called as balanced cohesion (e.g., *Family members are supportive of each other during difficult times*), enmeshed (e.g., *We spend too much time together*), and disengaged (e.g., *Family members seem to avoid contact with each other when at home*). The other three sub-scales that comprise the family flexibility are called as balanced flexibility (e.g. *Our family tries new ways of dealing with problems*), rigid (e.g., *There are strict consequences for breaking the rules in our family*), and chaotic (e.g., *We never seem to get organized in our family*). Each of the sub-scales consists of seven items, and the items are evaluated with 5-point Likert scale. Cronbach-alpha coefficient was .89 for the balanced cohesion dimension, .87 for the disengaged dimension, and .77 for the enmeshed dimension. .84 for the balanced flexibility dimension, .82 for the rigid dimension, and .86 for the chaotic dimension in the original study (Olson, 2011). The measurement produces a circular ratio by dividing the balanced dimensions into extreme dimensions: *Cohesion*

*Ratio* = [*Balanced Cohesion*] / [(*Disengaged* + *Enmeshed*) / 2]; *Flexibility Ratio* = [*Balanced Flexibility*] / [(*Rigid* + *Chaotic*) / 2]; *Circumplex Total Ratio* = [(*Cohesion Ratio* + *Flexibility Ratio*) / 2]. Therefore, the higher level of circumplex total ratio refers more balanced family system according to this formulation (Olson, 2011).

Additionally, the quality of the communication skills between family members is measured with the family communication scale (e.g., *Family members can calmly discuss problems with each other*), and the quality of functioning of the family system is measured with the family satisfaction scale (e.g., *How satisfied are you with* **y***our family's ability to cope with stress?*). Each of these two scales consists of ten items, and the items are also evaluated with 5-point Likert scale. Cronbach-alpha coefficient was .90 for the family communication scale, and .92 for the family satisfaction scale (Olson et al., 2007). The high scores on these scales, reveal the quality of the communication and satisfaction in family system.

***The Multidimensional Scale of Perceived Social Support (MSPSS):*** The scale was developed by Zimet, Dahlem, Zimet & Farley (1988), and the Turkish adaptation of it was carried out by Eker, Arkar, &Yaldız (2001). The scale aims to measure the level of support that individuals perceive from social resources in their lives, and the items are evaluated with 7-point Likert scale. The perceived family support sub-scale was used for the convergent validity within the scope of this study, and Cronbach-alpha coefficient was reported as .85 for the family support in the original study (Zimet et al., 1988).

***Personal Information Form:*** This form includes the socio-demographic information of the participants such as gender, age, current marital status of the parents (married or divorced), educational level of the parents, income level of the family, number of children in family, and the family type [nuclear family, extended family, single-parent family, and parentless family (the siblings living together without their parents after divorce or any other reason, etc.)].

### 2.3. Procedure

After obtaining the legal and ethical permissions for the study, data were collected during the fall semester of the 2016-2017 academic year. Within the translation process, each item in the scale was first translated into Turkish by the three authors of the research individually. Then, three separate individual forms prepared by the researchers were put together to reach a common form that would express the Turkish translation of each item in a correct way. The language validity of the final form was examined by two lecturers who have doctoral degree in the field of English Language and Literature, and in the direction of the feedbacks, the Turkish form has been finalized. After that, a pilot study was carried out with the participants out of the current sample and it was concluded that scale items were adequately clear and understandable as a result of feedbacks from the participants.

### 2.4. Data Analyses

Within the scope of the adaptation study, confirmatory factor analysis (CFA) was used for examining the construct validity of the FACES IV, and maximum likelihood was used for the estimation method. Also, confirmatory factor analysis was performed via AMOS (Analysis of Moment Structures) statistical program. In confirmatory factor analysis, the items with low factor loadings below .30 become candidates for dropping. On the other hand, the items with low loadings may be retained at times to satisfy statistical identification requirements, or to meet the minimal number of items per factor (Hair et al., 2014). With respect to this, we firstly preferred to preserve the poorly performing items in the scale as long as they do not jeopardize the overall fit indices. However, we had to delete some items with low factor loadings when the overall fit indices could not be met. As a matter of fact, dropping one or two items from a

large battery of items can be tolerated for the construct validity, also the confirmatory test may not be jeopardized (Hair et al., 2014).

The internal consistency reliability of the scale was examined with Cronbach-alpha, average variance extracted (AVE), and composite reliability (CR) analyses. Pearson correlation values were used for the convergent validity. Also, t test and variance analyses were conducted to examine whether the families with disadvantaged socio-demographic characteristics can be distinguished by circumplex total ratio within the scope of criterion-related validity. Additionally, other descriptive statistics such as mean, standard deviation, and skewness were performed via SPSS (Statistical Package for the Social Sciences) statistical program.

## 3. FINDINGS

Preliminary assumptions (sample size, normality, linearity, multicollinearity, singularity, homoscedasticity, independence of the error terms) were checked and met before the analyses. Missing data under 5 percent for an individual case was ignored, and the missing values were imputed by the mean substitution method (Hair, Black, Babin, & Anderson, 2014; Tabachnick & Fidell, 2001). The descriptive statistics of the FACES IV scales are presented in Table 1. Pearson correlation coefficients between the FACES IV sub-scales are presented in Table 2.

**Table 1.** Descriptive Statistics of the FACES IV Scales (N = 1613)

| | *N* | Range | *M* | *SD* | Skewness | |
|---|---|---|---|---|---|---|
| | | | | | Statistic | *SE* |
| Cohesion Ratio | 1613 | .29 - 5.00 | 1.91 | .62 | .90 | .06 |
| Flexibility Ratio | 1613 | .24 - 5.00 | 2.29 | 1.01 | .49 | .06 |
| Circumplex Total Ratio | 1613 | .32 - 4.69 | 2.10 | .75 | .35 | .06 |
| Family communication | 1613 | 1.00 - 5.00 | 3.76 | .93 | -.78 | .06 |
| Family satisfaction | 1613 | 1.00 - 5.00 | 354 | .94 | -.61 | .06 |

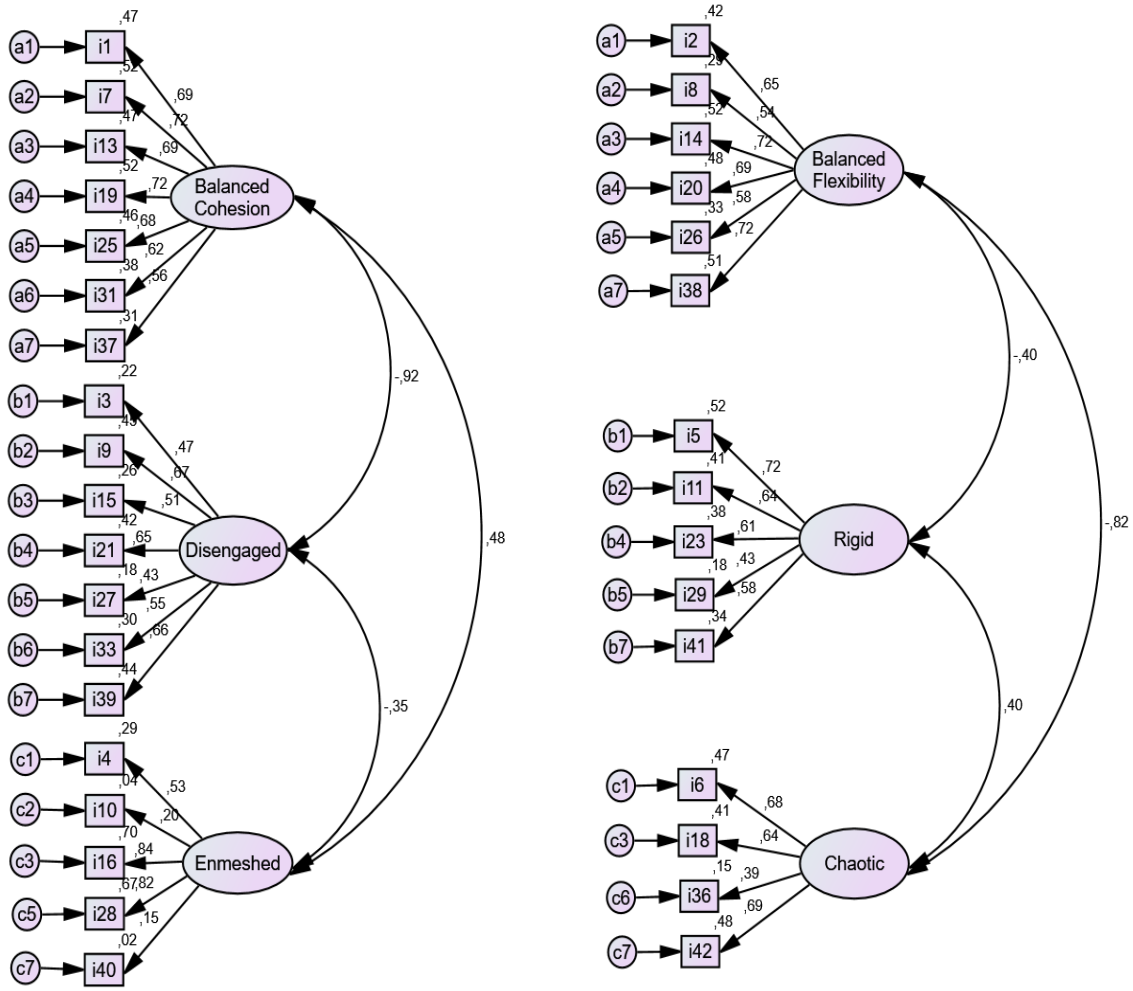
**Table 2.** Pearson Correlation Coefficients between the FACES IV Sub-Scales (N = 1613)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Balanced cohesion | - | | | | | | | | | |
| 2. Disengaged | -.72** | - | | | | | | | | |
| 3. Enmeshed | .31** | -.17** | - | | | | | | | |
| 4. Balanced flexibility | .79** | -.62** | .25** | - | | | | | | |
| 5. Rigid | -.27** | .34** | .26** | -.31** | - | | | | | |
| 6. Chaotic | -.65** | .69** | -.13** | -.60** | .33** | - | | | | |
| 7. Cohesion Ratio | .66** | -.74** | -.38** | .55** | -.43** | -.58** | - | | | |
| 8. Flexibility Ratio | .64** | -.65** | .01 | .75** | -.68** | -.75** | .65** | - | | |
| 9. Circumplex Total Ratio | .71** | -.75** | -.15** | .74** | -.64** | -.75** | .86** | .95** | - | |
| 10. Family communication | .74** | -.63** | .28** | .78** | -.31** | -.57** | .53** | .66** | .67** | - |
| 11. Family satisfaction | .72** | -.65** | .27** | .75** | -.32** | -.59** | .54** | .67** | .68** | .84** |

$p < .01$

### 3.1. The Construct Validity

The three-factor model about the family cohesion which includes balanced cohesion, enmeshed, and disengaged dimensions was examined with the confirmatory factor analysis, but the first analysis could not produce good fit indices [$\chi^2$ ($N$ = 1613) = 1231.12, $p$ < .0001; $\chi^2/df$ = 6.62; *GFI* = .93, *AGFI* = .91, *CFI*= .90, *RMSEA* = .60, *SRMR* = .078]. Therefore, after two items (item22 and item34) with very low factor loadings and insignificant explained variance values had been excluded from the enmeshed dimension, the construct validity for family cohesion was successfully confirmed [$\chi^2$ ($N$ = 1613) = 716.96, $p$ < .0001; $\chi^2/df$ = 4.81; *GFI* = .95, *AGFI* = .94, *CFI*= .94, *RMSEA* = .049, *SRMR* = .052]. Also, the results of the modified confirmatory analysis for the family cohesion are presented in Figure 1. Although the factor loadings of the item10 and item40 were low in the enmeshed dimension, these two items were kept in the scale because of the significant contributions to the explained variance and also having sufficient fit index of the scale (Hair et al., 2014).



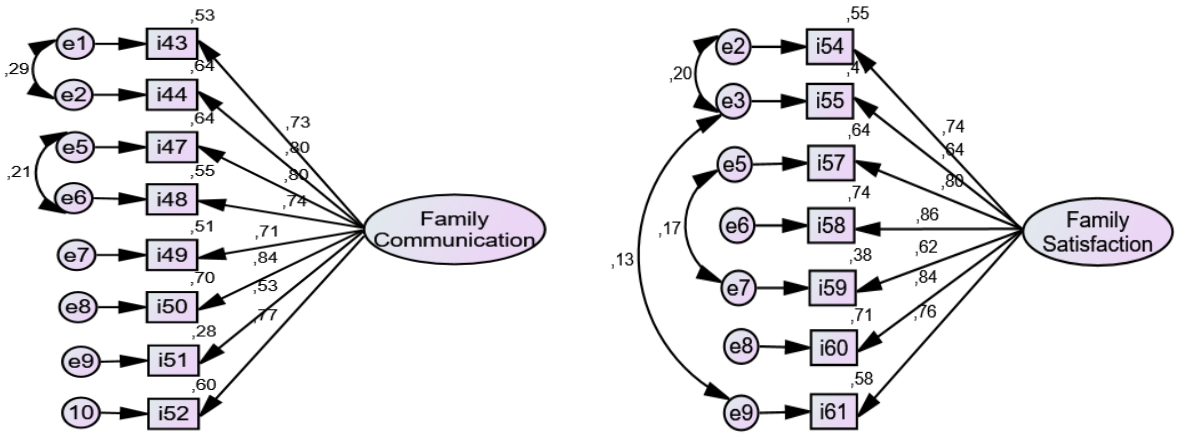

*Note (Family Cohesion):* $\chi^2$ ($N$ = 1613) =716.96, $p$ < .0001; $\chi^2/df$ = 4.81; *GFI* = .95, *AGFI* = .94 , *CFI*= .94, *RMSEA* = .049, *SRMR* = .052.

*Note (Family Flexibility):* $\chi^2$ ($N$= 1613)= 407.32, $p$ < .0001; $\chi^2/df$ = 4.68; *GFI* = .97, *AGFI* = .95 , *CFI*= .95, *RMSEA* = .048, *SRMR* = .042.

**Figure 1.** Modified confirmatory factor analysis results for family cohesion and family flexibility (Standardized solution; N = 1613).

The three-factor model about the family flexibility which includes balanced flexibility, rigid, and chaotic dimensions was examined with the confirmatory factor analysis, but the first analysis could not produce acceptable fit indices [$\chi^2$ ($N$ = 1613) = 2049.98, $p$ < .0001; $\chi^2/df$ = 11.02; *GFI* = .88, *AGFI* = .85, *CFI*= .80, *RMSEA* = .80, *SRMR* = .092]. After one item (item32) in the balanced flexibility dimension, two items (item17 and item35) in the rigid dimension, and three items (item12, item24, and item 30) in the chaotic dimension had been excluded because of the low factor loadings and their tendency to factor under other dimensions beside the factor they belong to, the construct validity for family flexibility was successfully confirmed [$\chi^2$ ($N$ = 1613) = 407.32, $p$ < .0001; $\chi^2/df$ = 4.68; *GFI* = .97, *AGFI* = .95 , *CFI*= .95, *RMSEA* = .048, *SRMR* = .042]. The results of the modified confirmatory analysis for the family flexibility are also presented in Figure 1.

One-factor model about the family communication was examined with the confirmatory factor analysis, but the first analysis could not produce good fit indices [$\chi^2$ ($N$ = 1613) = 437.99, $p$ < .0001; $\chi^2/df$ = 12.51; *GFI* = .95, *AGFI* = .92, *CFI*= .96, *RMSEA* = .85, *SRMR* = .030]. After two items (item45 and item46) had been excluded because of the low factor loadings and their tendency to factor under another independent dimension besides the one-factor model, the construct validity for family communication was successfully confirmed [$\chi^2$ ($N$ = 1613) = 63.5, $p$ < .0001; $\chi^2/df$ = 3.53; *GFI* = .99, *AGFI* = .98, *CFI* = .99, *RMSEA* = .040, *SRMR* = .015]. The results of the modified confirmatory analysis for the family communication are presented in Figure 2.




*Note (Family Communication):* $\chi^2$ ($N$ = 1613) = 63.57, $p$ < .0001; $\chi^2/df$ = 3.53; *GFI* = .99, *AGFI* = .98, *CFI* = .99, *RMSEA* = .040, *SRMR* = .015.

*Note (Family Satisfaction):* $\chi^2$ ($N$ = 1613) = 53.90, $p$ < .0001; $\chi^2/df$ = 4.90; *GFI* = .99, *AGFI* = .98, *CFI* = .99, *RMSEA* = .049, *SRMR* = .012.

**Figure 2.** Modified confirmatory factor analysis results for family communication and family satisfaction (Standardized solution; N = 1613).

One-factor model about the family satisfaction was examined with the confirmatory factor analysis, but the first analysis could not produce good fit indices [$\chi^2$ ($N$ = 1613) = 805.47, $p$ < .0001; $\chi^2/df$ = 23.01; *GFI* = .90, *AGFI* = .84, *CFI*= .93, *RMSEA* = .12, *SRMR* = .042]. After three items (item53, item56, and item62) had been excluded because of the low factor loadings and their tendency to factor under another independent dimension besides the one-factor model, the construct validity for family satisfaction was successfully confirmed [$\chi^2$ ($N$ = 1613) = 53.90, $p$ < .0001; $\chi^2/df$ = 4.90; *GFI* = .99, *AGFI* = .98, *CFI* = .99, *RMSEA* = .049, *SRMR* = .012]. The results of the modified confirmatory analysis for the family satisfaction are also presented in Figure 2.

### 3.2. The Convergent Validity

Convergent validity of the FACES IV was examined with the Perceived Social Support Scale among a sub-sample of 97 students chosen with convenience sampling method. According to the findings, positive and strong correlation values were reached between perceived family support and balanced cohesion ($r_{(n=97)} = .74$, $p < .001$), and also balanced flexibility ($r_{(n=97)} = .70$, $p < .001$) dimensions which are defined as the healthy dimensions of the family structure. Moreover, negative correlation values were reached between the level of perceived family support and the disengaged ($r_{(n=97)} = -.69$, $p < .001$), rigid ($r_{(n=97)} = -.28$, $p < .01$), and chaotic ($r_{(n=97)} = -.62$, $p < .001$) dimensions which are defined as the unhealthy dimensions of the family structure. On the other hand, the positive relationship between perceived family support and enmeshed dimension was not significant ($r_{(n=97)} = .11$, $p > .05$). Also, positive and strong correlation values were found between the perceived family support and the total scores about the cohesion ratio ($r_{(n=97)} = .59$, $p < .001$), flexibility ratio ($r_{(n=97)} = .59$, $p < .001$), and also circumplex total ratio ($r_{(n=97)} = .63$, $p < .001$). Moreover, positive and strong correlation values were found between perceived family support and family communication ($r_{(n=97)} = .65$, $p < .001$), and also family satisfaction ($r_{(n=97)} = .65$, $p < .001$).

### 3.3. The Criterion-Related Validity

Marital status of the parents, family type, parents' education level and reading ability, number of children in the family, also income level and perceived economic stress are the remarkable socio-demographic factors affecting family system in the literature (Eccles & Harold, 1996; Klein & Forehand, 2000; Mikolajczak, Raes, Avalosse, & Roskam, 2018; Trent & South, 1992). Therefore, t test and variance analyses were conducted to examine whether the families with disadvantaged socio-demographic characteristics can be distinguished by circumplex total ratio within the scope of criterion-related validity.

The important finding supporting the criterion-related validity was that circumplex total ratio which indicates the general balance of the family system was significantly lower in the families with disadvantaged socio-demographic characteristics. Also the t test and variance analyses results about the circumplex total ratio and socio-demographics of the family are presented in Table 3. As a matter of fact, circumplex total ratio was significantly lower among the participants from the divorced families ($t_{(1611)} = 5.40$, $p < .001$). In addition, circumplex total ratio of the participants who indicated their current family type as parentless-family (the siblings living together without their parents after divorce or any other reason, etc.) was significantly lower than of all groups with the other family types ($F_{(3\text{-}1609)} = 9.03$, $p < .001$).

Moreover, circumplex total ratio was significantly lower among the participants who have illiterate mothers ($F_{(6\text{-}1606)} = 4.77$, $p < .001$). However, circumplex total ratio was significantly higher among the participants who have university graduate fathers ($F_{(7\text{-}1605)} = 4.57$, $p < .001$). Also, circumplex total ratio of the participants from families with five or more children is significantly lower than of all groups which have a less number of children ($F_{(4\text{-}1608)} = 4.84$, $p < .001$). Finally, circumplex total ratio of the participants from families that income level is lower than a minimum wage is significantly lower than of all other groups which the family income level is above a minimum wage ($F_{(3\text{-}1609)} = 7.76$, $p < .001$).

**Table 3.** The T Test and Variance Analyses Results about the Circumplex Total Ratio and Socio-Demographics of the Family (N =1613)

| | Circumplex Total Ratio | *N* | *M* | *SD* | *t / F* | *p* |
|---|---|---|---|---|---|---|
| Marital status of the parents | Married | 1524 | 2.12 | .74 | 5.40*** | .000 |
| | Divorced | 89 | 1.69 | .79 | | |
| Family type | Nuclear family | 1366 | 2.12 | .74 | 9.03*** | .000 |
| | Extended family | 133 | 2.09 | .75 | | |
| | Single-parent family | 83 | 1.98 | .83 | | |
| | Parentless family | 31 | 1.45 | .65 | | |
| Educational level of the mother | Illiterate | 77 | 1.79 | .65 | 4.77*** | .000 |
| | Literate | 66 | 1.84 | .69 | | |
| | Primary school | 865 | 2.10 | .74 | | |
| | Middle school | 245 | 2.22 | .75 | | |
| | High school | 270 | 2.13 | .75 | | |
| | College | 25 | 2.07 | .77 | | |
| | University | 65 | 2.13 | .81 | | |
| | Master/ Doctorate | - | - | - | | |
| Educational level of the father | Illiterate | 7 | 1.63 | .52 | 4.57*** | .000 |
| | Literate | 34 | 1.77 | .70 | | |
| | Primary school | 620 | 2.04 | .71 | | |
| | Middle school | 307 | 2.13 | .75 | | |
| | High school | 386 | 2.10 | .72 | | |
| | College | 57 | 2.08 | .81 | | |
| | University | 195 | 2.31 | .83 | | |
| | Master/ Doctorate | 7 | 2.42 | .81 | | |
| Number of children in family | Only child | 83 | 2.09 | .73 | 4.84** | .001 |
| | Two children | 772 | 2.16 | .77 | | |
| | Three children | 454 | 2.08 | .73 | | |
| | Four Children | 164 | 2.06 | .76 | | |
| | Five or more children | 140 | 1.88 | .61 | | |
| Income level of the family | Under a minimum wage | 435 | 1.96 | .73 | 7.76*** | .000 |
| | Between one or two minimum wages | 802 | 2.14 | .75 | | |
| | Between two or three minimum wages | 276 | 2.19 | .75 | | |
| | Three or more minimum wages | 100 | 2.17 | .75 | | |

### 3.4. The Internal Reliability

The internal consistency reliability of the scale was examined with Cronbach-alpha, average variance extracted, and composite reliability analyses. Also, the values are presented in Table 4. The Cronbach-alpha coefficients ranged between .65 and .91, the average variance extracted coefficients ranged between .33 and .57, and the composite reliability coefficients ranged between .66 and .90 for the FACES IV sub-scales. Actually, the values with .60 to .70 are deemed the lower limit of acceptability for the Cronbach-alpha and composite reliability analyses, and the lower limit of the average variance extracted is deemed as .50 (Hair, et al. 2014).

**Table 4.** Cronbach-alpha (α), Average Variance Extracted (AVE), and Composite Reliability Coefficients for the Internal Reliability

| | *α* | *AVE* | *CR* |
|---|---|---|---|
| Balanced cohesion | .85 | .45 | .85 |
| Disengaged | .77 | .33 | .77 |
| Enmeshed | .65 | .34 | .66 |
| Balanced flexibility | .82 | .43 | .82 |
| Rigid | .73 | .36 | .74 |
| Chaotic | .69 | .38 | .70 |
| Family communication | .91 | .48 | .89 |
| Family satisfaction | .91 | .57 | .90 |

Accordingly, the coefficients for the Cronbach-alpha and composite reliability were emerged at an acceptable level in this study, but coefficients for the average variance extracted were emerged at a low level. On the other hand, in such cases that the average variance extracted is less than .50, but the composite reliability is higher than .60 the convergent validity of the construct may still be deemed adequate (Fornell & Larcker, 1981).

### 3.5. The Test-Retest Reliability

The stability coefficient of the scale was examined with test-retest method within a three-weeks interval among a sub-sample of 129 students chosen with convenience sampling method. The test-retest coefficient was .85 for the balanced cohesion dimension, .81 for the disengaged dimension, and .71 for the enmeshed dimension. Also, test-retest coefficient was .81 for the balanced flexibility dimension, .76 for the rigid dimension, and .74 for the chaotic dimension. Finally, the test-retest coefficient was .84 for the family communication, and .86 for the family satisfaction. To sum up, the test-retest coefficients of the sub-scales ranged between .71 and 86 for the FACES IV sub-scales.

## 4. DISCUSSION AND CONCLUSION

The current study aimed to examine the psychometric properties of the Turkish version of Family Adaptability and Cohesion Scale IV Questionnaire Package, and also to achieve a valid and reliable assessment tool for the future investigations of the Circumplex Model of Marital and Family Systems in Turkish culture. Also, an important encouragement for the study was the suggestion that highlights the need for new studies investigating the major hypotheses of the Circumplex Model in collectivist Asian cultures (Kouneski, 2002; Mirnics et al., 2010).

The adaptation study of the scale was completed with satisfactory psychometric properties. Specifically, it can be said that a very rigorous examination has been carried out for the construct validity of the scale. Although we tried to preserve the original form of the scale as a whole, the confirmatory factor analyses produced serious modification suggestions. On the other hand, leaving some items out of assessment because of their low factor loading or their tendency to factor under other dimensions besides the factor they expected to fit is a tolerable modification which is commonly observed in other adaptation studies of the FACES IV (Baiocco et al., 2013; Koutra et al., 2012; Mirnics et al., 2010; Pirutinsky & Kor, 2013; Rivero, Martinez-Pampliega, & Olson, 2010). Eventually, we have preserved the original form of the scale with all items for the future research in order not to ignore the suggestion of the developers of the scale in that way (Gorall et al., 2006; Olson et al., 2007; Olson, 2011).

Supportive findings have been reached for the convergent validity of the scale in regard to the correlations between the perceived family support scale and the sub-scales of FACES IV. The direction of the correlations was positive for the balanced dimensions of the cohesion and flexibility scales whereas the direction of the correlations was negative for the unbalanced dimensions such as disengaged, rigid, and chaotic. However, there was not a negative correlation

between the enmeshed dimension and perceived family support, although the enmeshed dimension was conceptualized as an unbalanced dimension in the Circumplex Model. In other words, the enmeshed dimension was not perceived as an unhealthy dimension in regard to the perceived family support in Turkish culture. As a matter of fact, the enmeshed dimension may be perceived as a facilitating dimension which positively contributes to family functioning in the cultures that family togetherness is strongly emphasized (Kouneski, 2002).

Within the scope of criterion-related validity, the scale has successfully discriminated the families with disadvantaged socio-demographic characteristics in regard to the circumplex total ratio. According to these findings, being a member of the divorced or the parentless-family, having an illiterate mother, having five or more siblings in the family, and having a family income lower than a minimum wage were significant risk factors for the balance of the family system. Moreover, having a university graduate father was significantly a protective factor.

The internal reliability coefficients of the Cronbach-alpha and composite reliability appeared with advanced values, with the exception of enmeshed and chaotic dimensions. As a matter of fact, there are many research findings that reporting the similar results about the enmeshed and chaotic dimensions in the literature (Koutra et al., 2012; Marsac & Alderfer, 2011; Olson, 2011; Pereira & Teixeira, 2013). Also, the Cronbach-alpha and the composite reliability coefficients of the enmeshed and chaotic dimensions were emerged at an acceptable level in this study. However, the average variance extracted values were emerged at low levels for the other sub-scales. At this juncture, Fornell and Larcker (1981) who are the developers of the average variance extracted actually described AVE coefficient as a conservative measure, and they recommended that the composite reliability may be deemed adequate for the construct validity of the scale. Finally, the test-retest findings within in a three-weeks interval successfully supported the stability reliability for all sub-scales.

Within the scope of the current study, a valid and reliable scale was obtained with sufficient psychometric properties to evaluate the family cohesion, family adaptability, family communication, family satisfaction, and the family functioning as a whole in Turkish culture. It is also expected that this study will enable further researches to generate positive contributions to the literature that investigating the cross-cultural validity of the Circumplex Model of Marital and Family Systems. Additionally, it is expected that this study will enable further intervention studies aiming to increase the family functioning based on the Circumplex Model among Turkish families.

**Acknowledge**

The summary of the study was presented with the title of "Turkish adaptation of Family Adaptability and Cohesion Scale IV: Validity and reliability study" at IV$^{th}$ International Eurasian Educational Research Congress, 11-14 May 2017, Denizli-Turkey.

**ORCID**

Turgut Turkdogan https://orcid.org/0000-0001-5440-341X

## 5. REFERENCES

Baiocco, R., Cacioppo, M., Laghi, F., & Tafa, M. (2013). Factorial and construct validity of FACES IV among Italian adolescents. *Journal of Child & Family Studies*, *22*, 962-970.

Barber, B. K., & Buehler, C. (1996). Family cohesion and enmeshment: Different constructs, different effects. *Journal of Marriage and the Family*, *58*, 433-441.

Beavers, R., & Hampson, B. (1990). *Successful families: Assessment and intervention*. New York: W. W. Norton.

Beavers, R., & Hampson, B. (2000). The Beavers Systems Model of family functioning. *Journal of Family Therapy, 22*, 128-143.

Costa-Ball, D., Gonzalez-Tornaria, M., delArca, D., Masjuan, N., & Olson, D. H. (2013). Psychometrıc propertıes of FACES IV: Validity study in Uruguayan population. [Abstract]. *Ciencias Psicologicas, 7*(2), 119-132.

Doherty, W. J., & Hovander, D. (1990). Why don't family measures of cohesion and control behave the way they're supposed to? *American Journal of Family Therapy, 18*, 5-18.

Eccles, J. S., & Harold, R. D. (1996). Family involvement in children's and adolescents' schooling. In A. Booth & J. F. Dunn (Eds.), *Family school links: How do they affect educational outcomes?* (pp. 3-34). Mahwah, NJ: Erlbaum.

Eker, D. Arkar, H., & Yaldız, H. (2001). Factorial structure, validity, and reliability of revised form of the multidimensional scale of perceived social support. *Türk Psikiyatri Dergisi, 12*(1), 17-25.

Epstein, N., Baldwin, L., & Bishop, D. (1983). The McMaster family assessment device. *Journal of Marital and Family Therapy*, *9*(2), 171-180.

Epstein, N., Baldwin, L., & Bishop, D. (1983). The McMaster family assessment device. *Journal of Marital and Family Therapy*, *9*(2), 171-180.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Gladding, S. T. (2011). *Family therapy: History, theory and practice*. (5th ed). Upper Saddle River, New Jersey: Pearson Education.

Gorall, T. M., Tiesel, J., & Olson, D. H. (2006). *FACES IV: Development and validation*. Minneapolis, MN: Life Innovations.

Gulerce, A. (2007). *The transformational family model and the psychological patterns of families in Turkey*. Istanbul: Bogazici University Library Cataloguing.

Hamilton, E., & Carr, A. (2016). Systematic review of self-report family assessment measures. *Family Process*, *55*(1), 16-30.

Klein, K., & Forehand, R. (2000). Family processes as resources for African-American children exposed to a constellation of sociodemographic risk factors. Journal of Clinical Child Psychology, *29*(1), 53-65.

Kouneski, E. (2002). *Circumplex model and FACES: Review of literature*. Twin Cities, MN: University of Minnesota.

Koutra, K., Triliva, S., Roumeliotaki, T., Lionis, C., & Vgontzas, A. N. (2012). Cross-cultural adaptation and validation of the Greek version of the family adaptability and cohesion evaluation scales IV package (FACES IV Package). *Journal of Family Issues*, *34*(12), 1647-1672.

Krejcie, R.V., & Morgan, D.W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30*, 607-610. doi: 10.1177/001316447003000308

Loriedo, C., Di Nuovo, S. F., & Visani, E. (2013). FACES-IV reliability and validity in an adult Italian sample. Retrieved from http://www.buildingrelationships.com/facesiv_studies/italy_facesiv_2013.pdf

Marsac, M. L., & Alderfer, M. A. (2011). Psychometric properties of the FACES-IV in a pediatric oncology population. *Journal of Pediatric Psychology*, *36*(5), 528-538.

Mikolajczak, M., Raes, M. E., Avalosse, H., & Roskam, I. (2018). Exhausted parents: Sociodemographic, child-related, parent-related, parenting and family-functioning correlates of parental burnout. *Journal of Child and Family Studies*, 27(2), 602-614.

Mirnics, Z., Vargha, A., Toth, M., & Bagdy, E. (2010). Cross-cultural applicability of FACES IV. *Journal of Family Psychotherapy*, *21*(1), 17-33.

Moos, R., & Moos, B. (2009). *Family environment scale manual* (4th ed.). Palo Alto, CA: Consulting Psychology Press.

Okman-Fisek, G. (1990). *Family interconnectedness*. Graduate seminar. Bogazici University, Department of Educational Sciences, Istanbul.

Olson D. H., & Gorall D. M. (2006). FACES IV and the circumplex model. Retrieved from https://www.societyofpediatricpsychology.org/sites/default/files/files/3_innovations.pdf

Olson D. H., Gorall D. M., Tiesel J. W. (2007). *FACES IV Manual*. Minneapolis, MN: Life Innovations.

Olson, D. H. (2000). Circumplex model of marital and family systems. *Journal of Family Therapy*, *22*(2), 144-167.

Olson, D. H. (2011). FACES IV and the circumplex model: Validation study. *Journal of Marital and Family Therapy*, *7*(1), 64-80.

Olson, D. H., McCubbin, H. I., Barnes, H., Larsen, A., Muxen, M., & Wilson, M. (1989). *Families*: *What makes them work* (2nd ed.). Newbury Park, CA: Sage Publications.

Olson, D. H., Russell, C. S., & Sprenkle, D. H. (1983). Circumplex model of martial and family systems: VI. Theoretical update. *Family Process*, *22*, 69-83.

Olson, D. H., Russell, C. S., & Sprenkle, D. H. (1989). *Circumplex Model*: *Systemic assessment and treatment of families*. New York, NY: Haworth Press.

Olson, D.H., & Gorall, D. M. (2003). Circumplex model of marital and family systems. In F. Walsh (Ed.) *Normal family processes*: *Growing diversity and complexity* (3rd ed., pp. 514-548). New York, NY: Guilford Press.

Olson, D.H., Sprenkle, D.H., & Russell, C. S. (1979). Circumplex model of marital and family system: I. Cohesion and adaptability dimensions, family types, and clinical applications. *Family Process*, *18*(1), 3-28.

Pereira, M. G., & Teixeira, R. J. (2013). Portuguese validation of FACES-IV in adult children caregivers facing parental cancer. *Contemporary Family Therapy*: *An International Journal*, *35*, 478-490.

Pirutinsky, S., & Kor, A. (2013). Relevance of the circumplex model to family functioning among Orthodox Jews in Israel. *The New School Psychology Bulletin*, *10*(2), 25-38.

Rivero, N., Martinez-Pampliega, A., & Olson, D. H. (2010). Spanish adaptation of the FACES IV questionnaire: Psychometric characteristics. *The Family Journal*, *18*(3), 288-296.

Sebokova, G., Jurisova, E., Popelkova, M., Uhlarikova, J., & Zatkova, M. (2016). Factorial validity and reliability of the Slovak version of the F ACES IV in adolescents. [Abstract]. *Ceskoslovenska Psychologie*, *60*(3), 278-289.

Skinner, H., Steinhauer, P., & Sitarenios, G. (2000). Family assessment measure (FAM) and process model of family functioning. *Journal of Family Functioning*, *22*, 190-210.

Stratton, P., Bland, J., Janes, E., & Lask, J. (2010). Developing an indicator of family function and a practicable outcome measure for systemic family and couple therapy: The SCORE. *Journal of Family Therapy*, *32*, 232-258.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn & Bacon.

Tolan, P., Gorman-Smith, D., Huesmann, R., & Zelli, A. (1997). Assessment of family relationship characteristics: A measure to explain risk for antisocial behaviour and depression among urban youth. *Psychological Assessment*, *9*(3), 212-223.

Trent, K., & South, S. J. (1992). Sociodemographic status, parental background, childhood family structure, and attitudes toward family formation. *Journal of Marriage and Family, 54*(2), 427-439.

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. *Journal of Personality Assessment, 52*(1), 30-41.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 645–658

DOI: 10.21449/ijate.463409

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# The Effect of Teaching quality and teaching practices on PISA 2012 Mathematics Achievement of Turkish Students

**Safiye Bilican Demir** [1,*]

[1] Koceli University, Faculty of Education, Department of Educational Science, Kocaeli, Turkey

**ARTICLE HISTORY**

*Received: 19 April 2018*
*Revised: 16 September 2018*
*Accepted: 20 September 2018*

**KEYWORDS**

PISA, Mathematical literacy,
Teaching quality,
Teaching practices,
Hierarchical Linear Modeling

**Abstract:** The purpose of this research is to explore the relationship mathematics teaching qualification and teaching activities with mathematics scores of 8th grade Turkish students who participated in PISA (Program for International Student Assessment) 2012 using Hierarchical Linear Modeling. PISA 2012 mathematics literacy test and student questionnaire were used to collect data and these were obtained from the official website of PISA. In this study, random effects one-way ANOVA and random coefficient regression model were developed to analyze the data. The results of the research show that there are significant differences in terms of mathematics literacy between schools. It also indicates that while the increase in the student-centered instructions within the context of classroom activities decreased students' mathematics literacy scores, teachers' frequency of activating cognitive processes and a good disciplined environment increased students' mathematics literacy skill. The impacts of significant variables were discussed and implications for further research were provided.

## 1. INTRODUCTION

Education policies that aim to raise students who can make independent decisions, develop critical thinking and analytical skills are becoming more important. Therefore, countries participate in international measurement practices where students' achievements in different fields can be assessed (Anderson, Lin, Treagust, Ross & Yore, 2007). PISA (Program for International Student Assessment) is one of those surveys, which is administered by the OECD for 15-year-old students. PISA application; is a useful tool that can be used to improve quality in education, equality and productivity and is a good predictor of student success, explaining some common features of students, schools and education systems (Schleicher, 2007). Therefore, the examination of PISA results is of great importance in terms of education systems. Since 2000, every three years fifteen-year-old students participate in PISA. Students are

CONTACT: Safiye Bilican Demir ✉ safiyebilican@gmail.com Koceli University, Faculty of Education, Department of Educational Science, Kocaeli, Turkey

**ISSN-e: 2148-7456 /© IJATE 2018**

selected randomly from schools and take tests in three main subjects: Reading, mathematics, and science literacy. For each assessment, one of reading, mathematics and science is chosen as the major domain and given greater emphasis. The remaining two areas, the minor domains, are assessed less thoroughly and the PISA 2012 focus was on mathematics literacy (OECD, 2014a).

In parallel with the changing understanding of education systems, PISA's assessment is forward-looking; rather than focusing on students' specific school curriculum, it assesses students' ability to use their knowledge and skills in real-life challenges. This reflects a change in curricular goals and objectives, which are increasingly concerned with what students can do with what they learn at school. For example, PISA 2012 survey is concerned more about how learners can apply their mathematics knowledge in real-life situations, rather than identifying what they know by using formulas or calculations. In this regard, the structure measured in PISA survey is defined as "mathematical literacy".

Mathematical literacy includes the ability to deal with mathematics and to understand how mathematics is firmly embedded in individual's private, academic and social life (OECD, 2014a). Similarly, McCrone and Dossey (2007) have defined mathematical literacy as being able to understand the role of mathematics in the world, make solid judgements and use mathematics in order to address the needs of life. These definitions locate mathematical literacy not only in the academic-school life of the individual, but also in the daily and business life. In accordance with these definitions, PISA is expected to use mathematical skills in situations where students may encounter real-life situations (e.g. shopping, travel, financial calculations or political problems, etc.) other than typical mathematical problems which they come across at school. Students need to use their mathematical skills in solving these problems. Mathematical skills require higher order thinking skills such as inference and modeling of problem and problem solving. So, PISA reports student performance not just as numerical scores, but also in terms of skills, by describing what students who achieve a given level on a PISA scale typically know and can do. (OECD, 2014a).

Raising individuals with mathematics literacy becomes more important as occupations that require abstract thinking skills; graph-reading or making prediction skills in daily life and importance of economic-monetary data increase. International surveys predict an increase of almost 1% in annual Gross Domestic Product (GDP) growth per capita with half a standard deviation's increase in individual mathematics and science performance (OECD, 2010). Also, mathematical skills are associated with socio-economic well-being. For example, longitudinal research in the UK suggests that people with poor mathematical skills are more than twice as likely as those with better skills to be represented at the lowest level of employment and are at increased risk of poor mental and physical health (Bynner & Parsons, 2005). So, the aim of mathematics teaching programs is to raise mathematical literate individuals (MEB, 2005).

In addition to students' mathematical literacy, PISA collects data from students and school administrators in the context of possible affective variables (school-class environment, classroom teaching activities, curricula, etc.) that may affect students' cognitive skills. When the relevant literature is examined, it is seen that there are many studies on the student, teacher and school related factors affecting the mathematical literacy of 15 years old Turkish students. In these studies, variables such as region and school type, time to learn mathematics, gender, socio-economic level, attitudes and class-school characteristics showed significant relationships with mathematical literacy (e.g. Aksu, Güzeller, & Eser, 2017; Akyüz & Pala, 2010; Berberoğlu & Kalender, 2005; İlgün-Dibek, 2015; Usta, 2014). As can be seen, the variables that may influence the mathematical literacy of students in PISA are addressed in a very broad context. Because the nature of the teaching and learning process has an increasing emphasis on student achievement and quality of education, the relevance of factors related to

this process to mathematical literacy has been addressed in this study. Many studies have shown that the mathematics achievement of students is influenced by the learning-teaching environment (e.g. Akyüz & Berberoğlu, 2010; Bloom, 1976; Eccles &Roeser, 2011; Hill & Rowe, 1998; Lamb & Fullarton 2002; Wentzel et al., 2010). Characteristics of the teachers who carry out teaching in schools have an important and inevitable influence on student achievement. Therefore, teachers are the basic components of education systems. In this case, it is not possible to ignore the quality of teachers when considering the learning of students and how schools can be improved (Blanton et al., 2002; Wayne & Youngs, 2003). Deal (2010) and Gallagher (2002) indicate that teachers have a high impact on student achievement when differences in student characteristics are controlled. In recent years, studies on teacher qualifications have focused on teachers' behaviors (Darling - Hammond, 2000; Wenglinsky, 2000). Accordingly, in PISA 2012, data were gathered from students about teachers' classroom practices within the context of mathematical teaching activities and teaching quality.

It is inevitable that the practices of teachers and their use of different materials or tools for a certain purpose are the main determinants of student performance. There are studies in the literature on the relationship between students' mathematical achievements and teaching activities. For example, the study by Davis-Langston (2012) examined the relationship between different style of teaching and mathematics success. The results of the research show that there is a meaningful and positive relationship between the teaching style in which that the teacher is the role model and the students are guided and the mathematics success of the students; on the other hand, there is a negative relationship between the teaching style in which the teacher conveys only knowledge as an expert and the mathematical success. Kirkpatrick (2002) showed that teaching activities, such as using materials in class, asking for problems with more than one solution, encouraging students to discuss groups, linking topics, etc., influenced students' mathematics success positively. In the longitudinal study of Palardy and Rumberger (2008) the students were followed from the pre-school period to the end of the fifth grade and the teaching activities revealed that the mathematics achievement of the students were affected.

Another variable discussed in this study is teaching quality. In this context, what kind of problems has emerged in the classroom during mathematics teaching, the ways teachers follow, and emotional and social support and cognitive activation are discussed. Studies of the relationship of these variables to mathematical achievement showed that a positive, non-disordered classroom environment improves mathematics achievement for students (e.g. Akyüz & Satıcı 2013; İlgün-Dibek, 2015; Sortkær & Reimer, 2016; Usta, 2014). Tennant and colleagues (2015) suggested that students tend to have better math achievement when teachers focus on helping and guiding them in understanding mathematical concepts and solving questions. As students spend much of their time with their teachers in school, teacher support can be vital to students' academic development. Emotional support domain is defined as teacher-student interactions that promote social connection and cohesion, convey concern for students' feelings and interest in their individuality, and honor students' desire to learn meaningful material and have a say in their learning (Ruzek et al., 2016). Students who are cognitively active and who are involved in the learning-teaching process are more willing to learn and have a positive effect on their success (Archambault, Janosz & Chouinard, 2012; Baumert & Kunter, 2013; Davis-Langston, 2012; Kirkpatrick, 2002; Palardy & Rumberger, 2008).

For many countries, student performance is an important indicator of the quality of education systems. In addition to evaluating the education systems as a whole, assessments are being carried out to determine the success of students at the national and international scale. In this context, international assessments such as PISA, and the performance of students are significant in many countries. Turkey, since 2003, regularly participates in the PISA application. PISA

results showed that 8th grade Turkish student performance in mathematics was below the international evaluation criteria and the average mathematics literacy level of students in OECD countries, while the third level, the average mathematics literacy level of students in the sample Turkey is second (OECD, 2014b). Similarly, in PISA 2003 where the focus was on mathematics, 75% of the Turkish students were in the second skill level and below in terms of mathematical literacy performance (EARGED, 2008). The extremely low percentage of students in the upper levels of proficiency samples in Turkey is also a major problem. Since the implementation of PISA 2003, there has been a slight increase in mathematics literacy scores of Turkish students. However, this increase did not affect much on Turkey's place in the rankings (OECD, 2004; OECD, 2014b). This situation necessitates identifying the reasons of the difficulties of Turkish students in terms of mathematical cognitive domain and the variables influencing it. Identifying the factors that influence the development of this skill will provide valuable information to teachers about how to improve this performance. Furthermore, taking opinions of students will provide data in developing teaching programs according to student expectations and in forming appropriate education environments. It would be appropriate to identify teacher behaviors that lead to behavioral changes in teachers' education policies, curricula, or teacher training programs. The aim of this research is to explore the relationship mathematics teaching qualification and teaching activities with mathematics scores of 8th grade Turkish students who participated in PISA 2012 using Hierarchical Linear Modeling.

## 2. METHOD

### 2.1. Research design

The research investigates the performance of the Turkish students who participated in PISA 2012 in the field of mathematics literacy. The studies based on hierarchical linear modeling aim to explore to what extent variables of interpenetrating structures predict other variables. For this reason, the research employs relational survey model.

### 2.2. Sample

The target population was 4848 15-year-old students who participated in PISA 2012 from Turkey. 49% of the students are female and 51% are male. The students were selected using two-stage stratified sampling. First, 170 schools were identified and then students were randomly selected from these schools (OECD, 2014a).

### 2.3. Instruments

PISA 2012 mathematics literacy test and student questionnaire were used to collect data and these were obtained from the official website of PISA. The mathematical cognitive test pays attention to the students' ability to analyze mathematical problems in different situations, to interpret the knowledge, to understand cause-effect relationships, and to find relationships between ideas.

Students participating in PISA applications have not responded to all questions of the mathematics literacy test. There are several possible alternative approaches for making this inference. PISA uses the imputation methodology usually referred to as plausible values (PVs). Using item parameters anchored at their estimated values from the international calibration, the plausible values are random draws from the marginal posterior of the latent distribution for each student (OECD, 2014a). In this research, five possible values for mathematical literacy were separately analyzed. The reliability estimates for the cognitive tests were high and the coefficients ranged between 0.88 and 0.90 for Turkey.

The research also used mathematics student data surveys to identify the variables related to student performance. The independent variables were identified using indices variables and

these were the scales consisting of several questionnaire items to measure teaching quality and teaching practices. Brief descriptions of these indices from PISA technical report are given below (OECD, 2014a):

**Teaching quality indice included four scales:**

**Cognitive activation strategies (COGACT).** Nine items measured cognitive activation in mathematics lessons (e.g. The teacher gives problems that require us to think for an extended time, The teacher helps us to learn from mistakes we have made). Response categories were "Always or almost always", "Often", "Sometimes" and "Never or rarely". All items were reversed

**Mathematics teacher support (MTSUP).** This scale provides information on mathematics teacher support. There are four items in this scale (e.g. My teacher lets us know we need to work hard, My teacher helps students with their learning). The four response categories vary from "Strongly agree" to "Strongly disagree". All items were reversed.

**Classroom management (CLSMAN).** This scale provides information on classroom management and consists of four items (e.g. My teacher keeps the class orderly, My teacher starts lessons on time). The four response categories vary from "Strongly agree" to "Strongly disagree". All items except the last one (ST85Q04) were reversed.

**Disciplinary climate (DISCLIMA).** This scale provides information on disciplinary climate in the classroom based on five items (e.g. There is noise and disorder, Students don't start working for a long time after the lesson begins). The four response categories were "Every lesson", "Most lessons", "Some lessons", to "Never or hardly ever". All items were reversed

The teaching practices items about the frequency of using 13 different teaching practices from the OECD Teaching and Learning International Survey (TALIS) survey were adapted for use in PISA 2012. The items were reframed for use with students and some practices that are specific to mathematics were added. Teaching practices scale indices had three dimension:

**Teacher-centered instruction (TCHBEHTD).** Five items measured teacher behaviour when giving directed instruction (e. g. The teacher sets clear goals for our learning, The teacher tells us what we have to learn). Response categories were "Every lesson", "Most lessons", "Some lessons" and "Never or hardly ever". All items were reversed.

**Teachers' use of formative assessment (TCHBEHFA).** Four items measuring teacher behaviour when conducting formative assessment (e.g. The teacher tells me about how well I am doing in my mathematics class, The teacher tells us what is expected of us when we get a test, quiz or assignment). Response categories ranged from "Every lesson" to "Never or hardly ever". All items were reversed.

**Teachers' student orientation (TCHBEHSO).** Four items measured teacher behaviour when performing student orientation (e.g. The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster, The teacher asks us to help plan classroom activities or topics.) Response categories were "Every lesson", "Most lessons", "Some lessons" and "Never or hardly ever". All items were reversed.

The reliability estimates for internal consistency of the independent variables were generally high and the Cronbach's alpha ranged between 0.57 and 0.85 for Turkey (OECD, 2014a).

### 2.4. Data analysis

In PISA, sampling was selected from different units such as student, teacher (class) and school. In studies involving such sampling structures, it is assumed that the previous level is not independent of its subsequent levels (Raudenbush & Bryk, 2002). In terms of validity and

reliability of the results, it is important to use appropriate analysis methods for such data structures (Moerbeek, Breukelen & Berger, 2002). For this reason, Hierarchical Linear Models (HLM) which do not ignore the relations between the observation units have been used as a data analysis method in the research. In terms of the independence and homoscedasticity assumptions, the HLM analysis allows determining standard errors and thus provides more reliable results. Furthermore, the final weights of students were included in the analysis. In this study, the following two models were developed to analyze the data:

**Random effects one-way ANOVA model.** One-way ANOVA Model was used to test whether there was a difference between schools in mathematics literacy. In addition, intracluster correlation coefficient was calculated to understand how much of the variance derived from the variables added to the model

**Random coefficient regression model.** This model helped identifying the variables included in the student questionnaire and impacted on students' achievement in mathematics literacy. Then, it was calculated to find out the extent these variables explain the percentage change in students' mathematics literacy scores.

## 3. FINDINGS

### 3.1. Random effects one-way ANOVA model

This model determines if there was enough school variance to justify the use of HLM. The general model is represented for mathematics scores of students:

Level 1 $( Y i j \ | F / M PV 1 -5 ) = \beta 0 j + r ij$

Level 2 $\beta 0 j = \gamma 00 + u0 j$

Expanded model: $Y ij = \gamma 00 + u0 j + r ij$

where $\beta 0 j$ was the intercept and $\gamma 00$ is the average mathematics scores for students. Further, $r ij$ and $u0 j$ were the random effect terms at student and teacher level models. The results of the one-way ANOVA with Random Effects model were presented in Table 1.

**Table 1.** Fixed and random effects for mathematics scores in one-way ANOVA model

| Fixed Effect | Coefficient | SE | t-ratio | df | p |
|---|---|---|---|---|---|
| Intercept $\beta_{0j}$, | | | | | |
| Overall school mean, $\gamma 00$, | 439.79 | 5.77 | 76.14 | 168 | <0.001 |
| Random Effect | | Variance | df | $\chi^2$ | p |
| School mean, $u_{0j}$ | | 5344.20 | 168 | 4984.20 | <0.001 |
| Level 1 effect, $r_{ij}$ | | 3589.84 | | | |

As revealed in Table 1, according one-way ANOVA model, the mathematics scores of students significant differed according to schools and these differences among schools in terms of mathematics scores were also random ($\chi^2_{168}$=4984.20, $p$<.001) and overall school mean was 439.79. The variability within the school and between the schools were estimated to be 5344.20 and 3589.84, respectively. In this regard, intraclass correlation was calculated as 0.40, which indicates that 40% of the variability in the mathematics scores of students was explained by the mean mathematics achievement of the school.

### 3.2. Random Coefficient Regression Model

The second model is random coefficient regression model. This model aimed to identify the student level variables that had an effect on mathematics scores of students. The following models were obtained:

**Level 1**

$$(Y\,i\,j \mid F / M\,PV\,1 -5\,) = \beta 0\,j + \beta_1 j\,(\text{COGACT}\,) + \beta_{2\,j}\,(\text{DISCLIMA}) + \beta_{3j}\,(\text{TCHBEHSO})$$

**Level 2**

$$\beta\,0\,j = \gamma 00 + u0\,j$$
$$\beta\,1j = \gamma 10 + u1\,j$$
$$\beta\,2\,j = \gamma 20 + u2\,j$$
$$\beta\,3\,j = \gamma 30 + u3\,j$$

where $\beta_{0\,j}$ is the intercept, $\beta_{1\text{-}3\,j}$ are slopes or effects of predictors. The term *r ij* is the random effect for student i nested in teacher j. The error terms were $u_{0\,j} - u_{3\,j}$ at the models. The results of the random coefficient regression models were presented in Table 2.

**Table 2.** Fixed and random effects for mathematics scores in random coefficient model

| Fixed Effect | Coefficient | SE | t-ratio | df | p | Effect size |
|---|---|---|---|---|---|---|
| Average mathematics achievement, γ00 | 444.255 | 5.535 | 80.251 | 168 | <0.001 | |
| TCHBEHSO, $\gamma_{30}$ | -9.856 | 1.493 | -6.601 | 50 | <0.001 | 0.167 |
| COGACT, $\gamma_{10}$ | 5.624 | 1.265 | 4.445 | 5137 | <0.001 | 0.094 |
| DISCLIMA, $\gamma_{20}$ | 4.927 | 1.466 | 3.360 | 59 | 0.002 | 0.084 |
| Random Effect | SD | Variance | $\chi^2$ | df | p | |
| INTRCPT1, u0 | 69.27 | 4799.37 | 4588.60 | 168 | 0.000 | |
| level-1, r | 59.33 | 3520.25 | | | | |

As it can be seen in Table 2, the findings show that the effects of the teacher behaviors student oriented, teachers' use of cognitive activation strategies and disciplined attitudes towards mathematics scores of students were statistically significant. The remaining variables were removed from the model. Teacher behavior that was perceived to be more students oriented (*TCHBEHSO*) had the highest correlation with mathematics scores of students. There was a negative significant relationship between these two variables (γ30= -9.856, *SH*=1.49, *p*<0.01). If all significant variables were controlled at the model, one-unit increase in the level *TCHBEHSO* caused 9.85 point decrease in mathematics scores of students. Considering the effect size of *TCHBEHSO*, an increase of one standard deviation in this variable would result in an decrease of 0.167 standard deviation in their mean scores.

On the other hand, a positive relationship was found between the frequency of the cognitive activation in mathematics lessons and the mathematics scores of students (γ10=5.624, *SE*=1.26, *p*<0.01). If all significant variables were controlled in the model, one-unit increase in the level *COGACT* caused 5.62 point increase in mathematics scores of students. Considering the effect

size of *COGACT*, an increase of one standard deviation in this variable would result in an increase of 0.094 standard deviation in their mean scores. Similarly, a better disciplinary climate was positively correlated with the mathematics scores of students ($\gamma_{20}$=4.92, *SE*=1.46, $p$<0.01). If all significant variables were controlled in the model, one-unit increase in the level *DISCLIMA* caused 4.92 point increase in mathematics scores of students. Considering the effect size of *DISCLIMA*, an increase of one standard deviation in this variable would result in an increase of 0.084 standard deviation in their mean scores.

Adding these three variables into student-level analysis reduced school variability from 5344.20 to 3520.25, which indicated that these variables explained 39.40% of within school variability in the mathematics scores of students.

## 4. DISCUSSION and CONCLUSION

This research explores the mathematics literacy performance of Turkish students in PISA 2012 within the context of teachers' class activities. HLM analysis shows that there are significant differences in terms of mathematics literacy between schools. It also indicates that while the increase in the student-centered instructions within the context of classroom activities decreased students' mathematics literacy scores, teachers' frequency of activating cognitive processes and a good disciplined environment increased students' mathematics literacy skills.

Studies in the relevant literature also report similar results. For example, many research results in Turkey display a very heterogeneous distribution in terms of the performance of schools (Berberoğlu & Kalender, 2005; MEB, 2007; 2010; 2011). This heterogeneity can be explained by teacher characteristics as studies show that the main factor determining the quality of education in schools is teachers (e.g. Harris & Sass, 2011; Hill & Rowe, 1996; Rowe & Hill, 1998; Lamb & Fullarton, 2002). Several studies have concluded that classrooms as well as schools are important and that teacher and classroom variables account for more variance than school variables (Scheerens et al., 1989; Scheerens, 1993). Schmidt et al. (1999) in their comparison of achievement across countries using TIMSS (Trends in International Mathematics and Science Study) data reported that classroom-level differences accounted for a substantial amount of variation in several countries including Australia and the United States. For example, in the United Kingdom, a recent study of 80 schools and 170 teachers measured achievement growth over the period of an academic year, using start-of-year and end-of-year attainment data. Using multi-level modelling techniques, the impact of teachers on achievement growth was measured. They claimed that over 30 per cent of the variance in student progress was due to teachers. They concluded that teacher quality and teacher effectiveness, rather than other classroom, school and student factors, are large influences on student progress (Hay McBer, 2000).

In this study, it was found a negative correlation between the effects of teacher behaviors related to student orientation and student achievement. Student-centered approaches require motivated and self-directed learners (Lee, 2000) and the learning environment should maximize these psychological needs. In this learning environment, teachers shift their roles from "instructors" to "supporters" of a learning process (Schaal & Bogner, 2005). From a logistic point of view, student-centered approaches might enhance students' motivation and interest and increase their cognitive achievement as outlined above. However, recent studies have linked student centered learning environments not necessarily to an overall high cognitive learning outcome, but rather to conventional approaches (Jones, 2012; Randler & Bogner, 2002; Schaal & Bogner, 2005). Many studies comparing teacher-centered and student-centered learning environments produced controversial results with no consistency in the explanation of effects of different learning environments on achievement (Sturm & Bogner, 2008). Potential reasons for such results could be related to the lack of experience in hands-on activities as well as open or

learner-centered approaches. These approaches may create unfamiliar environments and cause anxiety which inhibits learner achievement (Kagan & Fasan, 1988; Randler & Bogner, 2002). In Turkey, teachers experience problems in adopting student-centered approaches in their classes and need training on assessment and evaluation techniques that align with such approaches (e.g. Gözütok Akgün & Karacaoğlu 2005; Gelbal & Kelecioğlu, 2007). Futhermore, according to the results of the Teaching and Learning International Survey (TALIS), there is a negative relationship between class size and frequency of student-centered practices (OECD, 2009). So, teachers may use less student-centered approaches due to the larger classes in Turkey and this result in lower student achievement.

According to the findings of this study, there is a positive relationship between teachers' use of cognitive activation and mathematics achievement. Similarly, in the literature, a positive relationship has been reported between this variable and mathematics achievement (e.g. Baumert & Kunter, 2013; Kunter et al., 2013; Lipowsky et al., 2009). Further support for this finding comes from TALIS which notes that in order to foster cognitive activity, teachers need to use deep and challenging content (Mayer, 2004). It appears that argumentation and non-routine problem solving develop pupils' ability to make connections between mathematical facts, procedures, ideas and representations. As students do more cognitive analysis, a conceptual and deeper understanding of the content may be reached (Chi & Wylie, 2014; Mayer, 2004).

In this study, discipline climate was found to be positively related with their mathematics achievement. This is consistent with the results reported by previous research (e.g. Arum & Velez, 2012; Ning et al., 2015; Sortkær & Reimer, 2016; Usta, 2014). It cannot be expected that student achievement is independent of the characteristics of the classroom environment. A classroom which has teacher-student and student-student based communication can provide a safe environment for students and encourage student participation. For an effective classroom management, teachers should control the classroom, diminish the disruptions and be aware of students' behaviors (Brophy, 1986; Peart & Campbell, 1999). Otherwise, students may feel stressed and will not have the ideal environment for learning and teachers will spend time to deal with disciplinary problems. Accordingly, the division of the lesson may lead to the distraction of the teacher and the student and the reduction of interest (Smith & Laslett, 1992). Students in classrooms where there is no problem of discipline are more willing to learn, feel more academic and better able to focus on what is taught, which can positively affect mathematical achievements (EARGED, 2010).

When the results of the study are taken into account, it can be argued that a positive learning environment that eliminates attention distracting factors are needed as well as approaches that could increase student participation and prevent undesirable student behaviours should be used. In addition, teachers should use methods and techniques that could stimulate students' cognitive skills. Moreover, results of this study also have some implications for teacher educators. Policy-makers should develop effective teacher training programs.

**Acknowledgements**

This study was presented at the IIIrd International Eurasian Educational Research Congress in 31 May- 3 June 2016 in Mugla, Turkey.

**ORCID**

Safiye Bilican Demir http://orcid.org/0000-0001-9564-9029

## 5. REFERENCES


Aksu,G., Güzeller, C.O.,& Eser, M. T (2017). Analysis of maths literacy performances of students with Hierarchical Linear Modeling (HLM): The Case of PISA 2012 Turkey. *Education and Science*, *42* (191),247-266.

Akyüz, G., & Berberoğlu, G. (2010). Teacher and classroom characteristics and their relations to mathematics achievement of the students in the TIMSS, *New Horizons in Education*, *58*(1),77-95.

Akyüz, G., & Pala, N. M. (2010). PISA 2003 sonuçlarına göre öğrenci ve sınıf özelliklerinin matematik okuryazarlığına ve problem çözme becerilerine etkisi (The effect of student and class characteristics on mathematics literacy and problem solving in PISA 2003). *İlköğretim Online*, *9*(2), 668-678.

Akyüz, G., & Satıcı, K. (2013). PISA 2003 verilerine göre matematik okuryazarlığının çeşitli değişkenler açısından incelenmesi: Türkiye ve Hong Kong-Çin modelleri (Investigation of the Factors Affecting Mathematics Literacy Using PISA 2003 Results: Turkey and Hong Kong-China). *Kastamonu Üniversitesi Kastamonu Eğitim Dergisi*, *21*(2), 503-522.

Anderson, J. O., Lin, H-S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large scale assessment data sets for research in science and mathematics education: Programme for international student assessment (PISA). *International Journal of Science and Mathematics Education*, *5*, 591-614.

Archambault, I., Janosz, M., & Chouinard, R. (2012). Teacher beliefs as predictors of adolescents' cognitive engagement and achievement in math. *The Journal of Educational Research, 105*(5), 319-328.

Arum, R., & Velez, M. (2012). *Improving learning environments - school discipline and student achievement in comparative perspective*. Stanford: Stanford University Press.

Baumert, J., & Kunter, M. (2013). The effect of content knowledge and pedagogical content knowledge on instructional quality and student achievement. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project* (pp. 175–206). New York, NY: Springer.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi (Investigation of student achievement across years, school types and regions: The SSE and PISA analyses). *Eğitim Bilimleri ve Uygulama*, *4*(7), 21-35.

Blanton, L. P., Sindelar, P. T., Correa, V., I., Hardman, M., McDonnell, J., & Kuhel, K. (2002). *Conceptions of beginning teacher quality: Models for conducting research.* (COPSSE Document Number RS-6E). Gainesville, FL: University of Florida, Center on Personnel Studies in Special Education. Retrieved from http://copsse.education.ufl.edu/copsse/docs/RS-6/1/RS-6.pdf

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

Brophy, J. (1986). Teacher behavior and student achievement. *American Psychologist, 41*(10), 1069-1077. Doi:10.1037/0003-066X.41.10.1069

Bynner J., & Parsons S. (2005). *Does numeracy matter more?* London: National Research and Development Centre for adult literacy and numeracy, Institute of Education, University of London. Retrieved from http://discovery.ucl.ac.uk/10004758/1/parsons2006does.pdf

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. Doi:10.1080/00461520.2014.965823

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives*, *8*(1), 1–46. Retrieved from https://epaa.asu.edu/ojs/article/viewFile/392/515

Davis-Langston, C. (2012). *Exploring relationships among teaching styles, teachers' perceptions of their self efficacy and students' mathematics achievement.* (Doctoral dissertation). Virginia, Liberty University. Retrieved from https://digitalcommons.liberty.edu/doctoral/497

Deal, S. R. (2010). Graduate students' perceptions of teacher effectiveness. *The Journal of the Oklahoma Association of Teacher Educators*, 14, 9-18. Retrieved from https://files.eric.ed.gov/fulltext/ED524989.pdf#page=9

EARGED. (2008). PISA'da okuma becerileri, PISA'da matematik okuryazarlığı (Reading skills in PISA, Mathematical literacy in PISA). Retrieved from http://earged.meb.gov.tr/pisa/dokuman/2009/2009pisa.pdf

EARGED (2010). EARGED PISA Bülteni 3. Retrieved from http://yegitek.meb.gov.tr/earged/dosyalar%5Cpisa%5Cpisa_bulten3.pdf

Eccles J.S., & Roeser R.W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, *21*(1), 225–241. Doi:10.1111/j.1532 7795.2010.00725.x

Gallagher, H. A. (2002, April). *The relationship between measures of teacher quality and student achievement: The case of Vaughn elementary*. Paper presented at the annual meeting of American Educational Research Association, New Orleans, L. A. Retrieved from https://files.eric.ed.gov/fulltext/ED468254.pdf

Gelbal, S., & Kelecioğlu, H. (2007). Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront. *Hacettepe University Journal of Education, 33*, 135-145.

Gözütok, F. D., Akgün, Ö. E., & Karacaoğlu, C. (2005). İlköğretim programlarının öğretmen yeterlikleri açısından değerlendirilmesi (Evaluation of primary education programs in terms of teacher competencies). *Eğitimde Yansımalar: VIII Yeni İlköğretim Programlarını Değerlendirme Sempozyumu Bildiri Kitabı* (ss.17-40), Ankara: Sim Matbaası.

Harris, D.N., & Sass, T.R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8), 798-812. Doi: 10.1016/j.jpubeco.2010.11.009

Hay, McBer. (2000). *Research into teacher effectiveness: A model of teacher effectiveness.* (Report no: 210), Nottingham: Department for Education and Employment. Retrieved from http://ateneu.xtec.cat/wikiform/wikiexport/_media/formgest/equips_directius/st02/bloc_5/5_rr216investigacio_professors_eficients.pdf

Hill, P.W., & Rowe, K.J. (1996). Multilevel modelling in school effectiveness research. *School effectiveness and school improvement, 7*, 1-34.

Hill, P.W., & Rowe, K.J. (1998). Modelling student progress in studies of educational effectiveness. *School effectiveness and school improvement, 9*(3), 310-333.

İlgün-Dibek, M. (2015). *PISA 2012 matematik okuryazarlığı ile öğrenme ve öğretme süreci değişkenleri arasındaki ilişkiler* (Relationship between learning and teaching variables and PISA 2012 mathematics literacy). Unpublished Master's thesis. Ankara University, Ankara.

Jones T. (2012). Community in the classroom: An approach to curriculum and instruction as a means for the development of student cognitive, social, and emotional engagement in a high school classroom. *Educational Perspectives, 44* (1-2), 58-64.

Kagan, D., & Fasan, V. (1988). Stress and the environment. *College Teaching*, *36*, 75–80.

Kirkpatrick, N.D (2002). *The effects of teacher quality variables on students' mathematics achievement* (Unpublished doctoral dissertation), Available from ProQuest Dissertations and Theses database. (UMI No. 3067771)

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. Doi:10.1037/a0032583

Lamb, S. & Fullarton, S. (2002). Classroom and teacher effects in mathematics achievement: Results from TIMSS. *Australian Journal of Education*, *46*(2), 154-171. Doi:10.1177/000494410204600205

Lee, C. Y. (2000). Student motivation in the online learning environment. *Journal of Educational Media & Library Sciences*, *37*(4), 367–375.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. Doi: 10.1016/j.learninstruc. 2008.11.001

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning?. *American Psychologist*, *59*(1), 14–19. Doi: 10.1037/0003-066X.59.1.14

McCrone, S., & Dossey, J. A. (2007). Mathematical literacy–It's become fundamental. *Principal Leadership*, 7(5), 32–37.

MEB. (2005). İlköğretim matematik 6-8. sınıflar öğretim programı kitabı (Primary Mathematics 6-8. Classes Curriculum Book). Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı. Retrieved from http://talimterbiye.mebnet.net/Ogretim%20Programlari/ortaokul/2010-2011/Matematik%20-%206%20.pdf

MEB. (2007). PISA *2006 Uluslararası Öğrenci Değerlendirme Projesi, Ulusal Ön Rapor (Program for international student assessment PISA 2006, national report)*. Ankara: Eğitim Araştırma ve Geliştirme Dairesi Yayınları. Retrieved from http://www.earged.meb.gov.tr/ dosyalar/dokumanlar/uluslararasi/PISA2006.pdf

MEB. (2010). *PISA 2009 uluslararası öğrenci değerlendirme projesi, ulusal ön rapor (Program for international student assessment PISA 2009, national report)*. Ankara: Eğitim Araştırma ve Geliştirme Dairesi Yayınları. Retrieved from http://pisa.meb.gov.tr/wp-content/uploads/2013/07/pisa-2009-ulusal-on-rapor.pdf

MEB. (2011). *TIMSS 2007 ulusal matematik ve fen raporu 8. Sınıflar (TIMSS* 2007 national mathematics and science report Grades 8). Ankara: EARGED. Retrieved from http://timss.meb.gov.tr/wp-content/uploads/timss_2007_ulusal_raporu.pdf

Moerbeek, M., Breukelen, G.J.P., & Berger, M.P.F. (2002). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*, 341-350.

Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X., & Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement: A cross country comparative study. *School Effectiveness and School Improvement*, *26*(4), 1–26. Doi:10.1080/09243453.2015.1025796

OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD Publishing.

OECD (2009). *Creating effective teaching and learning environments first results from TALIS*. Paris: OECD Publication. Retrieved from http://www.oecd.org/education/school/43023606.pdf

OECD. (2010). The high cost of low educational performance: The long-run economic impact of improving educational outcomes. Paris: OECD Publication. Retrieved from https://www.oecd.org/pisa/44417824.pdf

OECD. (2014a). *PISA 2012 Technical Report*. Paris: OECD Publication. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2014b*). PISA 2012 Results: What students know and can do – Student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014). Paris: OECD Publication. Retrieved from https://www.oecd.org/pisa/keyfindings/pisa_2012-results-volume-I.pdf

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in the first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis,30*, 111-140.

Peart, N. A., & Campbell, F. A. (1999). At-risk students' perceptions of teacher effectiveness. *Journal for a Just and Caring Education, 5*(3), 269–284.

Randler, C., & Bogner, F. X. (2002). Comparing methods of instruction using bird species identification skills as indicators. *Educational Research, 36*(4), 181–188.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage.

Rowe, K.J., & Hill, P.W. (1998). Modelling educational effectiveness in classrooms: The use of multi-level structural equations to model students' progress. *Educational Research and Evaluation, 4*(4), 307-347.

Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42*, 95–103. doi: 10.1016/j.learninstruc.2016.01.004

Schaal, S., & Bogner, F. X. (2005). Human visual perception—Learning at workstations. *Journal of Biological Education, 40* (1), 32–37. Doi:10.1080/00219266.2005.9656006

Scheerens, J. (1993). Basic school effectiveness research: Items for a research agenda. *School Effectiveness and School Improvement, 4*(1), 17-36. Doi:10.1080/0924345930040102

Scheerens, J., Vermeulen, C. J. A. J., & Pelgrum, W. J. (1989). Generalizability of instructional and school effectiveness indicators across nations. *International Journal of Educational Research, 13*(7), 789 - 799.

Schleicher, A. (2007). Can competencies assesed by PISA be considered the fundemental school knowledge:15-year-olds should possess? *Journal of Educational Change, 8*(4), 349-357.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education*. Dordrecht: Kluwer.

Smith, C.J., & Laslett, R. (1992). *Effective Classroom Management*. United States: Routlage.

Sortkær, B., & Reimer, D (2016). Disciplinary climate and student achievement: evidence from schools and classrooms. Aarhus: Danish School of Education, Aarhus University. Retrieved from http://edu.au.dk/fileadmin/edu/Udgivelser/Eboeger/Working_papers_DPU_2016_Disciplinary_Climate_and_Student_Achievement.pdf. ID: 2347254578

Sturm, H., & Franz X. Bogner, F.X. (2008) Student-oriented versus teacher-centred: The effect of learning at workstations about birds and bird flight on cognitive achievement and motivation. *International Journal of Science Education*, *30*(7), 941-959, DOI: 10.1080/09500690701313995

Tennant, J. E., Demaray, M. K., Malecki, C. K., Terry, M. N., Clary, M., & Elzinga, N. (2015). Students' ratings of teacher support and academic and social-emotional wellbeing. *School Psychology Quarterly*, *30*(4), 494-512. https://doi.org/10.1037/spq0000106

Usta, H.G. (2014). *PISA 2003 ve PISA 2012 Matematik okuryazarlığı üzerine uluslararası bir karşılaştırma: Türkiye ve Finlandiya* (PISA 2003 and PISA 2012 mathematics literacy on an international comparison: Turkey and Finland). Unpublished Master's thesis. Ankara University, Ankara.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89–122.

Wenglinsky, H. (2000). How teaching matters: Bringing the classroom back into discussions of teacher quality. Princeton. (ERIC Document Reproduction Service No ED447128).NJ: Educational Testing Service, Policy Information Center. Doi:10.1177/019263650408863811

Wentzel K.R., Battle A., Russell S.L., & Looney L.B. (2010). Social supports from teachers and peers as predictors of academic and social motivation. *Contemporary Educational Psychology*, *35*(3), 193–202. Doi: 10.1016/j.cedpsych.2010.03.002

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 659–667

DOI: 10.21449/ijate.463871

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# The Teacher Academic Buoyancy Scale: Is it possible to keep TABS on students' academic buoyancy?

**Diarmuid Verrier** [1,*], **Scott Johnson** [1], **Lisa Reidy** [1]

[1] Department of Psychology, Sociology, and Politics, Sheffield Hallam University, Sheffield, S10 2BQ, UK

**ARTICLE HISTORY**

*Received: 29 June 2018*
*Revised: 13 September 2018*
*Accepted: 24 September 2018*

**KEYWORDS**

Academic,
Buoyancy,
Resilience,
Teachers,
Scale

**Abstract:** Academic buoyancy (AB) is the ability to overcome minor academic setbacks. However, although it seems as though teachers would be well placed to comment on this characteristic in students, no teacher-report measure of AB exists. This study evaluates a teacher-report version of the widely used, student-report, Academic Buoyancy Scale (ABS). Confirmatory factor analysis supported the unifactorial nature of the Teacher Academic Buoyancy Scale (TABS), and the scale showed excellent internal reliability. However, while there was some evidence for the criterion-related validity of the TABS, it showed very poor convergent validity with the ABS. It also correlated better with academic achievement than should theoretically be the case for a measure of AB. Further, AB estimates from the two measures were moderated by demographic characteristics: teachers rated girls and those not facing adversity as more buoyant, but the opposite was the case for self-reports. In sum, this study suggests a significant disjunction between teacher- and self-reports of AB, and that teacher estimates of AB are likely to be affected by salient, non-AB-related, student characteristics.

## 1. INTRODUCTION

Resilience is defined as the ability to overcome adversity through adaptation (Howard & Johnson, 2000) and has clear implications for achievement and coping within an individual's life (Drapeau, Saint-Jacques, Lépine, Bégin & Bernard, 2007). Resilience considered in an academic context has led to the exposition of two fundamental constructs: academic buoyancy (AB) and academic resilience (AR; Martin, 2013). Both AR and AB are relevant to the ability of individuals to overcome setbacks that have the potential to limit motivation and performance (Martin & Marsh, 2008a). However, they are ontologically distinct in terms of their kind and degree, relevance to different populations (Martin & Marsh, 2008a), and relevance to academic outcomes (Martin, 2013). Specifically, AB applies to the majority of individuals in an academic setting whereas AR applies to individuals who may experience acute adversity (e.g. major illness or emotional/behavioural difficulties; Martin & Marsh, 2009). Poor AB is related to low-level negative outcomes such as achievement anxiety, isolated poor grades, temporary lapses in engagement and motivation, and minor negative interactions with teachers; poor AR is

CONTACT: Diarmuid Verrier ✉ d.verrier@shu.ac.uk Department of Psychology, Sociology, and Politics, Sheffield Hallam University, Sheffield, S10 2BQ, UK

**ISSN-e: 2148-7456 /© IJATE 2018**

predictive of high-level negative outcomes such as disengagement from school, chronic under achievement, sustained disaffection and truancy, and opposition to teachers (Martin, 2013). On the other hand, high levels of AB have been found to correlate well with attendance, engagement in lessons, positive regard for school, persistence, planning, confidence, and academic achievement (Martin, Colmar, Davey, & Marsh, 2010; Martin & Marsh, 2006; Martin, 2014b). Putwain and Daly (2013) looked at the effect of AB and test anxiety on performance on the General Certificate of Secondary Education (GCSE) in the UK. Test scores were shown to be best when academic buoyancy was high and test anxiety was low to medium. A number of student characteristics show very small associations with buoyancy. For example, being female, facing adversity, and being from a lower socio-economic background all predict lower buoyancy (Martin, 2013)

Despite its potential importance, academic buoyancy remains an understudied construct with most of the research done by a relatively small group of scholars. Further, there is a repeated concern throughout the AB literature that all data so far has been generated solely via self-report measures (Malmberg, Hall, & Martin, 2013; Martin, 2013; Martin et al., 2010). Bernard, Killworth, Kronenfeld and Sailor (1984) suggested that on average almost half of self-report responses may be inaccurate, while Fan et al. (2006) highlighted deliberately misleading answers from adolescent participants as a concern. The combined impact of these problems on validity means that AB research is in need of an alternative means of assessment. Given AB's academic context and the responsibility of teachers to monitor students, a student measure that could be completed by teachers would be appropriate.

However, while teachers are quite good at accurately predicting students' academic achievement (Sudkamp, Kaiser, & Moller, 2012), evidence for the validity of their assessments of students' internal, affective and motivational, states has been less impressive. In general, findings in this area have shown the association between teacher- and self-ratings to be low to moderate. For example, a meta-analysis by Renk and Phares (2004) found the average correlation between teacher- and self-ratings of social competence to be around .25, while a more recent study by Zhu and Urhahnes (2014) found a strong association for academic self-concept, a moderate one for learning effort and enjoyment, and no association at all for test anxiety. Teachers' ratings are likely to be affected by student characteristics (Meissel, Meyer, Yao, & Rubie-Davies, 2017) as well as their own characteristics and the school environment (Pas & Bradshaw, 2014).

Of course, the availability of relevant information is also essential for any evaluation of personality (Funder, 2012). As an observer-report instrument, a scale designed to be completed by a teacher would need to be linked to observable behaviours. Martin (2002), drawing on a number of theoretical perspectives, characterised resilient (buoyant) students as demonstrating optimism, proactivity, persistence, attention, effort, and a sense of control over academic outcomes; they do not show anxiety, self-doubt, procrastination, or disengagement. As all of these manifest behaviourally in the classroom, to a greater or lesser degree, a teacher-report measure does seem practical. As there already exists a well-used measure of AB (the Academic Buoyancy Scale; Martin & Marsh, 2008a), the current study used a reworded version of this measure.

The current research has two key aims. The first aim is the psychometric evaluation of this new teacher-report measure of AB -- the Teacher Academic Buoyancy Scale (TABS). Concurrent validity will be assessed via its correlation with the Academic Buoyancy Scale (Martin & Marsh, 2008a), the high school version of the Motivation and Engagement Scale (Martin, 2007), and (prior) academic achievement. The second aim is to investigate which factors might differentially predict teacher-ratings and self-ratings of AB.

## 2. METHOD

### 2.1. Participants

Participants were students and staff at an Academy in the north of England. The students were in the penultimate year (Year 10) of their GCSEs. There were 108 students (56 female; 5 from an ethnic minority group) with a mean age of 14.71 years. Each student's buoyancy was rated by one teacher who was familiar with the student. In total, 50 teachers representing 11 subjects acted as a rater. Each teacher rated at most four students. Eight participants failed to complete at least one entire scales' worth of data and were removed leaving 100 students participants, 24 of whom were identified as facing some form of acute adversity (behavioural difficulty, attendance issues, major illness, or special educational needs).

### 2.2. Measures

#### *2.2.1. Academic Buoyancy Scale*

Academic buoyancy was measured using the 4-item self-report Academic Buoyancy Scale (ABS; Martin & Marsh, 2008a), which uses a 7-point Likert scale. This has previously demonstrated excellent test re-test reliability (Martin et al., 2010), and internal consistency (Martin & Marsh, 2008b). Items assess student ability to bounce back from adversity (e.g. “I don’t let a bad mark affect my confidence”). In the present study, Cronbach’s α was .82.

#### *2.2.2. Motivation and Engagement Scale – High School (MES-HS)*

The 44-item MES-HS (Martin, 2014a) is a self-report instrument that assesses 11 motivation and engagement factors in students (see Table 2). It uses a 7-point Likert scale. Previous research has demonstrated good reliability and internal consistency (Martin, 2007).

#### *2.2.3. Academic Achievement*

End of Key Stage 2 and 3 results in English and Maths were used as markers of academic performance. End of Key Stage 2 (KS2) national tests provide standardised results for students at the age of 10-11. End of Key Stage 3 (KS3) results come from tests performed within the school at the age of 12-13. They are not nationally standardised but are internally moderated to check for consistency.

#### *2.2.4. Teacher Academic Buoyancy Scale*

The TABS measure consists of the four items from the ABS reworded to refer to the student in the third person (e.g., "The student doesn't let a bad mark affect their confidence"). A teacher who knows the student well can use it to rate the student's level of academic buoyancy. It uses a 7-point Likert scale (1 = “Strongly disagree” to 7 = “Strongly agree”). Given that the ABS already has an established unifactorial structure, confirmatory factor analysis (CFA) was used to assess whether the TABS was also unifactorial. The maximum likelihood CFA was done in jamovi (jamovi project, 2018) and the model was generally supported. The chi-square value was not significant; the CFI was .98; the TLI was .95; and the SRMR had a value of .03. The only index of fit that did not support the model was the RMSEA, which was .13 (90%CI: .00 - .26). However, this statistic has been shown to frequently reject true models given a low sample size (N≤250; Hu & Bentler, 1999). Intercorrelations between the four items can be seen in Table 1. The internal reliability of this four-item scale was excellent (Cronbach's α = .84).

**Table 1.** *Intercorrelations between items on the Teacher Academic Buoyancy Scale (N = 100)*

| | Mean (SD) | Correlations | | |
|---|---|---|---|---|
| Item | | 2 | 3 | 4 |
| 1. The student does not let stress get on top of them | 4.33 (1.48) | .44 | .72 | .70 |
| 2. The student is good at dealing with schoolwork pressures | 4.98 (1.29) | | .45 | .54 |
| 3. The student does not let a bad mark affect their confidence | 4.50 (1.25) | | | .63 |
| 4. The student is good at dealing with setbacks (e.g. negative feedback on their work, poor results) | 4.68 (1.39) | | | |

*Note*: p < .001 for all correlations.

## 3. RESULTS

### 3.1. Criterion-related validity

Prior to analysis, item scores were reversed where appropriate and summed to provide total scores for each scale and subscale. Four subscales from the MES-HS had skew values that exceeded twice the standard error of skew. A square root transform was sufficient to appropriately reduce this before further analysis.

**Table 2.** *Descriptive Statistics and Correlations between the Academic Buoyancy and Criterion Variables (N=100)*

| | | Correlations | | | | Correlations | |
|---|---|---|---|---|---|---|---|
| | Mean (SD) | TABS | ABS | | Mean (SD) | TABS | ABS |
| **Academic Buoyancy** | | | | **MES-HS** | | | |
| TABS | 18.49 (4.46) | - | - | Self-belief | 74.11 (15.95) | .13 | .30** |
| ABS | 17.09 (4.93) | -.02 | - | Valuing | 73.70 (14.90) | .05 | .26** |
| **Academic Achievement** | | | | Learning Focus | 75.36 (14.39) | .01 | .19 |
| Key Stage 2 | 4.28 (0.51) | .32* | -.22* | Planning | 56.90 (20.76) | .14 | .18 |
| Key Stage 3 | 6.29 (0.78) | .40** | -.24* | Task Management | 68.04 (19.97) | .21* | .11 |
| Progress | 2.03 (0.44) | .30* | -.12 | Persistence | 63.34 (16.79) | .06 | .13 |
| | | | | Anxiety | 67.11 (19.96) | .13 | -.45** |
| | | | | Failure Avoidance | 51.14 (20.11) | .01 | -.16 |
| | | | | Uncertain Control | 53.14 (16.13) | -.09 | -.12 |
| | | | | Self-Sabotage | 39.61 (18.58) | -.16 | -.21* |
| | | | | Disengagement | 41.64 (19.95) | -.16 | -.34* |

*p<.05 **p<.001

Notably, there was no association between the teacher and student ratings of student academic buoyancy, suggesting poor convergent validity. Evidence for concurrent validity was assessed through correlations between AB measures and the MES-HS subscales (see Table 2). With regards to the MES-HS subscales, the TABS showed a weak positive correlation with task management. No other correlations were significant (though there did appear to be weak effects that were generally in the expected direction). The ABS also correlated in the expected direction with a number of MES-HS subscales.

AB has been shown to have a weak positive association with academic achievement in certain contexts. Here, the ABS correlated negatively with (prior) academic performance, though these were, predictably, weak effects (see Table 2). In contrast, the TABS showed moderate positive associations with students' KS2 and KS3 results and their academic progress (difference between KS2 and KS3 results).

### 3.2. Moderators of AB rating

We investigated whether the estimates of buoyancy provided by the TABS and ABS were dependent on group membership. An inverse pattern was observed when comparing boys and girls using the two measures (see Fig. 1a). Boys rated themselves as more buoyant, $t(98) = 3.25$, $p = .002$, $d = .66$; however, they were rated by teachers as less buoyant, albeit not significantly so, $t(98) = 1.06$, $p = .290$, $d = .21$.

A similar pattern was seen when comparing students who were identified as facing particular adversity with those who were not (see Fig. 1b). Adversity-facing students rated themselves as more buoyant, $t(98) = 2.17$, $p = .032$, $d = .54$, but were rated as less buoyant by teachers, $t(98) = 4.25$, $p < .001$, $d = 1.04$. The interactions suggested by these two patterns of results (i.e., gender/adversity × AB measure) were found to be significant when tested via 2 × 2 ANOVAs (gender interaction: $\eta_p^2 = .09$; adversity interaction: $\eta_p^2 = .18$).

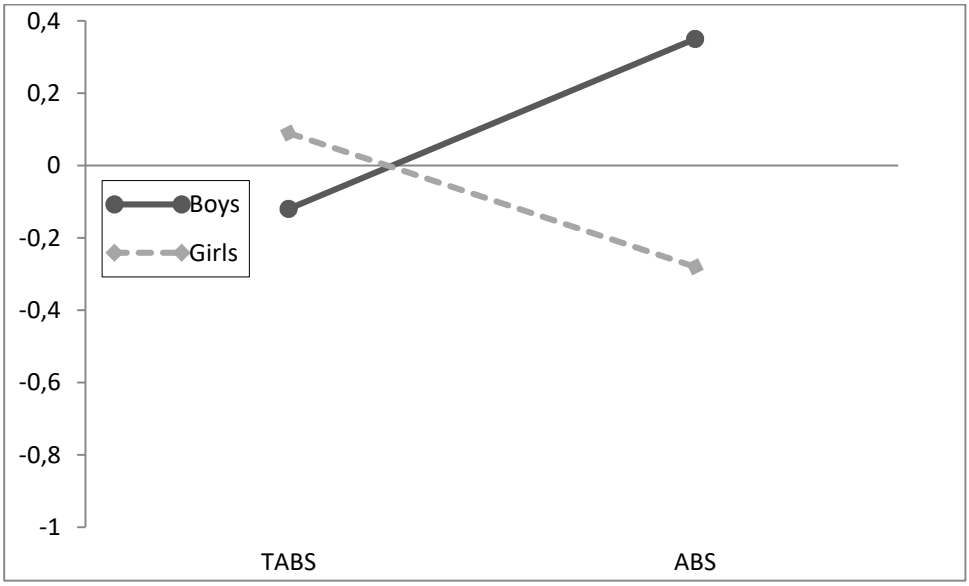



**Figure 1a.** Group differences in teacher-rated (TABS) and self-rated (ABS) academic buoyancy between boys and girls (standardised scores).

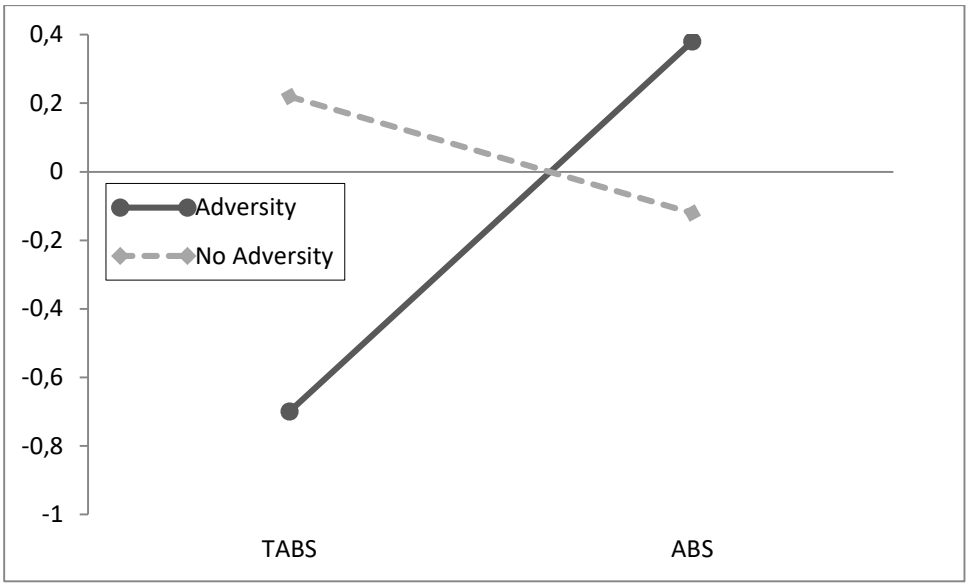



**Figure 1b.** Group differences in teacher-rated (TABS) and self-rated (ABS) academic buoyancy between those facing/not-facing adversity (standardised scores).

## 4. DISCUSSION

This study used a reworded version of the ABS to create a teacher-report measure of academic buoyancy. After confirming that the TABS shows the same unifactorial structure as the ABS, the psychometric properties of the scale were evaluated. The TABS had excellent internal consistency, but showed poor validity in a number of respects. In terms of convergent validity, one would expect a teacher-report of a student's academic buoyancy to correlate with the student's self-report, especially when the wording of the two scales differ so slightly. However, there was absolutely no association between the two scores. Concurrent validity was assessed in two ways: through the TABS' correlation with student motivation and engagement and through its association with prior academic achievement. Correlations between the TABS and the subscales of the MES-HS were uniformly weak, with only one (the association with task management) being significant. Conversely, the ABS showed moderate associations with several of the MES-HS subscales, and all correlations were in the expected direction (Martin, 2007, 2013). AB has been shown, at best, to be a weak predictor of attainment/achievement (Martin, 2014b) and any effect it has is likely to be indirect, through, for example, an enhanced sense of control (Collie et al., 2015). As such, one would expect a valid measure of AB to only weakly correlate with academic achievement. While this was the case for the ABS, there were moderate positive associations between the TABS and academic achievement. Thus, in terms of both convergent and concurrent validity, the TABS seems wanting.

In order to discern whether student characteristics might predict teacher bias (or, indeed, student bias), we also investigated the degree to which gender and facing adversity moderated estimates of AB. Adversity was defined broadly as having experienced behavioural difficulties, attendance issues, or major illness; or as having special educational needs. We found, in keeping with previous data (Martin, 2013) that women reported themselves as less buoyant than men. However, the opposite was the case for teacher reports, where women were perceived as being more buoyant. A similar effect was observed for adversity, where those facing adversity rated themselves as more buoyant (an unexpected finding), while being rated as less buoyant by teachers.

A prerequisite for accurate judgements of personality traits is that relevant behavioural information must be available and detected by the judge (Funder, 2012). The poor association between teacher reports of AB and self-reports of largely psychological states (buoyancy, motivation, and engagement) may suggest that the information teachers would need to make a valid estimate of AB is lacking, or less salient than it could be. In the absence of such information, teachers are likely to rely on more salient, non-AB-related, student characteristics. These could include academic performance, gender, and known adversity. Relying on academic performance as a guide would explain the observed, theoretically unexpected, association between this and TABS scores. Similarly, while women have been shown to report themselves as having lower AB than men do (Martin, 2013), they are seen by teachers to be "good" students: more verbally capable, conscientious, and engaged (e.g., Åhslund & Boström, 2018). Adversity, as defined here, included behavioural difficulties, which has been shown to be associated with unrealistically low judgements of academic performance (Bennet, Gottesman, Rock, Cerullo, & Levin, 1993). The same process may be at work here in teacher judgements of AB. This fundamental problem is likely to be endemic to any attempt to get an accurate observer-report measure of AB. Ensuring that the rater is blind to the student's past academic performance and behavioural history means that they will equally be blind to behaviours that would be relevant to AB.

While the present study suggests that it may not be possible to satisfactorily convert the ABS directly into a teacher-report measure of AB, the need for a teacher- or parent-report AB scale persists. Apart from the triangulationary value provided by such a measure alongside a self-report measure, it could actually be more accurate than self-report. For example, when Skinner, Kinderman, & Furrer (2009) looked at student self-report engagement scores, teacher-assessed engagement scores, and *in vivo* observations of engagement behaviours, the teacher scores were found to correlate better with observed engagement behaviour than self-report. It may be necessary to create such a scale entirely from scratch, paying particular attention to the space in which AB and observable behaviour overlap. Adding guidelines to the scale that can help the rater to disregard academic performance and other salient non-AB-related behaviours may also be of use.

Future research into the creation and validation of a teacher-report scale should use larger samples and evaluate its performance in different academic contexts. Looking at the relative predictive power of self-report and teacher-report measures of AB would help to ascertain whether one is strictly more valuable or whether they should be used alongside one another. To that end, the inclusion of additional criterion variables in future research would also be recommended. For example, the new subscale could be correlated against *in vivo* ratings of students' buoyant behaviours made by an independent observer.

In conclusion, although there is an apparent need for a teacher-report measure of AB, this new scale, a straightforwardly reworded version of the ABS, does not seem to be suitable. While it had a clear factor structure and excellent reliability, it showed suboptimal validity in several respects. It failed to correlate with the ABS at all, correlated poorly with measures of motivation and engagement compared to the ABS, and correlated better than it should have done with academic performance. In general, teacher estimates of internal psychological states in students are moderate at best (Zhu & Urhahnes, 2014), demonstrating how difficult it can be for teachers to make accurate inferences. The pattern of correlations here suggested that buoyancy-indicating behaviours, at least as described in the TABS, may not be apparent to teachers, who instead relied on more salient information to evaluate students.

**ORCID**

Diarmuid Verrier https://orcid.org/0000-0002-4148-3948

Lisa Reidy https://orcid.org/0000-0001-5442-2346

## 5. REFERENCES


Åhslund, I., & Boström, L. (2018). Teachers' perceptions of gender differences: What about boys and girls in the classroom? *International Journal of Learning, Teaching and Educational Research, 17*(4), 28-44.

Bennett, R., Gottesman, R., Rock, D., Cerullo, F., & Levin, Joel R. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347-356.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailor, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology, 13*, 495–517.

Collie, R. J., Martin, A. J., Malmberg, L.E., Hall, J., & Ginns, P. (2015). Academic buoyancy, student's achievement, and the linking role of control: A cross-lagged analysis of high school students. *British Journal of Educational Psychology, 85*(1), 113-130.

Drapeau, S., Saint-Jacques, M., Lépine, R., Bégin, G., & Bernard, M. (2007). Processes that contribute to resilience among youth in foster care. *Journal of Adolescence, 30*, 449–466. DOI: 10.1016/j.adolescence.2007.01.005

Fan, X., Miller, B.C., Park, K., Winward, B.W., Christensen, M., Grotevant, H.D., & Tai, R.H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-reporting surveys. *Field Methods, 18*, 223-244. DOI: 10.1177/152822X06289161

Funder, D. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177-182.

Howard, S., & Johnson, B. (2000). What makes the difference? Children and teachers talk about resilient outcomes for children 'at risk'. *Educational Studies, 26*, 321-337. DOI:10.1080/03055690050137132

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

jamovi project (2018). jamovi (Version 0.9) [Computer Software]. Retrieved from https://www.jamovi.org

Malmberg, L. E., Hall, J., & Martin, A. J. (2013). Academic buoyancy in secondary school: Exploring patterns of convergence in English, mathematics, science, and physical education. *Learning and Individual Differences, 23*, 262-266. DOI: 10.1016/j.lindif.2012.07.014

Martin, A. (2002). Motivation and academic resilience: Developing a model for student enhancement. *Australian Journal of Education, 46*, 34-49. DOI: 10.1177/000494410204600104

Martin, A.J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validity approach. *British Journal of Educational Psychology, 77*, 413-430. DOI:10.1348/000709906X118036

Martin, A.J. (2013). Academic buoyancy and academic resilience: Exploring 'everyday' and 'classic' resilience in the face of academic adversity. *School Psychology International, 34*, 488-499. DOI: 10.1177/0143034312472759

Martin, A.J. (2014a). *Motivation and Engagement Scale – High School*. Sydney, New South Wales, Australia: Lifelong Achievement Group.

Martin, A.J. (2014b). Academic buoyancy and academic outcomes: Towards a further understanding of students with attention deficit/hyperactivity disorder (ADHD), students without ADHD and academic buoyancy itself. *British Journal of Educational Psychology, 84*, 86-107 DOI: 10.1111/bjep.12007

Martin, A.J., Colmar, S.H., Davey, L.A., & Marsh, H.W. (2010). Longitudinal modelling of academic buoyancy and motivation: Do the '5Cs' hold up over time? *British Journal of Educational Psychology, 80*, 473-496. DOI:10.1348./000709910X486376

Martin, A.J. & Marsh, H.W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. *Psychology in the Schools, 43*, 267- 281. DOI: 10:1002/pits.20149.

Martin, A.J., & Marsh, H.W. (2008a). Academic buoyancy: Towards an understanding of students' everyday academic resilience. *Journal of School Psychology, 46*, 53-83. DOI: 10.1016/j.jsp.2007.01.002

Martin, A.J., & Marsh H.W. (2008b). Workplace and academic buoyancy: Psychometric assessment and construct validity amongst school personnel and students. *Journal of Psychoeducational Assessment, 26*, 168-184. DOI: 10.1177/0734282907313767

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48-60.

Pas, E., & Bradshaw, T. (2014). What affects teacher ratings of student behaviors? The potential influence of teachers' perceptions of the school environment and experiences. *Prevention Science, 15*(6), 940-950.

Putwain, D.W., & Daly, A.L. (2013). Do clusters of test anxiety and academic buoyancy differentially predict academic performance? *Learning and Individual Differences, 27*, 157-162. DOI: 10.1016/j.lindif.2013.07.010

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review, 24*(2), 239-254.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2008). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*, 493-525. DOI: 10.1177/0013164408323233

Martin, A. J. (2014). Academic buoyancy and academic outcomes: Towards a further understanding of students with attention-deficit/hyperactivity disorder (ADHD), students without ADHD, and academic buoyancy itself. *British Journal of Educational Psychology, 84*, 86–107. doi:10.1111/bjep.12007

Sudkamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A Meta-Analysis. *Journal of Educational Psychology, 104*(3), 743-762.

Zhu, M., & Urhahne, D. (2014). Assessing teachers' judgements of students' academic motivation and emotions across two rating methods. *Educational Research and Evaluation, 20*(5), 411-427.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 668–681

DOI: 10.21449/ijate.453337

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Gaining a Better Understanding of General Mattering Scale: An Application of Classical Test Theory and Item Response Theory

**Halil Ibrahim Sari** [1,*], **Mehmet Akif Karaman** [1]

[1] Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Department of Educational Sciences, Kilis, Turkey

**ARTICLE HISTORY**

*Received: 14 August 2018*
*Revised: 13 September 2018*
*Accepted: 24 September 2018*

**KEYWORDS**

Classical test theory,
Item response theory,
Generalized partial model,
Item analysis,
General mattering scale

**Abstract:** The current study shows the applications of both classical test theory (CTT) and item response theory (IRT) to the psychology data. The study discusses item level analyses of General Mattering Scale produced by the two theories as well as strengths and weaknesses of both measurement approaches. The survey consisted of a total of five Likert-type items. Each student chose the best answers from the given categories (e.g., not at all, a little, somewhat, very much). We specifically run generalized partial credit IRT model. Overall, we discussed that while CTT provides comparatively superficial information, IRT allows gaining deeper insight into the test items and response categories. We concluded that the meaning of item properties differs in CTT and IRT, and this difference may lead to different interpretations. We aimed to encourage psychologists and counsellors to give more consideration to the IRT applications when assessing the psychometric features of the items.

## 1. INTRODUCTION

The purpose of many tests in the area of educational, psychological or behavioral measurement is to observe and explain human behaviors or feelings, and draw accurate inferences from given answers to test items. However, this is not straightforward task since latent variables are intended to measure unobservable or indirect constructs, such as feelings or emotions. Since the latent variable is measured by a set of items, how much one knows about the items in the test batteries determines how much he/she knows about latent variables. Consequently, how one analyzes, and then makes inferences about test items play a vital role. This study showed how inferences and interpretations made from the items intended to measure students' feelings about mattering would differ across different measurement theories.

There are two types of measurement theories used to describe the relationships between observed (e.g., items) and unobserved variables (e.g., attitude or latent ability). The first one is

CONTACT: Halil İbrahim SARI ✉ hisari87@gmail.com Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Department of Educational Sciences, Kilis, Turkey

**ISSN-e: 2148-7456 /© IJATE 2018**

the classical test theory (CTT). The CTT is a simple measurement theory which also known as true score theory. During the history of psychometrics dating back to early 1900s (Crocker & Algina, 1986), the CTT has been very popular, and intensively used by researchers and practitioners from a variety of fields including counseling and psychology (e.g. Cho, Drasgow, & Cao, 2015; Riemer & Kearns, 2010). This is because the CTT based calculations can be easily computed, even by hand. However, it has been extensively criticized due to producing population or sample dependent interpretations (Lord, 1980). This means that when the same survey is taken by two different groups, item properties such as discrimination and difficulty, hereby, inferences might differ across the groups. The second one is item response theory (IRT; Baker, 1992). IRT, also known as latent trait theory, is a strong psychometric theory. IRT explains the statistical relationships between test taker's ability level and the response given to that item by item response category characteristic curves. Unlike CTT, IRT allows to draw population independent interpretations as long as one has enough sample size (e.g., a minimum of 300). It also separately estimates the errors in the observed scores for all individuals.

Although IRT requires complicated calculations, due to the advantages over CTT, it is seen as the evolutionary measurement theory by practitioners and measurement theorists. Thus, it is very common to see its applications in many studies from a variety of areas including health education (Hays, Morales, & Reise, 2000), teacher education (Brodin, Fors, & Laksov, 2010), psychology (Lee, Krishnan, & Park, 2012), childhood education (Gomez & Vance, 2015), counseling (Riemer & Kearns, 2010) and more. However, unlike CTT, IRT based application examples are limited in the fields of counseling and psychology, and as compared to the CTT, a little consideration has been given to the IRT-based analyses by psychologists.

This study illustrated an application of both CTT and IRT approaches to the psychology related data, and compared item level analyses and interpretations produced by the two measurement theories. This collaboration aimed to empirically support the notion that IRT-based inferences would be more meaningful, hereby its applications should take more place in the field of psychology. Our goal was to draw researchers' attention to the IRT, and encourage counselors and psychologists to give more consideration to the IRT models in their research for a better understanding of their surveys.

### 1.1. Related Studies

Even though the CTT has been the dominantly used theory to assess the psychometric characteristics of the measurement tools, IRT-based applications has gained popularity in recent years. There are some good examples that used IRT approach when assessing the outcomes of the psychological scales. For example, Zanon, Hutz, Yoo and Hambleton (2016) used IRT-based calculations on affect scale that aimed to explore students' emotions and feelings as pleasant or unpleasant. They specifically used the graded response model, and discussed the highlights of IRT for psychological test development. Hamzeh and Fatima (2016) run graded response model, and discussed the findings of IRT-based applications when validating marital satisfaction scale. Riemer and Kearns (2010) compared both CTT and IRT-based findings and interpretations on youth counseling impact scale, and discussed the advantages of IRT over CTT. Another great example was provided by Turner, Betz, Edwards and Borgen (2017). They examined item quality of self-efficacy scale by using both CTT and IRT, and concluded that IRT-based calculations provided better insight to detect low and high quality items. Furthermore, Tasca et al., (2016) used both CTT and IRT when validating the shortened version of the therapeutic factors inventory form. Similarly, Locke et al., (2012) showed CTT and IRT based calculations, and discussed the strengths and weaknesses of the interpretations drawn across the two measurement theories when shortening the counseling center assessment of psychological symptoms scale. Lastly, Lee, Krishnan and Park (2012) run nominal response IRT model, which is a member of Rasch models. They analyzed psychometric properties of the

children's depression inventory scale items, and reported item quality across the different age groups.

### 1.2. Theoretical Framework

When achieving the goal of measuring student feelings, typically Likert-type or polytomously scored items (aka multiple category items) are used. This is because polytomously scored items provide more information than dichotomously scored items (e.g., binary such as true/false, correct/incorrect, pass/fail; Embretson & Reise, 2000). Unlike dichotomously scored items (e.g., binary), polytomously scored items have more than two categories (e.g., strongly disagree, disagree, agree, and strongly agree). However, it is important to add that the choice of item format obviously depends on the purpose and consequences of the survey.

The CTT assumes that a test taker's test score is comprised of his/her "true" score and some measurement error. This is conceptualized as

$$X_{OBSERVEDTESTSCORE} = X_{TRUESCORE} + Error$$

The CTT assumes that the error in the observed score is constant across all test takers, meaning that everyone's true score is calculated by the same amount of error, which is unrealistic in practice. Another assumption is that there is no correlation between true scores and error (Crocker & Algina, 1986).

When items are dichotomously scored, item means represent item difficulty in CTT (i.e., proportion of examinees that answered an item correct). However, when the test items are polytomously scored, item means do not represent more than the average numerical mean score for an item. Item discrimination represents the ability to differentiate low proficiency and high proficiency examinees, and calculated by item-total correlations. If people get score low on an item also get low score on the test (or vice versa), this means that the item has better discrimination parameter.

In the IRT, there are two types of family modeling approaches. These are called cumulative (e.g., indirect) and adjacent (e.g., direct) models. Both type of models define step functions, the relationship between probability of reaching a certain category and latent variable, but differ in interpreting and calculating the steps. The cumulative models define step functions as the probability of scoring for cumulative score categories (e.g., probability of scoring more than 0, probability of scoring more than 1 etc.). Graded Response Model (GRM; Samejima, 1969) is a well-known example of polytomous IRT model that uses cumulative or indirect approach. The direct models, which is the interest in this study, define step functions as the probability of scoring for adjacent score categories (e.g., probability of scoring 1, for a person given that he/she already scored 0 or probability of scoring 2, given that person already scored 1). The Partial Credit Model (PCM; Masters, 1982) and the Generalized Partial Credit Model (GPCM; Muraki, 1992) are some examples of polytomous IRT models that use direct approach. If there are number of *k* response categories for an item, both PCM and GPCM specify *m* difficulty parameters (*b*), where *m*=*k*-1, unique to each step. However, while the GPCM specifies separate discrimination parameters for each item, and the PCM (i.e., also known as polytomous Rasch model) does not allow discrimination parameters to vary across the items, implying that the discrimination parameter equals to 1 for all items. It can be said that the two models are the nested models. The choice of the IRT model approach mainly depends on the theoretical and empirical considerations. One can refer to Huggins-Manley and Algina (2015) for more technical details about the polytomous item response theory.

In this study, we are particularly interested in direct or adjacent approach, and specifically run the GPCM due to empirical reasons (e.g., providing better model fit than PCM). One can refer to the Table 1 for the model-fit information for these two IRT models.

**Table 1.** *Item Response Theory Model-Fit Information*

| Model | AIC | BIC | Log Likelihood | *df* | *p* |
|---|---|---|---|---|---|
| Partial Credit Model | 18358.7 | 18436.6 | -9164.3 | 15 | |
| Generalized Partial Credit Model | 18297.2 | 18405.0 | -9128.6 | 20 | <.001* |

*There is a statistically significant difference between the nested models. Less restricted model (GPCM) fits better with the data.

The probability that an examinee scores the category $j$ on item $i$ is modeled by the GPCM as

$$P(Y = j|\theta) = \frac{exp^{a_i(\theta - b_{ij})}}{\sum_{k=0}^{m}\sum_{j=0}^{k} exp^{a_i(\theta - b_{ij})}}$$

Where $k$ is the number of response options, $m$ is the number of the steps, and $m = k-1$, $a_i$ is the difficulty parameter and $b_i$ is the difficulty parameter for item $i$ on the $j^{th}$ step. The difficulty of a step (aka step parameter) represents the point at which a person has an equal chance or probability of scoring $j$ or $j+1$ (advancement on the $j^{th}$ step). It is important to note that step parameters (e.g., $b_1$, $b_2$, $b_3$) are the interaction points of the two adjacent category response curves. The item response category characteristic curve given in Figure 1 visually shows an example of an item with four response options where $b_1$=-2.28, $b_2$=-0.96, $b_3$=0.85. The discrimination parameter determines the steepness of the category characteristic curves. As the discrimination of an item increases, the category response curves become steeper; hereby probability of selecting a response option rapidly changes.

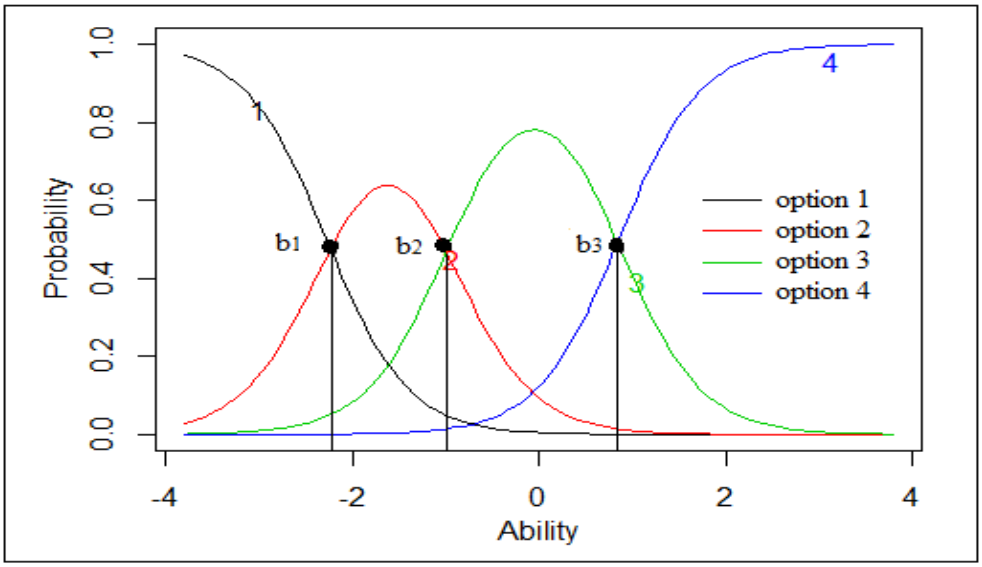



**Figure 1.** An example of item response category characteristic curve for a polytomously scored item.

### 1.3. Mattering

Many researchers in the fields of educational, social, and counseling psychology conducted studies on the concept of mattering and its relationship with other variables (Cha, 2016; Connolly & Myers, 2003; Flett, Galfi-Pechenkov, Molnar, Hewitt, & Goldstein, 2012; Rayle, 2006; Tovar, Simon, & Lee, 2009). However, no recent studies have explored this construct with Turkish participants or in the Turkish society. Rosenberg and McCullough conceptualized the term of mattering as a sense of belonging in 1980s (Tovar, 2013). After reviewing many research databases (e.g. PsychInfo, Academic Search Complete, Google Scholar, ULAKBIM) we have found that first time mattering was studied by Demir, Özen, and Doğan in 2012 with a group of Turkish participants. In other words, the term took attention of Turkish researchers 32 years after it was introduced by Rosenberg and McCullough. There are a few reasons that researchers in Turkey did not show sufficient interest in mattering.

First, Rosenberg and McCullough (1981) described mattering as one's sense of belonging to his/her immediate surroundings and society. We have found that there are many studies conducted on sense of belonging in the Turkish literature. This could be one of the reasons that researchers focused on the core concept of mattering but neglected the big picture- mattering. The second reason could be the lack of mattering measurements available in Turkish. The study conducted by Demir et al. (2012) used Mattering to Others Questionnaire (Marshall, 2001); however, the instrument is not available in Turkish and we assume that it was translated by Demir and colleagues for the study they conducted in 2012. One problem with this is the lack of information of the psychometric properties of the instrument they used because there is no adequate information about the validation process in their study. At our best knowledge, Haktanır, Lenz, Can, and Watson (2016) developed, and validated the first mattering measure which was General Mattering Scale (Marcus, 1991) into Turkish.

In parallel to the current study's aim, we believe that mattering will take more attention and both researchers and practitioners will benefit from its potential value. Moreover, measurements of mattering in different languages will help to conceptualize "our subjective perception and interpretation that we make a difference to others in our lives" (Tovar, 2013, p. 42). As stated by Rosenberg and McCullough (1981), three elements of mattering, which are attention, importance, and dependence, have continued to constitute the theoretical base of measurements and explain external validation of a person by others. This external validation comes true at the interpersonal and societal levels (Rosenberg & McCullough, 1981). The interpersonal dimension indicates individuals' level of mattering to people in their lives. In other words, the feeling of mattering they get from close relationships, such as, parents, friends, teachers. On the other hand, the societal dimension includes individuals' perception of mattering toward outer world, such as, schools, governmental institutions, religious institutions. As the number of mattering studies which have been conducted in different cultures increase, our understanding level of the concept and related variables increase.

## 2. METHOD

### 2.1. Participants and Data Collection

The relevant university ethics board approved the study, the data were collected from volunteer students attending a state university in southeast region of Turkey during 2017-2018 academic year. There were 1644 undergraduate students from five different faculties. Twenty-one cases were removed from the data file since participants left the instrument items blank. This ended the data with 1623 participants. Of the participants, 59.5% were female (n= 966) and 39.5% were male (n= 643), 14 participants failed to respond this demographic query. The ages of respondents were between 17 and 39, the average age of the total participants was 21.4.

Participants reported their academic levels as freshmen (n = 372, 23%), sophomores (n = 439, 27%), juniors (n = 524, 32.3%), and seniors (n=279, 17.2%). Nine participants failed to respond to this demographic query.

### 2.2. Measure

In the current study, we used the General Mattering Scale- Turkish Version (GMS-TR; Haktanir et al., 2016) to collect data. The GMS (Marcus, 1991) was developed to assess the degree individuals believe how they are important to others. This 5-point Likert-type assessment yields a single scale score based on participant responses that range from Very Much to Not at All. Possible scores on the scale range from five to 20, with higher scores being indicative of a greater perception of mattering. Mattering is accounted for by participant responses to items such as "How important do you feel you are to other people?" and "How interested are people generally in what you have to say?" Haktanır et al. (2016) reported moderate Cronbach's alpha coefficients of .74. For the original version, Rayle and Myers (2004) reported alpha coefficients between .74 and .86 among college students. For the current study, we calculated a Cronbach's alpha of .76. The survey is given in the Appendix.

### 2.3. Analysis Procedure

Since the IRT models presented above assume dimensionality (e.g., measuring one and only one construct), we first checked the dimensionality of the data to ensure there was a single construct of interest, named general mattering. Specifically, we run a confirmatory factor analysis (CFA) with five items in AMOS (Arbuckle, 2014). The fitted model that we run was shown in Figure 2. For the CTT interpretations, we calculated the means for each response category across all items, CTT based-observed score distribution, and item level means and discrimination parameters. We also calculated CTT-based reliability index known as Cronbach's alpha. For the IRT interpretations, we run the GPCM, and calculated similar parameters. For the reliability measures in the IRT context, we calculated item and test information function which corresponds to the reliability index in IRT. We used the SPSS version 22.0 (IBM, 2013) for the CTT calculations, and the "ltm" package (Rizopoulos, 2017) in R Software (R Core Team, 2016) for the IRT calculations.

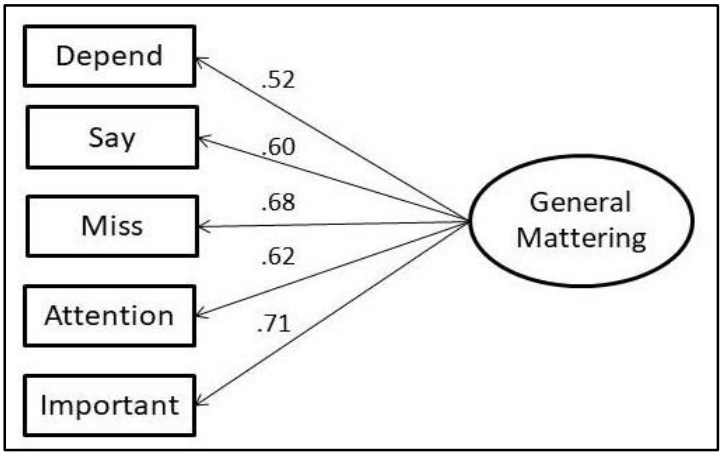



**Figure 2.** The specified CFA model of general mattering

## 3. RESULTS

The results of CFA indicated that the single factor model of general mattering scale (see Figure 2) had a strong fit; $\chi^2(5)= 47.22$, $p< .05$; GFI= .99, CFI= .97, TLI= .95, RMSEA= .07, and SRMR= .03. This shows that we met the unidimensionality assumption. Item and scale level findings produced by the CTT and IRT are discussed in the following paragraphs.

### 3.1. CTT Based Findings

Item means, standard deviations, item discrimination parameters and Cronbach's alpha statistics if item deleted across the five items, and scale level statistics produced under the CTT are given in Table 2. The mean for the item related to depend were the highest. It is obvious that there were a lot of students picked "somewhat" or "very much" on this item. The lowest mean was found for the item related to the importance. Item means for the remaining four items ranged from 2.66 to 2.91. This implies that on these four items, students generally picked the two middle options- "a little" and "somewhat". Based on the scale level statistics given at the bottom of Table 2, on average, a student got a score of 2.88 out of 4 for an item. Hence, it is safe to say that students mostly believe that the degree to which others matter to them was close to somewhat. The descriptive statistics for the response categories given in Table 3 also support this. The discrimination parameter, the power of differentiating the respondents, was the lowest and highest for the items related to the depend, and important, respectively. This means that the depend and important items were the most and least effective items to distinguish the students having low and high general mattering score in total, respectively. The discriminations for all items in the scale were higher than acceptable rates (e.g., 0.4). The scale level (e.g., across the five items) Cronbach's alpha measure was 0.76 which is not very high but at acceptable rate. It was also found that deleting any item would increase the alpha value. However, it was seen while the effect of deleting the first four items on the Cronbach's alpha would be more serious and, the effect of removing the depend item from analysis would be negligible.

**Table 2.** *Item and scale level statistics produced under classical test theory model*

| *Item Level* | Item Mean | Standard Deviation | Item-Total Correlation (Discrimination) | Cronbach's Alpha If item deleted |
|---|---|---|---|---|
| Important | 2.66 | 0.92 | .59 | .70 |
| Attention | 2.75 | 0.87 | .52 | .72 |
| Miss | 2.88 | 0.96 | .57 | .71 |
| Say | 2.91 | 0.87 | .53 | .72 |
| Depend | 3.18 | 0.87 | .45 | .75 |
| *Scale Level* | Mean | Standard Deviation | Variance | Cronbach's Alpha |
| | 2.88 | 0.64 | 0.41 | 0.76 |

**Table 3.** *Descriptive statistics for the response options*

| | Response Option | | | |
|---|---|---|---|---|
| Item related to the | Not At All | A Little | Somewhat | Very Much |
| Important | 235 (14.6%) | 344 (21.4%) | 751 (46.8%) | 276 (17.2%) |
| Attention | 140 (8.7%) | 433 (27.0%) | 706 (44.0%) | 325 (20.3%) |
| Miss | 164 (10.3%) | 350 (22.1%) | 581 (36.7%) | 490 (30.9%) |
| Say | 120 (7.5%) | 341 (21.2%) | 706 (43.9%) | 441 (27.4%) |
| Depend | 102 (6.3%) | 207 (12.8%) | 594 (36.8%) | 711 (44.1%) |

Note: Percentages are the valid percentages after removing missing cases

### 3.2. IRT Based Findings

The item parameters for the five items under the generalized partial credit model are given in Table 4, and the item response category characteristic curves that show the probability of selecting each response option are shown in Figure 3. By looking at the Table 4, we saw that $b_1$ parameter (e.g., first step parameter) was the highest for the important item, meaning that the chance of feeling "a little important to other people" for a person who scored "not at all" was more difficult compared to items. This also means that someone with $\theta$= -1.06 has an equal chance of picking the response category 1 (Not at all) and response category 2 (A little) on this item. In other words, any student with ability score of less than $\theta$ = -1.06 will more likely pick "not at all", higher than $\theta$ =-1.06 will more likely pick "a little". Whereas, it was the lowest for the say item, meaning that someone with $\theta$ = -1.86 has an equal chance of picking "not at all" and "a little". As told before, this is because the corresponding value of the ability scale at the interaction point of the curves for these two options was -1.86 for the say item. Thus, we can conclude that stepping from the response option "not at all" to "a little" is easier or requires less amount of general mattering ability score on the say item than any other items. The $b_1$ parameters for the remaining items can be interpreted similarly.

**Table 4.** *Item parameters produced by the Generalized Partial Credit Model*

| | Step Parameter | | | Discrimination |
|---|---|---|---|---|
| Item related to the | $b_1$ | $b_2$ | $b_3$ | $a$ |
| Important | -1.06 | -0.64 | 1.24 | 1.64 |
| Attention | -1.81 | -0.60 | 1.17 | 1.16 |
| Miss | -1.46 | -0.65 | 0.53 | 1.37 |
| Say | -1.86 | -0.94 | 0.81 | 1.10 |
| Depend | -1.79 | -1.66 | -0.05 | 0.83 |

When we looked at the transitions from the option "a little" to the option "somewhat", we saw that $b_2$ parameter was the lowest for the depend item, meaning that someone with $\theta$ =-1.66 has an equal chance of picking both middle categories on this item. When the latent ability score is higher than $\theta$ =-1.66, the likelihood of picking "somewhat" increases. Whereas, it was the highest for the attention item, meaning that someone with a theta of -0.60 has an equal chance

of picking the both categories. More specifically, any person with ability score of less than $\theta$ = -0.60 (but less than $\theta$ =-1.81 because $b_1$ for this item is –1.81) would more likely pick the response option of a little. To sum up, the chance of stepping from a little to somewhat requires less and higher amount of general mattering ability score on the depend and attention items, respectively.

We found that $b_3$ parameter was the lowest for the depend item, and highest for the important item. This means that the transitions from "somewhat" to "very much" for students who already selected "somewhat" requires lowest and highest general mattering latent trait score for the depend and important items, respectively.

We also noticed that the probability of selecting "a little" was the lowest on any point of the ability scale for the depend item (e.g., the curve numbered 2 was always under the other curves on any points of the scale). This means that regardless of the amount of general ability score, students always tented to select other three options on the depend item. This was also true for the other items. This is because the likelihood of selecting the option "a little" was higher just for a limited range of ability scores (i.e., the curve numbered 2 was above the other curves across a short range on the ability scale). Thus, we can conclude that the option "a little" was the least efficient option, especially for the depend item, and regardless of the magnitude of the general mattering ability score, the students always tended select the other options.

In terms of item discriminations, the best discriminating item was important item, and the worst discriminating item was depend item. When we look at the Figure 3, it is clear that item category curves were steeper on the important item, and flatter on the depend item.

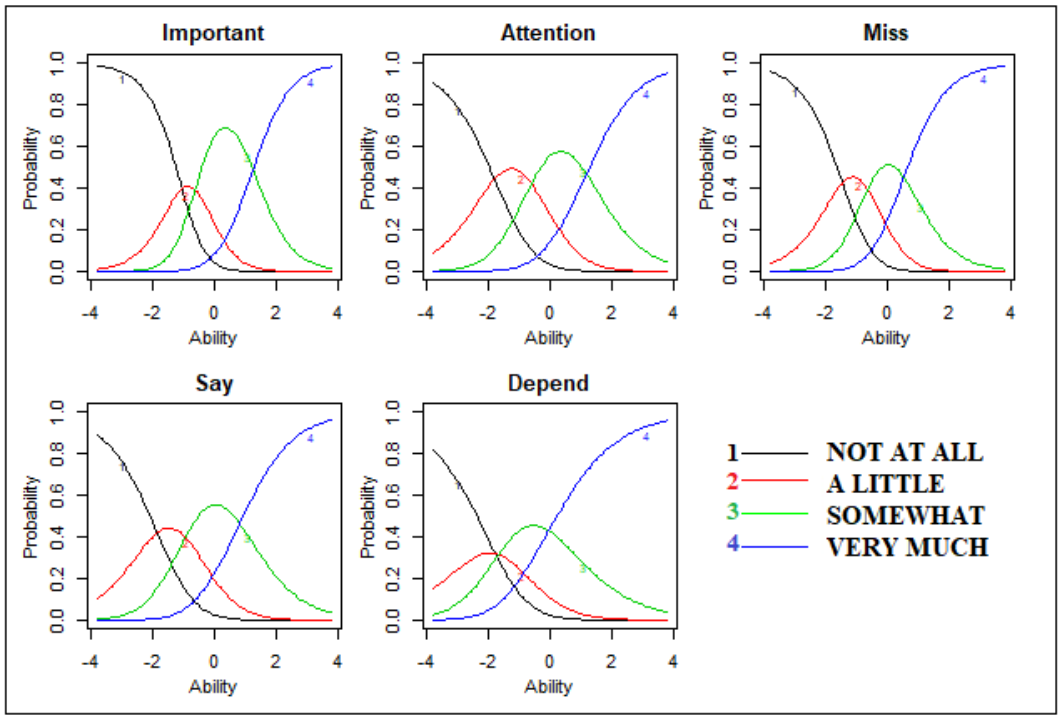



**Figure 3.** Item response category characteristic curves for all items.

We also provided item and test or scale level information function that shows (1) the amount of information an item or a scale provides across the different points on the ability scale, and (2) the point where the information peaks, and showed in Figure 4. Based on the Figure 4, the first finding was that item information function always peaks around the theta point of -1 for all items. This means that the test items were more appropriate for the students having ability

score of -1. In other words, the latent ability scores were more precisely or accurately calculated for the students with the ability level of around -1. The test level information chart in the same figure also supports this. The second finding was that the depend item provides the lowest information, and important item provides the highest amount of information for the students across the ability levels. This means that the depend and important items were the worst and best items, respectively, to be able to measure general mattering ability with a less amount of measurement error.

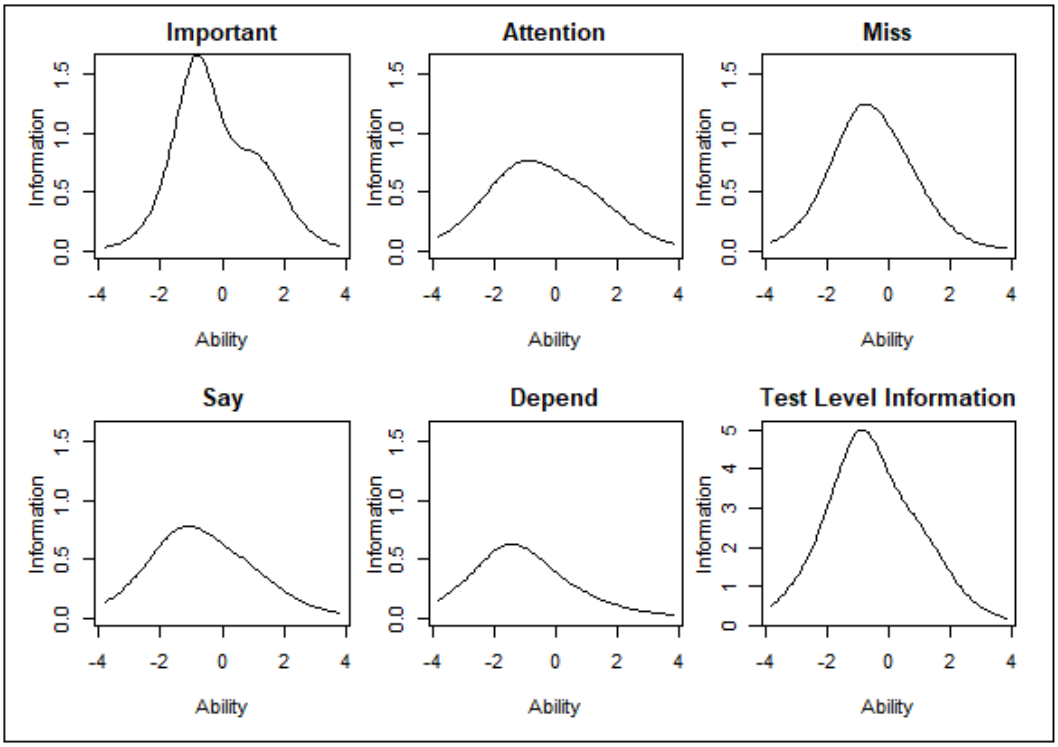



**Figure 4.** Total item information functions for the five items.

## 4. DISCUSSION

More likely due to computational easiness, classical test theory is more commonly used by the psychologists and counselors. However, because of producing sample dependent outcomes, it does not always allow researchers and practitioners to generalize findings to other facets. Furthermore, due to its test-focused feature (e.g., focusing total score), CTT does not tell much about item characteristics in detail and how persons are well on specific items. Whereas, item response theory can easily overcome these problems and can provide more meaningful and concrete interpretations (de Ayala, 2009). However, there is a lack of item response theory applications in the area of psychological measurement. This study showed an application of both classical test theory and item response theory applications to the counseling data, and presented the inferences made by the two.

Based on the CTT findings, for all items, students generally selected the second and third response options (e.g., a little and somewhat), but always the lowest percentage of the students picked the first response option (e.g., not at all). The distribution of selected response options was relatively closer on the important item which made it the most discriminating item, but varied on the depend item which led it to have the lowest discrimination parameter. Thus, it is safe to conclude that the in terms of the contribution to the scale for a better measurement, the depend item was the least effective item.

Item response theory also found that that the depend item was the most problematic item in the scale. This is because, this item provided the lowest amount of item information across the different ability levels, and had the lowest discrimination power. Moreover, the chance of reaching the highest level of response option (e.g., very much) was much easier on this item which was inconsistent with the other items. We believe that it is interesting to see that students at this age easily believe that people depend on them very much. There could be two possible explanations of this finding. First, from a psychological perspective, this means that it was more significant to be considered by others than someone else being depend on them. This was consistent with previous research conducted by Karaman, Balkin, and Juhnke (2018). Researchers conducted a life balance-related study with 453 Turkish participants, and indicated that there was a positive relationship between feeling important by others and a balanced life. In other words, participants concerned whether others cared about them or not. Second explanation of believing people depend on them very much could be that this was more likely because students in the current study did not have a better understanding of this item, and had difficulty interpreting. In order to eliminate the possible confusion, it would be beneficial to play with the structure of the item or to provide additional explanation in a parenthesis. On the other hand, the best item was the important item due to providing the highest amount of information and having the highest discrimination power. Also, the scale was more effective to measure students having general ability score of around -1. However, it was not easy to make these inferences by the CTT.

Furthermore, the IRT detected that the option “a little” is the least effective option, especially for the depend item. This was due to the fact that, the item category response curve for a little response option always stayed under the other response curves. Hence, the students always tended to select the other options, regardless of the magnitude of latent trait score. This might be because of that the students did not distinguish between the response options “a little” and “somewhat” when answering the items. We have concluded that this response option should be revised or removed.

This study does not argue against CTT-based calculations. Rather, this work simply shows how the meaning of item parameters would differ in two measurement approaches and how these differences may yield different interpretations. We aim to encourage psychologists and counselors to give more consideration to the item response theory applications when assessing the psychometric features of the items in the scales in the educational and counseling areas, and draw their attention to the IRT. This is because; as shown in the paper, IRT would provide more information about the items and response options. As the final note, the choice of measurement approach could be particularly very important when validating a psychometric tool.

In this work, we run GPCM only, but other studies should also examine properties of the items under other polytomously scored item response theory models using direct approach such as direct PCM or GRM.

## ORCID

Halil İbrahim SARI http://orcid.org/0000-0001-7506-9000

## 5. REFERENCES

Arbuckle, J. L. (2014). *Amos 23.0 user's guide.* Chicago, IL: IBM SPSS.

Baker F. (1992). *Item response theory*. New York, NY: Marcel Dekker, INC.

Brodin, U., Fors, U., & Laksov, K. B. (2010). The application of item response theory on a teaching strategy profile questionnaire. *BMC medical education, 10*, 14-14-doi:10.1186/1472-6920-10-14

Cha, M. (2016). The mediation effect of mattering and self-esteem in the relationship between socially prescribed perfectionism and depression: Based on the social disconnection model. *Personality and Individual Differences, 88*, 148-159. doi:10.1016/j.paid.2015.09.008

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment, 27*, 1241–1252. doi:10.1037/pas0000132

Connolly, K. M., & Myers, J. E. (2003). Wellness and mattering: The role of holistic factors in job satisfaction. *Journal of Employment Counseling, 40*, 152-160. doi:10.1002/j.2161-1920.2003.tb00866.x

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

de Ayala, R. J. (2009*). The theory and practice of item response theory*. New York, NY: Guilford.

Demir, M., Özen, A. & Doğan, A. (2012). Friendship, perceived mattering and happiness: A study of American and Turkish college students. *The Journal of Social Psychology*, *152*, 659-664. doi: 10.1080/00224545.2011.650237

Dodeen, H., & Al-Darmaki, F. (2016). The application of item response theory in developing and validating a shortened version of the Emirate Marital Satisfaction Scale. *Psychological Assessment*, *28*, 1625-1633. doi:10.1037/pas0000296.supp

Embretson S.E, & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, IN: Lawrence Erlbaum Associates, Inc.

Flett, G. L., Galfi-Pechenkov, I., Molnar, D. S., Hewitt, P. L., & Goldstein, A. L. (2012). Perfectionism, mattering, and depression: A mediational analysis. *Personality and Individual Differences, 52*, 828-832. doi:10.1016/j.paid.2011.12.041

Gomez, R., & Vance, A. (2015). Item response theory properties of the internalizing disorders in adolescents. *Journal of Childhood & Developmental Disorders, 1*, 1-10. doi:10.4172/2472-1786.100006

Haktanir, A., Lenz, A. S., Can, N., & Watson, J. C. (2016). Development and evaluation of Turkish language versions of three positive psychology assessments. *International Journal for the Advancement of Counselling, 38*, 286-297. doi:10.1007/s10447-016-9272-9

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*, II28-II42. doi:10.1097/00005650-200009002-00007

Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in M Plus. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 308-318. doi:10.1080/10705511.2014.937374

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

Karaman, M. A., Balkin, R. S., & Juhnke, G. A. (2018). Turkish adaptation of the Juhnke–Balkin Life Balance Inventory. *Measurement and Evaluation in Counseling and Development, 51*, 141–150. doi:10.1080/07481756.2017.1308226

Lee, Y. S., Krishnan, A., & Park, Y. S. (2012). Psychometric properties of the Children's Depression Inventory: An item response theory analysis across age in a nonclinical, longitudinal, adolescent sample. *Measurement and Evaluation in Counseling and Development, 45*, 84-100. doi: 10.1177/0748175611428329

Locke, B. D., McAleavey, A. A., Zhao, Y., Lei, P. W., Hayes, J. A., Castonguay, L. G., ... & Lin, Y. C. (2012). Development and initial validation of the Counseling Center Assessment of Psychological Symptoms–34. *Measurement and Evaluation in Counseling and Development*, *45*, 151-169. doi: 10.1177/0748175611432642

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Marcus, F. M. (1991). *Mattering: Its measurement and theoretical significance for social psychology*. Paper presented at the annual meeting of the Eastern Sociological Association, Cincinnati.

Masters, G. (1982). A Rasch model for partial credit scoring. Psychometrika, *47*, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Rayle, A. D. (2006). Mattering to others: Implications for the counseling relationship. *Journal of Counseling & Development, 84*, 483-487. doi:10.1002/j.1556-6678.2006.tb00432.x

R Core Team (2016). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Riemer, M., & Kearns, M. A. (2010). Description and psychometric evaluation of the Youth Counseling Impact Scale. *Psychological Assessment*, *22*, 259-268. doi:10.1037/a0018507

Rizopoulos, D. (2017). Package "ltm", Latent Trait Models under IRT. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Rosenberg, M., & McCullough, B. C. (1981). Mattering: Inferred significance and mental health among adolescents. *Research in Community & Mental Health, 2*, 163-182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*, 1–97. doi:10.1007/bf03372160

Tasca, G. A., Cabrera, C., Kristjansson, E., MacNair-Semands, R., Joyce, A. S., & Ogrodniczuk, J. S. (2016). The therapeutic factor inventory-8: Using item response theory to create a brief scale for continuous process monitoring for group psychotherapy. *Psychotherapy Research*, *26*, 131-145. doi:10.1080/10503307.2014.963729

Tovar, E., Simon, M. A., & Lee, H. B. (2009). Development and validation of the college mattering inventory with diverse urban college students. *Measurement and Evaluation in Counseling and Development, 42*, 154-178. doi:10.1177/0748175609344091

Tovar, E. (2013). *A conceptual model on the impact of mattering, sense of belonging, engagement/involvement, and socio-academic integrative experiences on community college students' intent to persist* (Doctoral Dissertation). Retrieved from CGU Theses & Dissertations (Paper 81). doi: 10.5642/cguetd/81

Turner, B. M., Betz, N. E., Edwards, M. C., & Borgen, F. H. (2010). Psychometric examination of an inventory of self-efficacy for the Holland vocational themes using item response theory. *Measurement and Evaluation in Counseling and Development*, *43*, 188-198. doi:10.1177/0748175610384810

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*, 1-10. doi:10.1186/s41155-016-0040-x

## APPENDIX

The following five questions are designed to measure the degree to which you believe that you matter to others. Please circle the appropriate response for what YOU believe.

| **GENERAL MATTERING SCALE** | **NOT AT ALL** | **A LITTLE** | **SOMEWHAT** | **VERY MUCH** |
|---|---|---|---|---|
| **1. How *important* do you feel you are to other people?** | **1** | **2** | **3** | **4** |
| 2. How much do you feel other people pay *attention* to you? | **1** | **2** | **3** | **4** |
| **3. How much do you feel others would *miss* you if you went away?** | **1** | **2** | **3** | **4** |
| 4. How interested are people generally in what you have to *say*? | **1** | **2** | **3** | **4** |
| **5. How much do people *depend* on you?** | **1** | **2** | **3** | **4** |

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 682–700

DOI: 10.21449/ijate.472185

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Automating Simulation Research for Item Response Theory using R

**Sunbok Lee** [1,*], **Youn-Jeng Choi** [2], **Allan S. Cohen** [3]

[1] Department of Psychology, University of Houston, Houston, TX 77204
[2] Department of Educational Studies in Psychology, Research Methodology & Counseling, The University of Alabama, Tuscaloosa, AL 35487
[3] Department of Educational Psychology (Quantitative Methodology), University of Georgia, Athens GA 30602

**ARTICLE HISTORY**

*Received: 22 August 2018*
*Revised: 12 October 2018*
*Accepted: 17 October 2018*

**KEYWORDS**

IRT, Simulation, R

**Abstract:** A simulation study is a useful tool in examining how validly item response theory (IRT) models can be applied in various settings. Typically, a large number of replications are required to obtain the desired precision. However, many standard software packages in IRT, such as MULTILOG and BILOG, are not well suited for a simulation study requiring a large number of replications because they were developed as a stand-alone software package that is best suited for a single run. This article demonstrated how built-in R functions can be used to automate the simulation study using the stand-alone software packages in IRT. For a demonstration purpose, MULTILOG was used in the example codes in the appendices, but the overall framework of a simulation study and the built-in R functions used in this article can be applied for a simulation study using other stand-alone software packages as well.

## 1. INTRODUCTION

Item response theory (IRT) provides a family of statistical models that establish the correspondence between item responses and latent variables (De Ayala, 2009). When applying IRT models to real datasets, researchers often want to know how validly their IRT models can be applied to their datasets at hand (Harwell, Stone, Hsu, & Kirisci, 1996). For examples, researchers may concern the small sample sizes or multidimensionality of their datasets. Analytic solutions can provide exact and rigorous answers to those questions, but may not be always tractable because of the complexity of the problem. In such a case, a simulation study can be a useful alternative.

Typically, a simulation study requires a large number of replications to obtain the desired precision. However, traditional standard software packages in IRT, such as MULTILOG (Thissen, Chen, & Bock, 2003) and BILOG (Zimowski, Muraki, Mislevy, & Bock, 1996), are not well suited for a simulation study requiring a large number of replications because they were developed as stand-alone software packages for a single run. Running such stand-alone software packages requires input command files, which make it difficult to run those software

CONTACT: Sunbok Lee slee95@Central.UH.EDU Department of Psychology, University of Houston, Houston, TX 77204

**ISSN-e: 2148-7456 /© IJATE 2018**

packages a large number of times. More importantly, it is very difficult to extract the values of interest, such as parameter estimates and fit indices, from the output files. It would be an extremely time-consuming task to manually extract the values from a large amount of output files. Therefore, for a simulation study using stand-alone software packages, it is necessary to automate the simulation procedure.

Recently, many R packages, such as LTM and MIRT, have been developed for IRT. With those packages, a simulation study can be made much simpler by repeatedly running IRT estimation functions within a loop and also extracting values of interest from R objects. However, there is still need for conducting a simulation using stand-alone IRT software packages. Those stand-alone IRT software packages, such as MULTILOG and BILOG, have been widely used in psychometrics, and still are considered as standard software packages by many researchers. Although there are other options such as R and Mplus, MULTILOG and BILOG are still widedly used for IRT simulation studies on various topics such as linking (Kim & Lee, 2006), goodness of fit statistics (Stone, 2000), initial valud and convergence criterian (Nader, Tran & Voracek, 2015), speicification of ability distribution and numerical integration (Kim & Lee, 2017), and structure zeros (Kim, Brennan, & Lee, 2017). Moreover, when researchers want to develop new estimation algorithms, they may want to compare the results from their new algorithms with those from traditional stand-alone software packages. Given the typical recommendation for the large replications, using the standalone software packages such as MULTILOG and BILOG for simulation studies require us to automatically run the software packages a large number of times and also automatically extract necessary information from a large number of output files. The purpose of this paper is to demonstrate how the built-in functions in R software package (R Core Team, 2015) can be used to automate a simulation study using standard stand-alone software in IRT. The procedure for a simulation study can be divided into four parts: generating item responses, preparing input command files, running stand-alone software, and extracting statistics from output files. Some functions in R that are useful in automating each part of a simulation study are introduced. For a demonstration purpose, MULTILOG is used in the example code in the appendices, but the framework of the simulation study and the R functions used in this article can also be applied for a simulation study using other stand-alone software.

## 2. A SIMULATION STUDY IN IRT

A simulation study is a computer intensive procedure to evaluate statistical methods, given a known model and parameters. Most frequently, this is done as a Monte Carlo simulation study in which random numbers are used to generate data to be analyzed with the model. In a simulation study, the random samples thus generated enable us to create an empirical sampling distribution of the parameters under the conditions in which researchers are interested (Bandalos, 2006). The empirical sampling distribution can be used in evaluating the estimating procedures (e.g., maximum likelihood estimation) or properties of the statistics (e.g., goodness of fit). Harwell et al. (1996) provides a useful reference for a simulation study in IRT, illustrating advantages, limitations, and major steps of IRT simulation. The simulation steps consist of formulating a problem, designing a simulation study, generating item responses and estimating model parameters, and analyzing the results of a simulation. In this paper, we present how some steps of a IRT simulation study can be automated using the R software package.

In most simulation studies, handling a large number of replications is a typical problem. Use of stand-alone software, such as MULTILOG and BILOG, to estimate IRT parameters can pose some additional difficulties that are not present when implementing the whole simulation procedures within a single platform such as C, FORTRAN, or R. Within a single programming language, the intermediate results produced in a given step can be easily used by the following

step of the simulation. For example, in a simulation study in which the whole process is implemented within R, the generated data sets can be saved as array variables which are then passed to the next step in the process. Using the generated data sets in the later step of the simulation is just a matter of referencing these variables within R. The estimated parameters are also saved as array variables in R, making it easy to summarize the results of parameter estimation.

In contrast to using a single platform of programming language, using stand-alone software for a simulation study poses some practical problems, as such software is typically designed for a single run in which the input command file needs to be provided by the user. Given an input command file, results are usually saved as a text file. In order to obtain the desired number of replications using stand-alone software, it is necessary to prepare data and input command files for as many replications as needed. Furthermore, when the subsequent output files are produced, statistics of interest such as parameter estimates, standard errors, and fit indices need to be extracted from output files and organized into structured data for summary. Considering the large number of replications required for a simulation study, it is not normally practical to handle these procedures manually. In the following sections, we describe how some built-in functions in the R software package can be used to automate the simulation study using stand-alone IRT software.

## 3. R FOR THE SIMULATION STUDY IN IRT

R is a freely available open source language for statistical computing (R Core Team, 2015) and can be used to automate the simulation using stand-alone software. R has built-in functions which can manipulate character strings and run external stand-alone software. R also supports the regular expressions which are useful for extracting the statistics of interest from output files. These functions of R will be discussed in this section, and the example code in the appendices that was written using these functions will be discussed in detail in the next section.

### 3.1. Functions for String Manipulations in R

Stand-alone software is typically developed for a single run in which data and input command files need to be prepared by a user. In order to obtain the desired number of replications, stand-alone software should be run as many times as the number of replications using data and input command files specific to each replication. Therefore, to automate the procedure for preparing those files, file names and some commands in the input command file need to be automatically modified from replication to replication. In R, paste() and strsplit() are functions for string manipulations and can be used to modify the character strings for file names and commands. The gsub() is also useful in replacing a string in a text.

*paste()*. In R, the paste() function concatenates or combines an arbitrary number of arguments to form a combined string after converting each argument to a character string. For example, in the following R code, the paste() function combines character string ‘item’ and the number 7 to form the string ‘item7’. The sep option is used to specify the character string to separate the arguments. In the following example, the sep option is used to indicate no space by placing two double quotes together, with no space between them:

```
> paste("item", 7, sep="") [1] "item7"
```

A character vector which contains item names can be generated easily by using a numeric vector. In the example code below, the numeric vector is generated by 1:7.

```
> paste("item", 1:7, sep="")
```

```
 [1] "item1" "item2" "item3" "item4" "item5" "item6" "item7"
```

The paste() function is useful for changing text, such as file names and commands, in command files. The following example shows how the data file name can be changed from replication to replication using the paste() function within a for loop.

```
for (replications in 1:1000) {
      ...
filename <- paste("MULTILOGI20S100R", replications,
            ".dat", sep="")
}
```

In the code above, replications is a for loop index variable that varies from replication to replication. If replications changes from 1 to 1000 , the variable filename will be changed from 'MULTILOGI20S100R1.dat' to 'MULTILOGI20S100R1000.dat'.

*strsplit()*. The strsplit() function splits a character string into substrings using a delimiter specified in the split option. It provides a convenient way for users to separate the file extension from the whole file name. Since a simulation using stand-alone software generates a large number of data, input command, and output files, it is necessary to manage a large number of files, each with separate names. In the examples in this paper, we have adopted the following naming rule: The body of the whole file name remains the same and only the file extension is changed for each of the files needed for each replication. For example, in the data file name 'MULTILOGI20S100R1.dat', MULTILOG is the body of the file name and does not change. I20 is used to indicate 20 items, S100 is used to indicate 100 subjects, and R1 is used to indicate the 1st replication. MULTILOG requires the file extension '.mlg' to indicate the input command file to the software. So, the name for the input command file for a 20-item test with 100 examinees would be 'MULTILOGI20S100R1.mlg'. In the R code below, the strsplit() function splits a character string 'MULTILOGI20S100R1.dat' into two substrings using a dot as a delimiter to form a character vector containing two elements, 'MULTILOGI20S100R1' and 'dat'.

```
> strsplit("MULTILOGI20S100R1.dat", split=".", fixed=T)

[1] "MULTILOGI20S100R1" "dat"
```

Combined with the paste() function discussed above, the strsplit() function can be used to change only the file extension of the whole file name. In the code below, the unlist() function is used to convert the list data type produced by the strsplit() function to the vector data type, and '[1]' is used to index the first element of a character vector, which is 'MULTILOGI20S100R1'. In all, the following R code can be used to replace the file extension '.dat' with '.mlg'.

```
> paste(unlist(strsplit("MULTILOGI20S100R1.dat",

+    split=".",fixed=T))[1],".mlg",sep="")

[1] "MULTILOGI20S100R1.mlg"
```

*gsub()*. The gsub() function replaces a string in a vector. More specifically, gsub(pattern, replacement, x) finds the string pattern in a vector x and replaces the string pattern with another

string replacement. For example, in the following example, gsub() replaces ‘apple’ with ‘orange’ in the vector x.

```
> x <- "I like apple"
> gsub("apple", "orange", x)
[1] "I like orange"
```

In a simulation study, the gsub() function can be used to prepare an input command file specific to each replication using a template file. For example, in MULTILOG, the number of examinees needs to be provided in an input command file. Suppose that a template file to generate the input command file contains the following line specifying the number of examinees: ‘NEXAMINEES =hnSubjectsi’ . Then, the R command below will replace hnSubjectsi with the specific number, 10:

```
> x <- "NEXAMINEES = <nSubjects>"
> gsub("<nSubjects>", 10, x)
[1] "NEXAMINEES = 10"
```

### 3.2. Functions for Executing External Software

Once data and input command files are prepared, researchers need to run stand-alone software the desired number of times to generate the output files. Running stand-alone software in batch mode is a useful way of automating such a process. In batch mode, users can run software without manual intervention by simply providing a command at the DOS prompt. For example, the following command at the DOS prompt will execute MULTILOG using ‘MULTILOGI5S500R9.mlg’ as an input command file. Note that the file extension need to be omitted. Also, progra~1 is the short name for ‘Program Files’ folder in the Windows operating system.

```
C:\progra~1\MULTILOG> mlg MULTILOGI5S500R9
```

Since our goal is to automate the procedure for running external software, the command above needs to be provided to the DOS operating system using R by changing the input command file from replication to replication. This can be done using the system() function in R, which can execute the system command from R. The following R code executes MULTILOG using the ‘MULTILOGI5S500R9.mlg’ as an input command file. Note that, in R, the double backslash represents backslash.

```
> system("C:\\Progra~1\\MULTILOG\\mlg MULTILOGI5S500R9")
```

### 3.3. A Regular Expression in R

Given a large number of output files, the key in the next step of the simulation is to extract the statistics of interest from the output files and to organize them into structured data. In order to extract specific information from the output files, it is often necessary to use a more sophisticated expression than a simple string. A regular expression is a powerful tool for extracting the statistics of interest from output files.

**An Example of the Regular Expression**. Suppose that a researcher is using BILOG for a simulation, and the following is the part of a BILOG phase 2 output file named ‘BILOGI20S200F2.PH2’.

```
  ITEM      INTERCEPT    SLOPE    THRESHOLD   LOADING   ASYMPTOTE   CHISQ
              S.E.       S.E.       S.E.       S.E.       S.E.     (PROB)
-------------------------------------------------------------------------


ITEM0001 | 0.292    |  2.090   | -0.140   |  0.902  | 0.000  |    4.1
         | 0.187*   |  0.339*  |  0.088*  |  0.146* | 0.000* | (0.3879)
         |          |          |          |         |        |
ITEM0002 | 0.043    |  0.802   | -0.053   |  0.626  | 0.000  |    2.9
         | 0.154*   |  0.180*  |  0.192*  |  0.140* | 0.000* | (0.8175)
         |          |          |          |         |        |
...
```

A researcher might want to extract thresholds or difficulty parameters from the output file. In the file above, the thresholds of item 1 and item 2, which are labeled ITEM0001 and ITEM0002, are -0.140 and -0.053 respectively. One way of extracting those numbers from the output files could be the following: i) extracting the lines that contain ITEM0001, ITEM0002, etc., and ii) extracting the numbers in the fourth column from those lines. In identifying the lines for extraction, it would be impractical to use a simple string such as ITEM0001, ITEM0002, etc. because a researcher needs to specify each string representing each item. Rather, it would be more practical if a researcher can specify the string for matching more flexibly like ITEMdddd, where d represents a single digit number. Also, in extracting thresholds, it would be convenient if there is a single expression representing any floating numbers to match -0.140 and -0.053.

A regular expression is the sequence of literal and special characters that describes a set of strings and provides a flexible means for matching strings in a text. In the example above, the lines that contain ITEM0001, ITEM0002, etc. can be matched using the regular expression ITEM[0-9]{4} in which [0-9] matches any single digit, and {4} matches 4 occurrences of the preceding element. Also, in R, the regular expression '[-+]?[0-9]+(\\.[0-9]+)?' matches a floating number. In a regular expression, the square bracket [] is used to match any single character listed within the bracket. For example, '[hs]eat' matches 'heat', 'seat', etc. and '[0-9]' match any single digit. Therefore, the leading [-+] allows a plus or minus sign. Since ? matches the preceding element zero or more times, a number with no sign will also be matched. [0-9]+ matches one or more digits since + matches the preceding element 1 or more times. The double backslash \\ is used for the escape sequence. In programming languages including R, many characters are reserved to represent a special meaning. Sometimes, however, users might want to use those special characters as literal strings. The dot is escaped from the special meaning by using the double backslash and just represents the decimal point in a floating number. ()? makes \\.[0-9]+ optional. In all, the regular expression, '[-+]?[0-9]+(\\.[0-9]+)?', matches any floating number such as -0.140 and -0.053. The actual R code for extracting the thresholds from 'BILOGI20S200F2.PH2' will be presented later in this section with some additional functions in R using the regular expression.

**The Regular Expression in General**. A regular expression is the sequence of literal and special characters that is used to match a specific string in a text. By using a regular expression, a user can locate the specific lines in output files by matching the string described in the regular expression with a string in the lines in a file. This, in turn, enables a user to extract the specific information matching that regular expression from unstructured data sources. Simply, the regular expression is a sequence of characters that forms a search pattern for matching. For example, the regular expression 'a.c' matches 'aac', 'abc', etc. since a dot character in a regular expression matches any single character. Also, the regular expression 'ab+' matches 'abb', 'abbb', etc. since the + character in a regular expression matches the preceding element, which is 'b' in this example, one or more times.

The regular expression is supported by many programming languages (e.g., R, SAS, C++, and Java) and command line utilities in UNIX (e.g., grep, sed, and awk) because of its usefulness in many applications. The wide applicability of the regular expression comes from its flexibility in matching a specific string of text using special characters. Some frequently used special characters are introduced with examples in Table 1. For more detailed discussion on the regular expression, readers are referred to other comprehensive literatures (Friedl, 2006; Spector, 2008).

**Table 1.** Special Characters in a Regular Expression

| Character | Meaning | Notes |
|---|---|---|
| . | Matches any single character | 'a.c' matches 'abc', 'acc', etc. |
| ^ | Start of string or line | '^app.e' matches apple but only at the beginning of the string or line. |
| $ | End of string or line | 'app.e$' matches apple but only at the end of the string or line. |
| [ ] | Matches any single character that is listed within the brackets | '[hs]eat' matches heat, seat, etc. |
| [^] | Matches any single character that is not listed within the brackets | [^hs]eat matches neat, beat, etc. |
| ? | Matches the preceding element 0 or 1 times | 'ab?' matches 'a' or 'ab', etc. |
| * | Matches the preceding element 0 or more times | 'ab*' matches 'a' or 'ab' or 'abb', etc. |
| + | Matches the preceding element 1 or more times | ab+ matches ab or abb or abbb, etc. |
| \w | Alphanumeric characters | Equivalent to [A-Za-z0-9] |
| \W | Non alphanumeric characters | Equivalent to [^A-Za-z0-9] |
| \d | Digits | Equivalent to [0-9] |
| \D | Non-digits | Equivalent to [^0-9] |
| \s | White space characters | Matches space, tab, etc. |
| \S | Non white space characters | Matches anything except white space |

R has several built-in functions that can be used in extracting the statistics of interest from output files using the regular expression. In the following, examples of the regular expression and functions in R supporting the regular expression will be presented.

*grep()*. The grep() function searches a file line by line and returns only those lines that match a particular pattern. The pattern is allowed to be a simple character string or the regular expression. For example, the following R command extracts only the lines that contain

ITEM0001, ITEM0002, etc. from the BILOG output file named 'BILOGI20S200F2.PH2' in the previous example.

```
> text <- readLines("BILOGI5S500R1.PH2")

> lines <- grep("ITEM[0-9]{4} \\|", text, value=TRUE)

> lines

 [1] " ITEM0001 | 0.292 |    2.090 | -0.140 |     0.902 |    0.000 | "
 [2] " ITEM0002 | 0.043 |    0.802 | -0.053 |     0.626 |    0.000 | "
 ...
```

In the example code above, the readLines() function reads 'BILOGI20S200F2.PH2' line by line and saves those lines in the character vector named text. Then, the grep() function returns the lines containing a character string that match the pattern described by the regular expression "ITEM[0-9]{4} \\|" from the character vector text and saves those lines in the character vector lines.

*str_extract_all()*. The str_extract_all() function extracts all pieces of a string that match a pattern. The user will need to ensure that the stringr package is installed to use this function. Given the vector named lines in the previous example, the floating numbers in the vector lines can be extracted using the following str_extract_all() function:

```
> numbers <- str_extract_all(lines, "[-+]?[0-9]+(\\.[0-9]+)?}")

> numbers

[[1]]

[1] "0.292" "2.090" "-0.140" "0.902" "0.000" "4.1"  "4.0"

[[2]]

[1] "0.043" "0.802" "-0.053" "0.626" "0.000" "2.9"  "6.0

…
```

Where [-+]?[0-9]+(\\.[0-9]+)? is the regular expression representing a floating number. Note that the result from the str_extract_all() function is a list object which is one of the data types in R. A list object has elements, each of which can contain any type of object in R. Since a matrix is easier to handle, the do.call() function is used to convert the list object to the matrix object as follows:

```
> numbers <- do.call("rbind", numbers)

> numbers

      [,1]    [,2]    [,3]     [,4]    [,5]    [,6]  [,7]

[1,] "0.292" "2.090" "-0.140" "0.902" "0.000" "4.1" "4.0"

[2,] "0.043" "0.802" "-0.053" "0.626" "0.000" "2.9" "6.0"

...
```

The do.call() function executes a function, which is the rbind() here, using a list of arguments to be passed to it. Since the rbind() combines vectors by rows, the do.call("rbind", numbers) command converts the list object numbers to a matrix object.

## 4. EXAMPLE CODE

In the previous section, we have discussed some useful functions of R that can be used to automate the simulation using stand-alone software. This section presents an example of a simple simulation study using those functions. The R codes for this example are presented in the appendices. In the example, item responses were generated using two-parameter logistic (2PL) model (Appendix A), input command files for MULTILOG were generated using a template file for the input command file (Appendix B), MULTILOG was automatically run as many replications as needed (Appendix C), and item parameters of 2PL were extracted and summarized using functions in R (Appendix D).

### 4.1. Setup and Data Generation

Appendix A provides the R codes that generate data sets for a simple example simulation study. The following information is required for this simulation: the number of items, the number of subjects, the number of replications, and the true item parameters for the 2PL model. All the files and results from this simulation will be saved in the folder specified by mF variable. The user also needs to specify the folder where MULTILOG is installed using multilogPath variable.

Item responses are generated in lines 29-39. The ability of each person is sampled from a standard normal distribution, and the probability of getting an item correct is calculated based on the equation for 2PL. Then, given a random number between 0 and 1 that is sampled from a uniform distribution, the item response for an item of a person becomes 1 if the sampled uniform random number is less than the probability of getting an item correct for a person and becomes 0 otherwise.

### 4.2. Preparing Input Command Files for MULTILOG

A template file and R codes for preparing input command files for MULTILOG are provided in Appendix B. A template file is used to generate the input command file for each replication by changing some parts of the template file to be specific for each replication. More specifically, a template file named ‘template.mlg’ contains MULTLOG commands for 2PL. Some parts of the template file are not specified and just marked using angle brackets such as <datafile> and <nItems>, which will be replaced with appropriate character strings using the gsub() function.

In line 5 of the R code in the second part of Appendix B, the list.files() function is used to read all the data file names in the main folder (mF). In line 11, each line of the template file is saved into text variable using readLines() function. In line 13, the name of data file for a specific replication represented by rep is saved into the variable dName. Then, in line 14, the string <datafile> in the template file is replaced with the specific file name, i.e., dName, for a given replication using the gsub() function.

```
5  dataFiles <- list.files(pattern = ".dat", file.path(mF))

...

11 text <- readLines("C:/template/template.mlg")

...

13 dName <- file.path(mF,dataFiles[rep])
```

```
14 text <- gsub(pattern = "<datafile>", replacement = dName, x = text)

...
```

The input command file in MULTILOG requires a string representing the format of a data file. In line 23, paste("(5A1,", nItems,"A1)", sep="") creates the string for the format of the data file. In our example, the number of items is 5, and therefore the string (5A1, 5A1) is saved into format variable. In line 24, the string <format> in a template file is replaced by the string saved in format.

### 4.3. Running MULTILOG in R

Given the data and input command files, the code in Appendix C runs MULTILOG as many replications as needed. In the previous section, we have introduced the system() function, and the following command executes MULTILOG using the input command file named 'MULTILOGI5S500R9.mlg':

```
> system("C:\\Progra~1\\MULTILOG\\mlg MULTILOGI5S500R9")
```

In order to repeatedly run MULTILOG many times , the system() function can be used within a for loop by changing the argument of the system() function using the paste() function.

```
...

dataFiles <- list.files(pattern = ".dat", file.path(mF))

...

for (rep in 1:nReplications) {

      system(paste(multilogPath,unlist(strsplit(dataFiles[rep],"\\.
             "))[1]))

}

...
```

In the codes above, the vector object dataFiles contains the names of all data filesin the main folder (mF). The for loop index variable rep changes from 1 to nReplications, and accordingly dataFiles[rep] indicates a different data file name for each replication. The strsplit() function is used to remove the file extension from the data file name, and the paste() function is used to combine the string for the path of MULTILOG and data file name.

### 4.3. Extracting and Summarizing Results

In the codes in Appendix D, item discrimination and difficulty parameters are extracted from output files and organized into structured format. It should be noted that the codes in Appendix D are just one way of extracting and summarizing parameters. Also, codes could be more complicated as the simulation becomes more complex. However, the functions used in this example would be useful in extracting statistics of interest from output files in other cases.

In Appendix D, the lapply() function is used through the codes. The lapply(x,fun), where the x is a list object and fun is any function in R, returns a list object of the same length as the list object x, each element of which is the result that can be obtained by applying the function fun to the corresponding element of x. Because of the implicit iterative nature of the lapply()

function, it can replace the for loop statement in R, which makes a code more simple and clear. The following simple example may be helpful in understanding the lapply() function:

```
> results <- lapply(1:3, function(x) x+1)

> results

[[1]]

[1] 2

[[2]]

[1] 3

[[3]]

[1] 4


> for (x in 1:3) {

      print(x+1)

  }

[1] 2

[1] 3

[1] 4
```

In the above example, the lapply() function picks up the first element of a vector 1:3, which is 1, and uses that element as an argument of the function(x), which is defined as x+1, to yield 2. Then, the result, which is 2 here, is saved as the first element of the list object results. The same procedure is repeated for the second and third elements of a vector 1:3. The same results can be obtained using a for loop. In this way, the lapply() function can replace a for loop, which makes codes more concise. In line 5 in Appendix D, the lapply() function is used to read output files:

```
15 outPut1 <- lapply(list.files(pattern=".OUT"), readLines)
```

In the code above, list.files(pattern=".OUT") creates a list object that contains file names having the file extension '.OUT'. Then, the lapply() function picks up the output file name from the list object one by one and use it as an argument of the readLines() function. In all, outPut1 is the list object, each element of which is the character vector containing each line of an output file as its element. For example, if there are 10 output files, each of which has 100 lines, then outPut1 would be the list object with 10 elements, each of which contains a character vector of size 100.

In line 13, the grep() function is used to extract the lines in output files that contain the discrimination parameters. For example, the following is a part of an output file from MULTILOG which contains the estimates for item discrimination of difficulty parameters:

```
 ITEM  1:   2 GRADED CATEGORIES
       P(#)  ESTIMATE (S.E.)
 A     1     1.06 (0.20)
 B( 1) 2    -1.76 (0.29)
```

Note that the estimates for discrimination and difficulty parameters are located at the second and third lines after the line that contains the string 'GRADED CATEGORIES'. To pick up the estimate for the discrimination parameter, which is 1.06 in this example, the following command in line 13 can be used:

```
13 outPut2<-lapply(outPut1,function(x) x[grep("GRADED CATEGORIES", x)+2])
```

In the code above, the lapply() function picks up each element of the list object outPut1, which is the character vector containing each line of an output file as its element, and uses it as an argument of the function defined by the grep() function. Hence, the argument x in the grep() function will be the character vector containing each line of an output file. Then, grep("GRADED CATEGORIES",x) will return the indices of the lines that matches the pattern 'GRADED CATEGORIES'. Also, grep("GRADED CATEGORIES",x)+2 will return the indices of the second lines after the line that match the pattern, and finally x[grep("GRADED CATEGORIES",x)+2]) will return the character strings in those lines. The following is the last part of the hypothetical results saved in outPut2 assuming 5 items and 100 replications:

```
> outPut2

> outPut2 [[1]]

 [1] " A     1      0.78   (0.19)" " A     3      1.55   (0.22)"
 [3] " A     5      0.53   (0.14)" " A     7      0.67   (0.15)"
 [5] " A     9      0.35   (0.13)"
 ...
 [[100]]
 [1] " A     1      1.47   (0.25)" " A     3      0.67   (0.15)"
 [3] " A     5      0.86   (0.15)" " A     7      0.44   (0.13)"
 [5] " A     9      0.57   (0.15)"
```

Note that each element of the list object outPut2 is the character vector containing the lines that contain estimates for discrimination parameters. Since the results in outPut2 are grouped by replication, it would be more convenient to sort the results by items. The following command in line 15 sorts the results by items:

```
15 outPut3<-lapply(1:nItems, function(x)unlist(lapply(outPut2,function(y)
                   y[x])))
```

Note that the lapply() function is used twice, which is much simpler than the nested for loop. The following is a part of the list object outPut3:

```
...

[[4]]

   [1]  " A   7      0.67   (0.15)" " A     7     0.07 (0.12)"

    ...

   [99] " A   7      0.59   (0.13)" " A     7     0.44 (0.13)"

[[5]]

    [1]  " A   9      0.35   (0.13)" " A     9      0.32   (0.14)"
     ...
    [99]" A   9      0.69   (0.15)" " A     9      0.57   (0.15)"
```

Given the list object outPut3, the code in line 18 will extract the numbers for the estimates of discrimination parameters as follows:

```
18 outPut4 <- sapply(outPut3, function(x)

+ as.numeric(str_extract(x,"[-+]?[0-9]+\\.[0-9]+")))

> outPut4

       [,1] [,2] [,3] [,4] [,5]

[1,]   0.78 1.55 0.53 0.67 0.35

...

[99,]  0.79 1.48 0.89 0.59 0.69

[100,] 1.47 0.67 0.86 0.44 0.57
```

Exactly the same strategy can be applied in obtaining estimates for item difficulty parameters and codes are provided in lines 22-32 in Appendix D. In lines 47-52, the same strategy is used to check whether output files contain the string 'NORMAL PROGRAM TERMINATION', which can provide the information about the convergency of model estimation for each replication

## 5. DISCUSSION

A simulation study is useful in investigating the behavior of IRT models in various settings. For a simulation study in IRT, traditional stand-alone software is often preferred because of its efficiency and reliability. In order to get the desired number of replications, it is almost essential to automate the simulation procedures. In this article, it was demonstrated how the R package can be used to automate a simulation study using stand-alone software in IRT.

Typically, a simulation using stand-alone software can be divided into four parts: generating data sets, preparing input command files, running software, and summarizing output files. Generating data sets is directly related to the research question of interest. In this article, we only presented the most simple case of item response generation. In preparing input command

files and summarizing output files, skills for handling text files are important. The R package provides many useful functions for text processing. In R,

character strings can be combined, split, and replaced using the paste(), strsplit(), and gsub() functions, respectively. Also, the grep(), str_extract_all(), and str_extract() functions are useful in extracting the statistics of interest from output files. In extracting the required information from output files, the regular expression is useful in forming flexible search patterns for matching. The system() function is also useful in running stand-alone software many times from R.

A couple of comments on some practical issues in a simulation study might be helpful. The estimation for a given replication could fail to converge. For the non-converged cases, researchers can exclude those cases from the summary statistics or re-run them using different starting values or different random samples (Harwell et al., 1996). Including some codes for checking the convergency of each replication might be useful. In our example codes, the convergency of the estimation for each replication was checked using the codes in lines 46-52 in Appendix D. On the other hand, printing some information about the current progress of a simulation might be helpful when the simulation study takes a long time. For example, IRT models can be formulated using the hierarchical generalized linear model, which can be estimated using the lme() function in the R package lme4. Sometimes, it may be necessary to use both stand-alone software and the lme() function. Usually, based on our experience, it took at least a couple of weeks to run those kinds of simulation because of the heavy computation in the lme() function. In that case, simply printing current values of the iteration could be helpful for checking the current status of a simulation. In our example, putting print(rep) within a for loop is enough to check the current status of a simulation.

Many useful R packages for IRT have been developed. For example, the R package irtoys (Partchev, 2009) provides the function est(), which can easily run BILOG and extract item parameters from an output file. If those packages can provide enough information for a simulation study, it would be more efficient to use built-in functions in those packages. However, sometimes, other information that is not summarized by the existing built-in functions may be necessary for a simulation study. More importantly, to our knowledge, there are not many R packages or other software that can help the automation of a simulation study using stand-alone software in IRT. In this paper, we have presented a framework of automating a simulation using stand-alone software and introduced some useful R functions that can be used in automating the simulation procedures. Also, an example of automating a simple simulation study using MULTILOG is provided in the appendices

## ORCID

Sunbok Lee http://orcid.org/0000-0002-0924-7056

## 6. REFERENCES

Bandalos, D. L. (2006). The use of monte carlo studies in structural equation modeling research. In Structural equation modeling: A second course (pp. 385–426).

Greenwich, CT: Information Age.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: The Guilford Press.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*, 225–245.

Friedl, J. (2006). *Mastering regular expressions*. Sebastopol, CA: O'Reilly Media, Inc.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101–125.

Kim, H. J., Brennan, R. L., & Lee, W. C. (2017). Structural Zeros and Their Implications With Log-Linear Bivariate Presmoothing Under the Internal-Anchor Design. *Journal of Educational Measurement*, *54*, 145-164.

Kim, K. Y., & Lee, W. C. (2017). The Impact of Three Factors on the Recovery of Item Parameters for the Three-Parameter Logistic Model. *Applied Measurement in Education*, *30*, 228-242.

Kim, S., & Lee, W. C. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. *Journal of Educational Measurement*, *43*, 53-76.

Nader, I. W., Tran, U. S., & Voracek, M. (2015). Effects of Initial Values and Convergence Criterion in the Two-Parameter Logistic Model When Estimating the Latent Distribution in BILOG-MG 3. *PloS one*, *10*, e0140163.

Partchev, I. (2009). irtoys: Simple interface to the estimation and plotting of irt models. R package version 0.1, 2.

R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*, 207–230.

Spector, P. (2008). *Data manipulation with r*. New York, NY: Springer.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). Multilog 7 for windows: Multiple-category item analysis and test scoring using item response theory [computer software]. lincolnwood, il: Scientific software international. IL: Scientific Software International.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). Bilog-mg: Multiple-group irt analysis and test maintenance for binary items. *Chicago: Scientific Software International*, *4*(85), 10.

# Appendix A

## Setup and Data Generation

```
# load R packages to use some functions in this code
library(gdata) # required for 'write.fwf()' function
library(stringr) # required for 'str_extract()' function
#########################################################################
# setting simulation conditions #
#########################################################################
# data, input files, and output files will be saved
# in the main folder (mF) below.
mF <- "C:/I5S500"
nReplications <- 100 # set the number of replications
nItems <- 5 # set the number of items
nSubjects <- 500 # set the number of subjects
# set the true item parameters for the 2PL model.
# a = discrimination parameters.
# b = difficulty parameters.
a <-c(1,1,1,0.5,0.5)
b <-c(-2,-1,0,1,2)
# specify the folder where MULTILOG is installed.
# 'Progra~1' is a short name for 'Program Files' folder
# in the Microsoft Windows operating system.
# \\ indicates escape sequence for \ in R.
multilogPath <- "C:\\Progra~1\\MULTILOG\\mlg"
#########################################################################
# generating item responses based on the 2PL #
#########################################################################
dir.create(file.path(mF)) # create main folder
for (replications in 1:nReplications) {
  # generating responses based on 2PL
  eta <- matrix(0,nSubjects,nItems)
  for (i in 1:nSubjects) {
     theta <- rnorm(1,0,1)
     for (j in 1:nItems) {
        eta[i,j] <- a[j]*(theta-b[j])
     }
  }
  randomMat <- matrix(runif(nSubjects*nItems,0,1),nSubjects,nItems)
  y <- ifelse(randomMat < (exp(eta)/(1+exp(eta))),1,0)
  colnames(y) <- c(paste("I",1:nItems,sep=""))
  # export responses y to text file with MULTILOG data format.
  # write.fwf() is used to create fixed width format data file for MULTILOG.
  # fN is the file name for data file.
  fN <-
paste("MULTILOGI",nItems,"S",nSubjects,"R",replications,".dat",sep="")
  write.fwf(data.frame(formatC(1:nSubjects,width=5,format="d",flag="0"),y)
            ,sep="",file=file.path(mF,fN),colnames=FALSE)
}
```

# Appendix B

## Preparing Input Files for MULTILOG

Contents of Template Input File, ”C:/template/template.mlg”

```
MULTILOG command file
for the Rasch model
>PROBLEM RANDOM,
INDIVIDUAL,
DATA = ’<datafile>’,
NITEMS = <nItems>,
NGROUPS=1,
NEXAMINEES = <nSubjects>,
NCHARS = 5;
>TEST ALL,
L2;
>END ;
2
01
<key>
N
<format>
```

Generating Input Files

```
#########################################################################
# Generating Input Files for MULTILOG using "template.mlg" #
#########################################################################
setwd(mF) # change working folder to main folder(mF)
dataFiles <- list.files(pattern = ".dat", file.path(mF)) # read data files
for (rep in 1:nReplications) {
  # template.mlg will be read and <datafile>, <nItems>, <nSubjects>
  # ,<key>, <format> in the template file will be replaced
  # with appropriate strings.
  # read template.mlg file
  text <- readLines("C:/template/template.mlg")
  # replace data file names in the template.mlg using gsub() function in R.
  dName <- file.path(mF,dataFiles[rep])
  text <- gsub(pattern = "<datafile>", replacement = dName, x = text)
  # replace number of items
  text <- gsub(pattern = "<nItems>", replacement = nItems, x = text)
  # replace number of subjects
  text <- gsub(pattern = "<nSubjects>", replacement = nSubjects, x = text)
  # replace key
  key <- paste(rep(1, nItems), collapse="")
  text <- gsub(pattern = "<key>", replacement = key, x = text)
  # replace format
  format <- paste("(5A1,", nItems,"A1)", sep="")
  text <- gsub(pattern = "<format>", replacement = format, x = text)
  # input file name is the same as the data file name
  inputFile <- paste(unlist(strsplit(dataFiles[rep], "\\."))[1], ".mlg", sep="")
  # write to text file
  writeLines(text, con = inputFile)
}
```

# Appendix C

## Running MULTILOG in R

```
1   ######################################################################
2   # Run MULTILOG using system() function in R #
3   ######################################################################
4   for (rep in 1:nReplications) {
5     # system(command) invokes the OS command specified by command.
6    # ex) the following command will run MULTILOG with input file 'I5S100R1.mlg':
7     # R> system("C:\\Progra~1\\MULTILOG\\mlg I5S500R9")
8      # strsplit() function is used to remove a file extension from data file
name.
9     system(paste(multilogPath,unlist(strsplit(dataFiles[rep],"\\."))[1]))
10  }
```

## Appendix D

## Extracting and Summarizing Results from Output files

```
1   ######################################################################
2   # Summarizing Results from MULTILOG Output Files #
3   ######################################################################
4   # read *.OUT files (MULTILOG output files) into outPut1 using lapply().
5   outPut1 <- lapply(list.files(pattern=".OUT"),readLines)
6
7   # extracting discrimination parameters from output files.
8   # extract the lines that containing discrimination parameters using grep().
9   # grep("GRADED CATEGORIES",x) returns indices for the lines that contains
10  # "GRADED CATEGORIES". Since discrimination parameters are located
11  # in the next second line after that, grep("GRADED CATEGORIES",x)+2
12  # will specify indices for the lines that contain discrimination parameters.
13  outPut2 <- lapply(outPut1, function(x) x[grep("GRADED CATEGORIES", x)+2])
14  # sort outPut2 according to the items
15  outPut3 <- lapply(1:nItems, function(x) unlist(lapply(outPut2,function(y) y[x])))
16  # extract only the decimal number from outPut3
17  # regular expression: [-+]?[0-9]+\\.[0-9]+ = any single decimal number
18  outPut4 <- sapply(outPut3, function(x) as.numeric(str_extract(x,"[-+]?[0-9]+\\.[0-
9]+")))
19  # put column name to outPut4
20  colnames(outPut4) <- paste("Item",1:nItems,"a",sep="")
21
22  # extracting difficulty parameters from output files.
23  # extract the lines that containing difficulty parameters using grep().
24  # Since difficulty parameters are located in the next third line after that,
25  # grep("GRADED CATEGORIES",x)+3 will specify indices for the lines
26  # that contain difficulty parameters.
27  outPut5 <- lapply(outPut1, function(x) x[grep("GRADED CATEGORIES",x)+3])
28  # sort according to item.
29  outPut6 <- lapply(1:nItems, function(x) unlist(lapply(outPut5,function(y) y[x])))
30  # extract decimal numbers.
31  outPut7 <- sapply(outPut6, function(x) as.numeric(str_extract(x,"[-+]?[0-9]+\\.[0-
9]+")))
32  colnames(outPut7) <- paste("Item",1:nItems,"b",sep="")
33
34  # calculate bias and RMSE across replications
35  mean_a <- apply(outPut4,2,mean)
36  bias_a <- mean_a - a
37  rmse_a <- sqrt(apply((outPut4-t(replicate(nReplications, a)))^2,2,mean))
38  mean_b <- apply(outPut7,2,mean)
39  bias_b <- mean_b - b
40  rmse_b <- sqrt(apply((outPut7-t(replicate(nReplications, b)))^2,2,mean))
41  results_a <- rbind(outPut4, mean_a, a, bias_a, rmse_a)
42  results_b <- rbind(outPut7, mean_b, b, bias_b, rmse_b)
43  finaloutput <- cbind(results_a, results_b)
44   rownames(finaloutput) <- c(1:nReplications, "mean", "true parameters", "bias",
"rmse")
45
46  # check whether the MULTILOG was terminated normally for each replication.
47  outPut8 <- lapply(outPut1, function(x) x[grep("NORMAL PROGRAM TERMINATION",x)])
48  outPut9 <- matrix(0,nReplications,1)
49  outPut9 <- lapply(1:nReplications, function(x) ifelse(length(outPut8[[x]])==0,
50              outPut9[x]<-NA, outPut9[x]<-outPut8[[x]] ) )
51  outPut9 <- do.call("rbind",outPut9)
52  outPut9 <- rbind(outPut9,0,0,0,0)
53  # export final result to text file, "finaloutput.csv"
54  finaloutput <- cbind(finaloutput,outPut9)
55  write.csv(finaloutput,file=file.path(mF, "finaloutput.csv"))
```

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 701–712

DOI: 10.21449/ijate.429123

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Using the Fuzzy Logic in Assessing the Programming Performance of Students

**Nihan Arslan Namlı** [1,*], **Ozan Şenkal** [1]

[1] The Department of Computer Education and Instructional Technologies, Faculty of Education-Cukurova University, Sarıcam-Adana/TURKEY

**ARTICLE HISTORY**

*Received: 31 May 2018*
*Revised: 05 October 2018*
*Accepted: 12 October 2018*

**KEYWORDS**

Absolute evaluation system,
Flexible percentage system,
Fuzzy logic,
Programming

**Abstract:** The overall objective of this study is to understand how the fuzzy logic theory can be used in measuring the programming performance of the undergraduate students, as well as proving the advantages of using fuzzy logic in evaluation of students' performance. 336 students were involved in the sample of this quantitative study. The first group was consisted of 150 students, whereas the second group was consisted of 186 students. Cluster analysis was also conducted in order to ensure the neutrality of sample. The rule-based intelligent fuzzy logic assessment logic (FLAL) system was developed. This system has a flexible database in order to assess the academic programming performances of students. Therefore, an absolute evaluation system was used in order to calculate the second group's performance. On the other hand, FLAL system was applied to the first group to determine their programming performance. A Mamdani-type fuzzy logic algorithm mechanism having two inputs and one output was utilized. An independent sample T test was used in analyzing the data sets. As a result, there was a significant difference between first and second groups' results in favor of the first group. While 29 students comprised of 19.3% of all the students failed in the flexible percentage system, 41 students comprised of 22% of all the students failed in the absolute evaluation system evaluating their grades via fuzzy logic system. By increasing the input parameters of the fuzzy logic rules, the results can be addressed more efficiently.

## 1. INTRODUCTION

The education is one of the most important and difficult matters of society. It has been defined in many different ways. Sönmez (1994) defines the education as the period of changing the behaviors. What education includes is to teach how to learn, assess and evaluate the process. Therefore, the assessment and evaluation are two fundamental components of any education period. Butt (2010) describes both of the assessment and evaluation as the action, which instructors take in order to obtain information about students' skills. The assessment and

CONTACT: Nihan Arslan Namlı ✉ nihanarslannamli@egitim.cu.edu.tr Bilgisayar ve Öğretim Teknolojileri Eğitimi Anabilim Dalı, Eğitim Fakültesi-Çukurova Üniversitesi, Sarıçam - Adana/TÜRKİYE

**ISSN-e: 2148-7456 /© IJATE 2018**

evaluation process ensures the progression of learning process of instructors and students (Kerka and Wonacott, 2000). Examining the assessment and evaluation process is important for several reasons. First of all, with the assessment process, the teachers decide what and how students learn (Rockman, 2002). Furthermore, it offers an analysis for classroom management, course guidance, and student support. Under favor of the assessments, the students and teachers receive feedback about their learning, experiences, practices, and performances. It also maintains the quality in measurement of students' efficiency in terms of subjects and method (Kibby, 1999).

It is important to take the significance and requirement of assessment and evaluation into consideration in educational context since they play significant role in both learning and teaching. Almost every teacher makes use of assessment tools in evaluating the academic performances of themselves and students. But, various paradigms emerge out of their diversified assessment methods. Before the constructivist learning approach, the behaviorist and cognitivist approaches dominated the teaching, learning, assessment, and evaluation methods. The teachers generally prefer using the objective tests in order to assess the students because the courses are generally based on conveying the facts and information. Furthermore, what is expected from the student is to memorize only the knowledge in order to make an evaluation (Ward, Stoker & Ward, 1996). The behaviorist paradigm was replaced by the constructivist approaches, which brought a change in assessment and evaluation processes. Constructivist approach designates varied qualities to knowledge. For instance; being temporary, progressive, and socially and culturally mediated are only some of them (Anderson, 1998).

There are two main approaches used in determining the grades of students. One of these approaches is the Absolute Evaluation System (AES), which uses a fixed standard in which students' performances are determined in accordance with their own peculiar information, competence, and understanding levels (CTL, 2001). In absolute grading system (AGS) used in Turkey, the lecturers assess the grades of students by using the scale that they prepared. The final score of this scale ranges between 0 and 100 points. The points are expressed by using the letters varying between AA and FF. According to AES, the AGS is presented in Table 1.

**Table 1.** Absolute grading system scale used in Turkey

| Grade | Score | Description |
|---|---|---|
| FF | 0-25 | Fail |
| DZ | 0-25 | Fail |
| DD | >35 to 40 | Marginal |
| DC | >40 to 50 | Poor |
| CC | >50 to 60 | Fair |
| CB | >60 to 70 | Good |
| BB | >70 to 80 | Very good |
| BA | >80 to 90 | Excellent |
| AA | >90 | Outstanding |

The other approach is the Relative Evaluation System (RES), in which the success of student is measured in relation with grades of other students studying in the same class (Keskin and Ertan, 2001). Another approach used is the use of computer-adaptive systems. These are a form of computer-based tests that measure the student's competence level. These instruments that are developed by using FL approaches in evaluation of student's performance have been popularly used (Lin, 2010) for last twenty years. Fuzzy logic approaches includes 0 and 1 as extreme cases of truth (or "the state of matters" or "fact") but also includes the various states of truth in between so that, for example, the result of a comparison between two things could be not "tall"

or "short" but ".38 of tallness." In general fuzzy approach was proposed to assess student performance based on several criteria which were created several algorithm. Samples can be seen from the literature. Echauz and Vachtsevanos (1995) used FL systems in converting the traditional scores into letter notes. Biswas (1995) proposed a method for evaluating the students by using the FL systems. Chen (1999) improved the Biswas method using the fuzzy sets. Kwok et al. (2001) developed a fuzzy set approach for collaborative assessment in a university. Hammadi and Milne (2003) used the neuro-fuzzy theory in determining the appropriateness of students for the education in an engineering department. In another study carried out by Ertuğrul (2006), the academic performances of academic staff in universities were evaluated via UN approach. Bai and Chen (2008) used the fuzzy logic in generating the concept maps and integrating them into the learning systems in automatized manner. Baba et al. (2009) developed a fuzzy decision-making system software for general use in their study. FL has recently been extensively used in the educational assessment and evaluation process (Rasmani and Shen, 2005, Bai and Chen, 2008; Lin, 2010). An example fuzzy logic decision making scale is seen in Table 2.

**Table 2.** Fuzzy logic decision making scale

| Grade | Score | Description |
|---|---|---|
| DC | >30 | Poor |
| CC | >30 to 45 | Fair |
| CB | >45 to 60 | Good |
| BB | >60 to 75 | Very good |
| BA | >75 to 90 | Excellent |
| AA | >90 | Outstanding |

Reference: Kakoty, S., Lal, M., & Sarma, S. K. (2012, August)

In Tables 1 and 2, the students' course grades were modified according to the scale that teachers use. If teachers think that AGS is not probably the most suitable tool in measuring the grade of student, then the teacher prefers the fuzzy logic decision making scale created by the teacher. Therefore, it can be concluded that the students' grades can be calculated by using various methods. "Most of the researchers inferred that any suitable study does not exist to measure the perceptive field. 'The fuzzy logic theory' should be called a system, which can eliminate the deficiencies of traditional system" (Semerci, 2004). The deficiencies of traditional system can be listed as: it is not an accurate representation of the performance and the knowledge gained. It is not an exact scoring system:for example the science subject is someone's weak point and with a tremendous effort, he got an A or a C for all his attempts, which would have made a vast disparity in his sense of accomplishment.also The traditional letter grade system considers that every alphabet is an inducement to perform good or better or the best. That is why flexible grading system is needed .The current study aims to evaluate the programming performance of students by using the fuzzy logic system (FLS). It was also tried to answer the following sub-goals:

1. Evaluating the students programming performance in programming course by using FLS, and
2. Determining the differences between the FLS and AES(Absolute Evaluation System) evaluation logics.

## 2. METHOD

This research is a research in the semi-experimental research model. The only difference between this method and the real experimental method is that the sample is not selected by

random assignment. In experimental research models, it is tried to determine how systematic changes in the independent variable affect the dependent variable eller (Karasar, 2009).

### 2.1. Participants

The participants ($n$ = 336; 48% females) were the undergraduate students studying at CEIT (Computer Education & Instructional Technology) department. The ages of participants were not systematically assessed, since all of the participants selected aged between 17 and 27 years as shown in Table 3.

**Table 3.** Demographic characteristics of participants

| Characteristics | *n* | % |
|---|---|---|
| *Gender* | 336 | 100 |
| Male | 174 | 51.78 |
| Female | 162 | 48.21 |
| *Age* | 336 | 100 |
| 17-19 | 214 | 63.69 |
| 20-22 | 107 | 31.84 |
| 23-27 | 15 | 4.46 |
| *Grade Level* | 336 | 100 |
| 1 | 71 | 21.13 |
| 2 | 63 | 18.75 |
| 3 | 169 | 50.29 |
| 4 | 33 | 9.82 |

In the present study, there are two different groups, which are familiar with programming languages. The first one consisted of 150 students whose rankings were calculated with flexible percentage system, whereas the second group consisted of 186 students, whose rankings were calculated using absolute evaluation system.

### 2.2. Data Collection and analysis

The data of first group were collected from the students' grades in homework, projects, midterm exams, final exams, attendance, and their presentations. However, in order to collect the data of second group, only the final and midterm grades were taken into consideration. Two programs used for the statistical analysis were Matlab and SPSS. Independent sample-T test was used in analyzing the data sets.

#### *2.2.1. Fuzzy Logic Algorithm Mechanism*

According to Mendel, the FLS can be identified as the nonlinear mapping of an input data set to a scalar output data (Mendel, 1995). A FLS can be divided into four segments. These are called fuzzifier (fuzzification), rules, inference engine, and defuzzifier (defuzzification). The general architecture of FLS is presented in Figure 3.

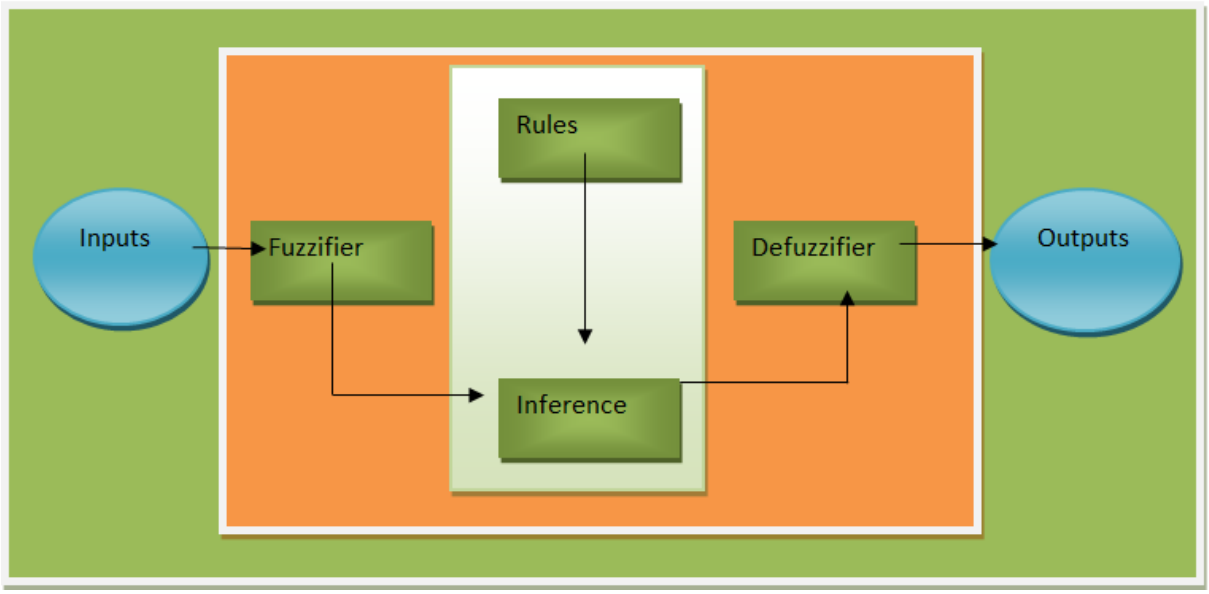



**Figure 3.** Fuzzy logic system (Tamilselvan and Shanmugam, 2014)

The fuzzy logic system incorporates a process. First of all, the input data are gathered. Then, they are converted into a fuzzy set. During this process, the system uses several fuzzy linguistic variables, fuzzy linguistic terms, and membership functions. The whole of this operation is called fuzzification. Then, an inference constitutes some set of rules. Finally, the resulting fuzzy output matches to a crisp output by using the membership functions in the defuzzification step.

### *2.2.2. Fuzzification*

The fuzzification of programming performance was achieved by using both of input variables and membership functions in the fuzzy sets. Every student receives feedbacks about his/her programming grades and attendance into the lesson. These become the input variables of FLS. First input variable is called “mark”, and the second one is called “attendance” in this system. The researchers used a flexible percentage system (FPS) for the first group (homework constitutes 15% of final grade, project 20%, midterm exam 20%, final exam 30%, and presentation 10%), and the percentages were degraded into a point system, so that it was easier to calculate the programming grades of students as an input variable for FLS. For the second group, the researchers calculated the grades in accordance with AES (40% of midterm and 60% of final exam). After calculating the grades of students, then the grades were entered into FLS as “mark”. Moreover, whether the students come to class or not was calculated out of 100 points. If the students’ attendance was below 60%, then the final grade was calculated but not taken into consideration. As for the output variable, the parameter showing the overall programming performance of students was named as a “programming performance”. As it can be seen in Figure 4, two inputs- one output Mamdani system was used to categorize the data set acquired from both of FPS and AES.

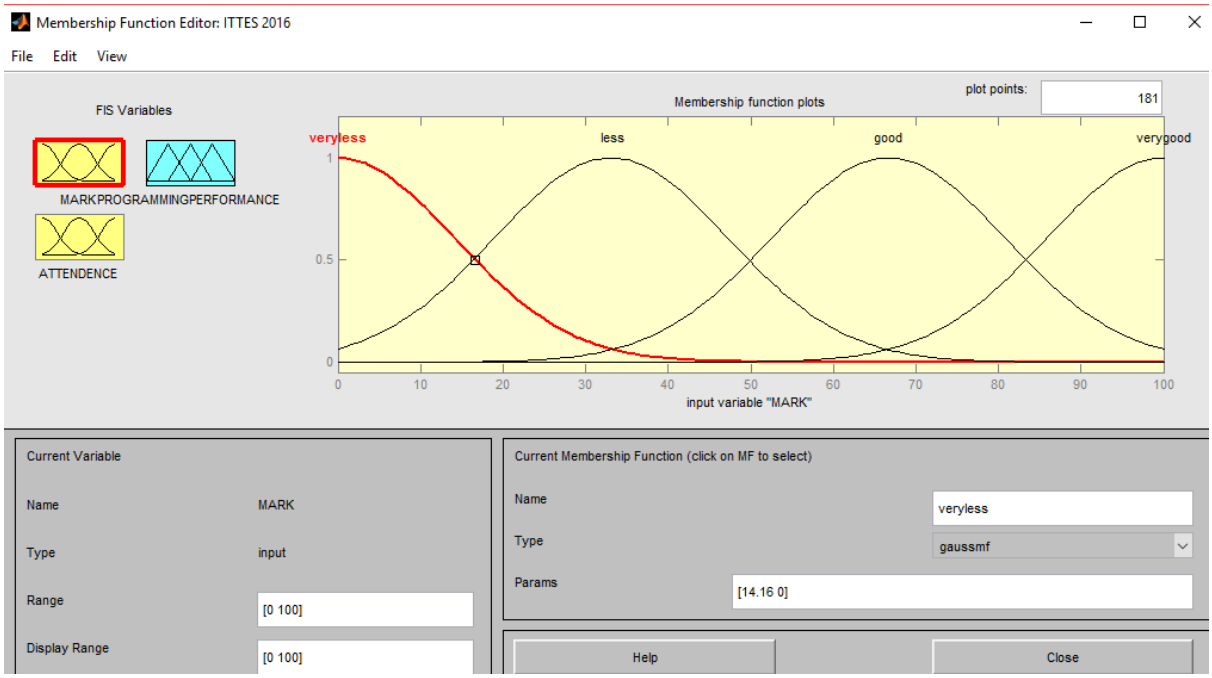



**Figure 4.** Membership function editor involving two inputs and one output

The input and output variables had four gauss membership functions namely very less, less, good, and very good. An input and output variable was placed in a scale, which was scored between 0 and 100. The membership functions of these input and output variables are also presented in Figure 4.

### *2.2.3. Rules and inference*

Some researchers such as Mamdami, Takagi-Surgeno, and Zadeh have developed a series of techniques for the fuzzy decision-making and fuzzy inference, but the Mamdami method was preferred in this research (Semerci, 2004, Zadeh, 1965, Rutkowski, 2004). The rules make use of research's inference process. These rules are non-graphical ones and also are called "If-Then" rules (Altrock, 1995, Semerci, 2004). These rules were constituted from the perspective of obtaining ideas from Computer Education and Instructional Technologies, computer engineering, and biomedical engineering experts in order to determine how the FLS makes a decision according to the importance of the inputs and output in order to determine the membership functions of FLS results (Figure 5).

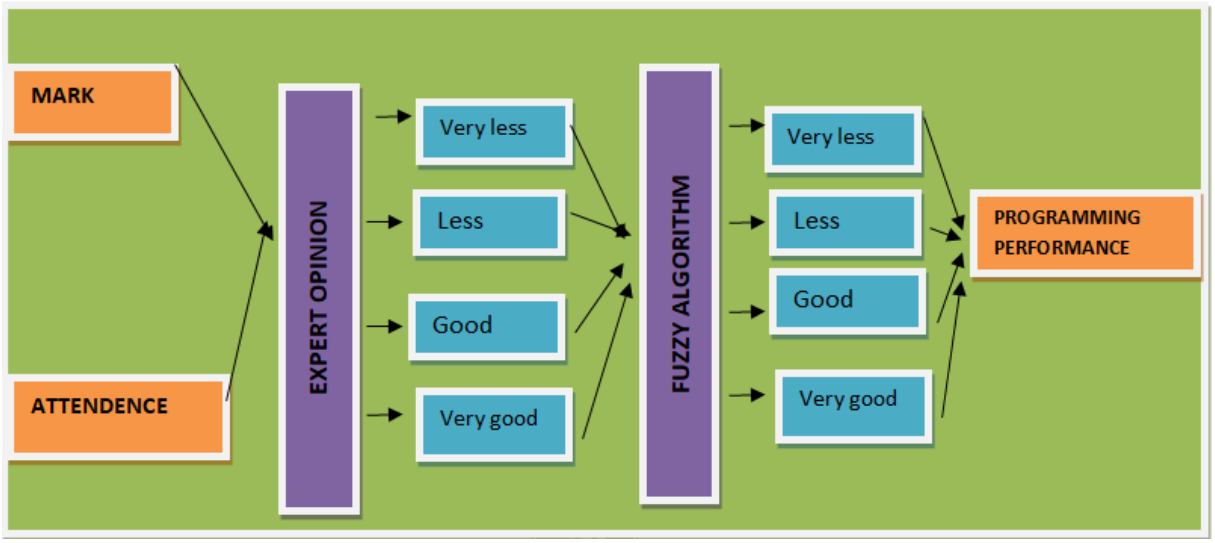



**Figure 5.** Fuzzy logic algorithm with expert opinion

Fuzzy rules for assessing students' performance in Fig. 6 are listed below.

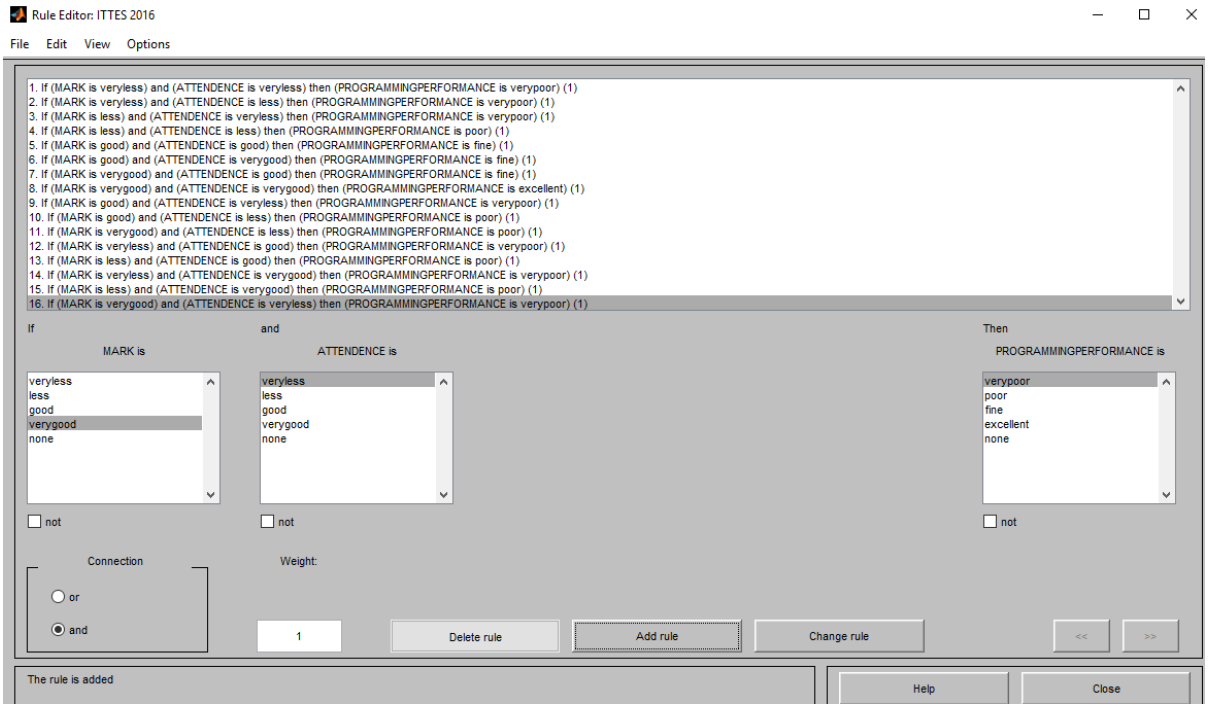



**Figure 6.** Rule editor

In FLS, a rule base was constructed in order to control the output variable. When one rule was active, an AND operation was carried out between the inputs. This method was repeated, so that the output membership functions were determined for each rule. A fuzzy rule is a simple If-Then rule with condition and conclusion. For instance; Rule-2: If (MARK is very less) and (ATTENDENCE is less) then (PROGRAMMING PERFORMANCE is very poor).

### *2.2.4. Defuzzyfication*

Following the rules and inference step, the final result became the fuzzy value. This result should be defuzzified in order to obtain the final output. The final output is the main aim of the defuzzifier component of FLS. The defuzzification was executed in accordance with the

membership function of output variable. The example of a student's rules and performance value is presented in Figure 6.

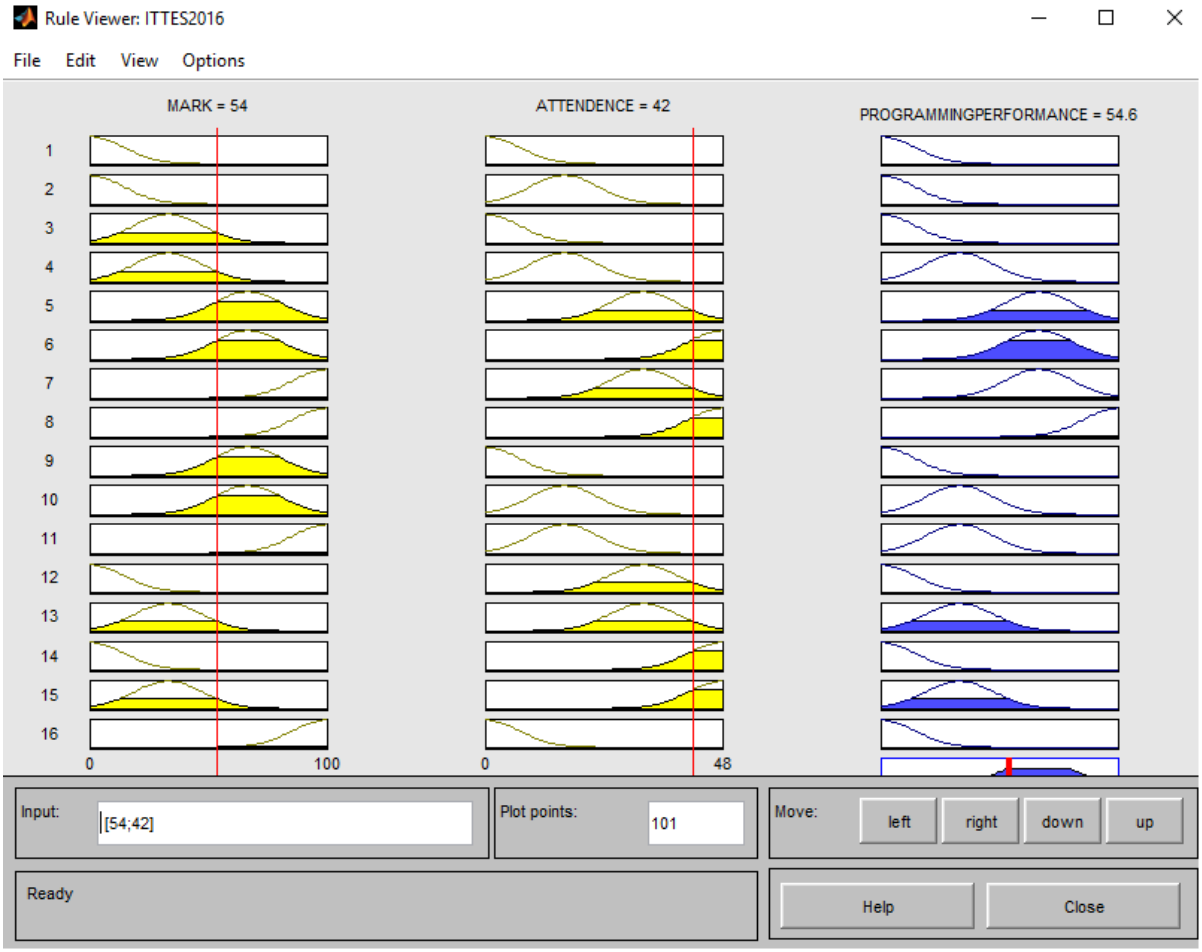



**Figure 7.** Example of the rules and performance value

The result was defuzzificated in accordance with the Mamdani method. The area between input and output axes of membership function is shaded in accordance with the accuracy of rules. One of the student's results in this deffuzifier is shown in Figure 7. The surface view of students is also presented in Figure 8.

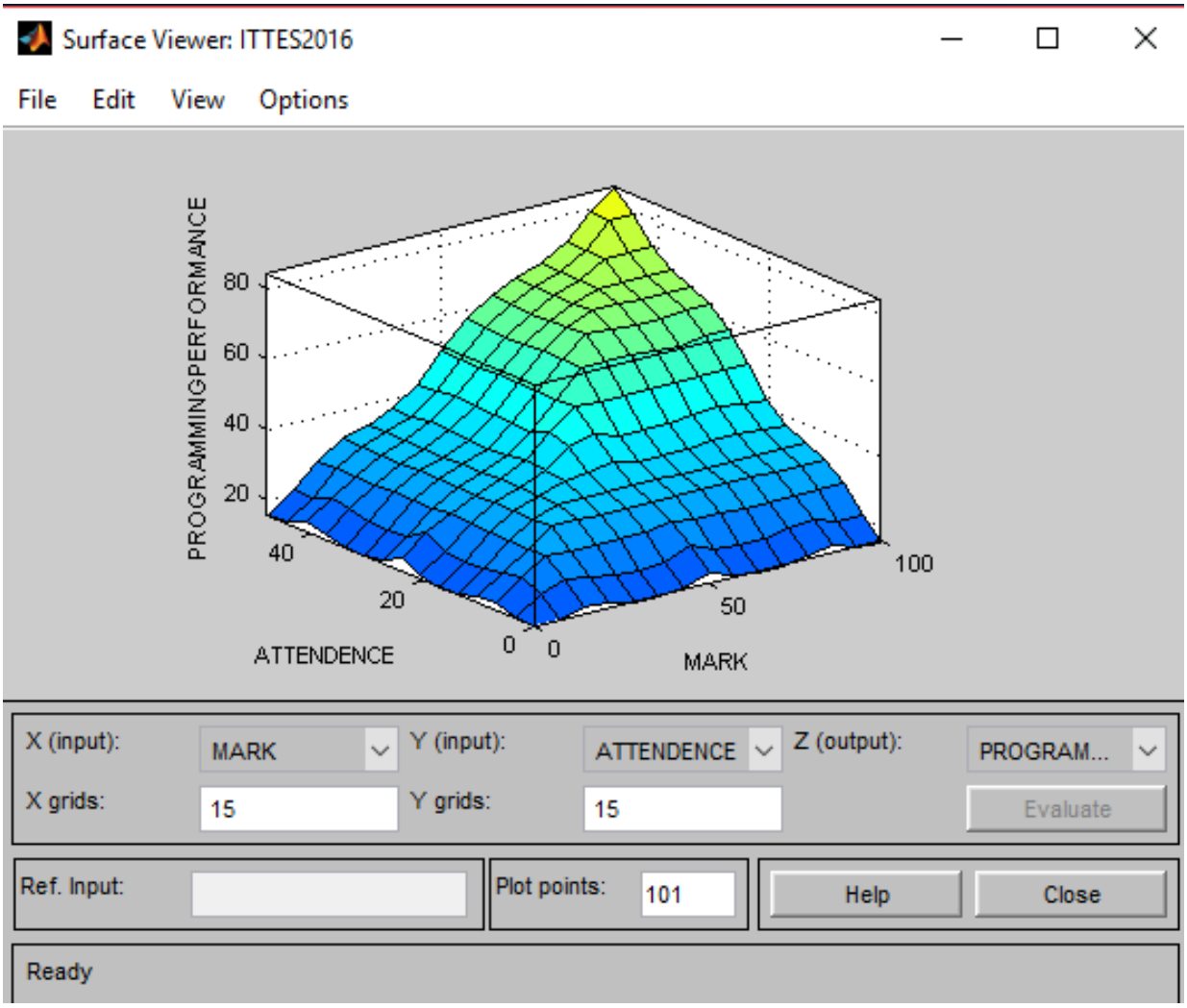



**Figure 8.** Surface view

## 3. RESULTS

The results of first and second groups (Tables 4 and 5) and experts' perspectives were added into the FLS (Figure 5.)

### 3.1. Datasets obtained from two groups for comparing their FLS weights

The independent T-test results of FLS grades using two different groups' calculation are shown in Table 4.

**Table 4.** Independent T-test FLS results of the two different groups

| Study Group | N | Mean | SD | t | p |
|---|---|---|---|---|---|
| 1. Group (FPS) | 150 | 51.76 | 21.90 | 4.62 | .000 |
| 2. Group(AES) | 186 | 39.77 | 17.00 | | |

As it can be seen in Table 4, the mean score of first group was calculated to be X=51.76, and that of second group was found to be X=39.77 (Table 5). A significant difference in favor of the first group was found between them ($X_{FPS}$ = 51.76, $X_{AES}$= 39.77 t=4.62; p<0.05).

Moreover, it can be seen that the second group was successful less than the first group was. The findings related to the comparisons of the AES and FPS performance based on FLS are presented in Table 5.

**Table 5.** Comparisons of students' AES and FPS programming performance grades according to AGS and FLS

| Stu. No | Stu. Marks | Stu. Attendance | 1.group Performance (FPS) | Final Grade of 1. Group (FLS) | Final Grade of 1. Group (AGS) | 2.group performance (AES) | Final Grade of 2. group (FLS) | Final Grade of 2. Group (AGS) |
|---|---|---|---|---|---|---|---|---|
| 1 | 91 | 46 | 65.70 | DZ | DZ | 73 | DZ | DZ |
| 2 | 89 | 78 | 77.80 | BA | BB | 84.6 | BA | BA |
| 3 | 54 | 44 | 44.19 | DZ | DZ | 50 | DZ | DZ |
| 4 | 91 | 50 | 50.11 | DZ | DZ | 74.6 | DZ | DZ |
| 5 | 55 | 86 | 54.80 | CB | CC | 67.4 | BB | CB |
| 6 | 95 | 93 | 79.27 | BA | BB | 94.2 | AA | AA |
| 7 | 41 | 94 | 70.60 | BB | CB | 68.2 | BB | CB |

In Table 5, the performance grades of first and second groups are presented. Here, S The same procedure was applied for the second group. As a result, 29 students (19.3% of the total) have failed in group 1, whereas 41 students (22% of total) have failed in the group 2. After the results were scored, it became easier to determine the difference between the group 1 and 2.

## 4. DISCUSSION and CONCLUSION

There are two complementary components in an educational process, namely the assessment and evaluation. Assessment is the systematic process of documenting and using empirical data on the knowledge, skills, attitudes and beliefs. By taking the assessment, teachers try to improve student learning. On the other hand, evaluation focuses on grades and may reflect classroom components other than course content and mastery level. Evaluation is a final review on your instruction to gauge the quality.The assessment and evaluation are performed at the end of every semester. It is predicted to find many areas of use in education systems, as well as the evaluation of student academic performance, including the curricula, teachers, educators, instructors, and lecturers (Bai & Chen, 2006; Pavani, Gangadhar & Gulhare, 2012; Rasmani & Shen, 2006). In the evaluation of students' academic performance, the fuzzy logic supplies new techniques that have been used for evaluation depending on the numeric data acquired in assessment and evaluation of exam marks (Rasmani & Shen, 2006).

So, this research presents the integration of fuzzy logic into assessment of students' programming performance grades. In this context, the new FLS has proposed an evaluation system based on fuzzy logic techniques for the students' programming performance. When the results are evaluated by using a fuzzy-expert system, the differences are observed between the outcomes of AES and proposed fuzzy logic-based expert systems. The EAGS (existed absolute grading system) adheres to the stable mathematical rules. On the other hand, the evaluation via FLS offers excellent flexibility and reliability. In this research, two groups of FLS results were shown and a comparison was made between the AES and FPS grades determined in accordance with AGS and FLS. As a result, the grades of some students decreased, but those of some others increased and there was no visible change in some of them (Özdemir & Tekin, 2016). Moreover, there was a significant difference in favor of the FPS. When compared to AES, it was seen that the proposed method, which was the fuzzy logic system, was more appropriate for evaluating the academic performance of students. This result is in parallel with those reported by Gawronski (1971), Çekiç (1991), Bowers (1987), and Kılıç (2002).

It can be concluded that this system can be useful in analyzing the performance of students by making use of membership functions. The improvement methodology can be applied to those students listed in the performance categories of "poor" or "very poor". It should be noted that

this improvement methodology varies between the universities. In the future, C-Means clustering algorithm can be suggested for students' academic performance evaluation because C-means clustering algorithms are qualified to generate and separate the membership function. This makes it easier to find experts for getting their opinion while preparing membership functions. Moreover, the dynamic fuzzy expert system can be developed in order to achieve quick feedback information about the overall performance of the student.

## ORCID

Nihan ARSLAN NAMLI http://orcid.org/0000-0002-5425-1468

## 5. REFERENCES

Altrock, V., C. (1995). Fuzzy Logic Applications in Europe, In J. Yen, R. Langari, and L. A. Zadeh (Eds.) Industrial Applications of Fuzzy Logic and Intelligent Systems, Chicago: IEEE Press.

Anderson, R. S. (1998). Why Talk About Different Ways to Grade? The Shift from Traditional Assessment to Alternative Assessment, New Directions for Teaching and Learning, (74), 5-16.

Baba, A. F., Kuşcu, D., & Han, K. (2009). Developing a software for fuzzy group decision support system: A case study. *The Turkish Online Journal of Educational Technology*, TOJET, 3(8), 22-29.

Bai, S. M., & Chen, S. M. (2008). Automatically constructing grade membership functions of fuzzy rules for students' evaluation. Expert Systems with Applications, 35(3), 1408–1414.

Biswas, R. (1995). An application of fuzzy sets in students' evaluation. Fuzzy Sets and Systems, 74, 187-194.

Bowers, P.S. (1987). *The Effects of the 4MAT System on Achievement and Attitudes in Science.* Unpublished PhD thesis, The University of Noith Caroîina at Chapel Hill.

Butt, G. (2010). Making Assessment Matter, NewYork, USA: Continuum International Publishing Group.

Chen, S. M. (1999). Evaluating weapon systems using fuzzy arithmetic operations. Fuzzy Sets and Systems, (77), 265-276.

CTL, (2001). Teaching at Carolina. Center for Teaching and Learning, University of North Carolina at Chapel Hill. Retrieved 2 February, 2018 from http://ctl.unc.edu/he2.html

Çekiç, H. (1991). *Matching learning and teaching styles in a Turkish EFL university classroom and its effect on language development.* Unpublished master thesis, Bilkent Üniversitesi, Ankara

Echauz, J. R., & Vachtsevanos, G. J. (1995). Fuzzy Grading System. IEEE Transactions on Education, 38(2), 158–165.

Ertuğrul, İ. (2006). Akademik performans değerlendirmede bulanık mantık yaklaşımı. *Atatürk Üniversitesi İİBF Dergisi*, *20*(1), 155-156.

Gawronski, J. D. (1971). *An investigation of the effect of selected learning styles on achievement in eighth grade mathematics.* Unpublished PhD thesis, University of Minnesota.

J. Mendel. Fuzzy logic systems for engineering: a tutorial. Proceedings of the IEEE, *83*(3), 345-377.

K. A. Rasmani & Q. Shen (2006). Data-Driven Fuzzy Rule Generation and its Application for Student Academic Performance Evaluation. *International Journal of Applied Intelligence*, *25*(3), 305-319.

Karasar, N. (2009). Bilimsel araştırma yöntemi. Ankara: Nobel Yayın Dağıtım.

Kerka, S., & Wonacott, M.E. (2000). Assessing learners online: Practitioner file. Washington, D.C.: Office of Educational Research and Improvement.

Keskin, M. & Ertan, H. (2001). İstanbul Üniversitesi'nin Bağıl Değerlendirme Sistemi, İstanbul.

Kılıç, E. (2002), *Web temelli öğrenmede baskın öğrenme stilinin öğrenme etkinlikleri tercihi ve akademik başarıya etkisi*. Yayınlanmamış doktora tezi, Ankara Üniversitesi, Ankara.

Kibby, M (1999). Assessing Students On-line. The University of Newcastle Retrieved June 3, 2008, from http://www.newcastle.edu.au/discipline/sociolanthrop/staff/kibbymarj/online/assess.html

Kwok, R. C.W., Ma, J., Vogel, D., & Zhou, D. (2001). Collaborative assessment in Education: An application of a Fuzzy GSS. Information Management, *39*, 243- 253.

Lin, H. F. (2010). An application of fuzzy ahp for evaluating course website quality. *Computers & Education*, *54*(4), 877-888.

Özdemir, O., & Tekin, A. (2016). Evaluation of the presentation skills of the pre-service teachers via fuzzy logic. *Computers in Human Behavior*, *61*, 288-299.

Rasmani, K. A., & Shen, Q. (2005). Subsethood-Based fuzzy rule models and their application to student performance classification. IEEE International Conference on Fuzzy Systems.

Rockman, I. F. (2002). The importance of assessment, Reference Services Review, *30*(3), 181-182.

R.S. Yadav, S. Kumar (2009). *The Effective Utilizations of Fuzzy Logic Approach for Student Academic Performance Evaluation*. M.Phil. Dissertation, YCMOU, Nashik, India, 6-7.

Rutkowski, L. (2004). Flexible Neuro-Fuzzy Systems: Structures, Learning and Performance Evaluation, Bostan: Kluwer Academic Publisher.

Semerci, Ç. (2004). "The Influence of Fuzzy Logic Theory on Students' Achievement." *TOJET: The Turkish Online Journal of Educational Technology*, 3, 2.

Bai, S.M. & Chen, S.M. (2006). A new method for students' learning achievement evaluation using fuzzy membership functions. Proceeding of the 11th Conference of Artificial Intelligence and Applications, Kaohsiung, Taiwan, Republic of China, 177-184.

Sönmez, V. (1994). Program geliştirmede öğretmen el kitabı, Ankara: PEGEM Yayınları. No:12, 7. Basım.

Pavani, S. Gangadhar, P.V.S.S., & K. K. (2012). Gulhare. Evaluation of Teacher's Performance Evaluation Using Fuzzy Logic Techniques. *International Journal of Computer Trends and Technology*, 3(2), 200-205.

Tamilselvan, G. M., & Shanmugam, A. (2014). "Fuzzy-Logic Based Medium Access Control Model for Battery Lifetime Enhancement in Wireless Body Area Networks." *Journal of Engineering and Technology* 4(2), 135.

Ward, A., Stoker, H. W., & Murray-Ward, M. (1996). Educational measurement: Origins, theories, and explications Volume 2. Maryland: University Press of America

Wu, M. H. (2003). *Research on applying fuzzy set theory and ıtem response theory to evaluate learning performance*. Master Thesis, Department of Information Management, Chaoyang University of Technology, Wufeng, Taichung County, Republic of China.

Zadeh (1965). "Fuzzy sets. Information and Control", (8), 338-354.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 713–730

DOI: 10.21449/ijate.478278

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Using a Transactional Model and Thematic Analysis to Evaluate a Minority Male Student Success Initiative to Improve Participants' Campus Experience and Retention

**Kathryn Todd Bliss [1], Richard Mensah [1], Kelly D. Bradley [1,*], Alexis Rodgers [1], Falynn Thompson [1]**

[1] University of Kentucky, College of Education, Department of Educational Policy Studies & Evaluation

**ARTICLE HISTORY**

*Received: 18 May 2018*
*Revised: 29 September 2018*
*Accepted: 19 October 2018*

**KEYWORDS**

Evaluation,
Minority students,
Higher education,
Transactional model

**Abstract:** While the national discourse about the call for increased effort to help strengthen retention and graduation rates among minority college students seems to be losing its vigor, concerned citizens continue to find other means to sustain the momentum. Similarly, the office for institutional diversity at a public southeastern university established a minority male student success initiative in 2010 to help improve the educational experience of its members. The evaluators' objectives for this transactional, formative, and qualitative evaluation were to understand this initiative from the different vantage points of its stakeholders, and to determine if they were satisfied thus far with its trajectory. The evaluators found that while some stakeholders held similar perspectives, others differed in their views on the initiative's goal(s). The majority, however, expressed satisfaction with the initiative's direction at the time of the evaluation.

## 1. INTRODUCTION

The difficulties that African American and Latino male populations face at the higher education level in the United States are ongoing national issues, and deserve special attention and support. Low retention and graduation rates are some of the leading issues. The present research was to examine the efforts of a public predominantly white southeastern university (PSEU) in its attempt to address the challenges its minority male students face in their campus life. The results shed light on methods to help address these issues on similar campuses.

In 2010, the administrative leadership of a PSEU authorized its head administrator of the Office for Institutional Diversity (OID) to commission a faculty member to lead efforts to establish a research-based initiative to improve the on-campus experience of its African American male students. Data from the PSEU's Office for Institutional Research had indicated that African American men performed lowest among all racial and gender groups. The initial general goal

CONTACT: Kelly D. Bradley ✉ kelly.bradley@uky.edu University of Kentucky, 131 Taylor Education Building, Lexington, KY 40506-001

**ISSN-e: 2148-7456 /© IJATE 2018**

was to focus efforts on this particular group of students to provide a place for them to discuss their campus experiences.

The leadership of this newly founded initiative established a research goal to uncover the concerns of the students and, to some extent, help the students expel any harmful assumptions they might have about campus experiences. A second research goal was to then publish any research findings for individuals on the PSEU's campus who were interested in helping minority students, as well as for other higher education institutions whose minority male students struggled with similar issues. The research results would, thus, not only benefit this PSEU, but they could also assist other institutions with similar initiatives. The PSEU recognized the value of African American men leading such an effort, given their parallel personal experiences to those of the students served by this program. The leadership named the undertaking the Black Male Initiative and an African American assistant professor was commissioned to direct it. In 2012, the leadership decided to add the Latino population to this effort because these students often experience similar difficulties as African American students. As a result, the leadership officially renamed this effort the Minority Male Student Success Initiative—henceforth referred to as the Initiative.

Most of the Black male initiatives (BMI) on higher educational campuses have very similar goals—increasing matriculation, retention, and graduation rates among male students of color. They have some form of mentoring programs that seek to shepherd students from matriculation to graduation. However, while some are specifically meant for Black male students (Office of Diversity and Inclusion-UCO, 2014), others include all underrepresented student groups (CUNY Black Male Initiative, CUNY BMI, 2005). Furthermore, some BMIs are established statewide (University System of Georgia's African-American Male Initiative, AAMI, 2002), while others are institution-specific, like the CUNY BMI. The Initiative is institution-specific but employs a modified version of the AAMI and CUNY BMI models, albeit AAMI is statewide. The AAMI and CUNY BMI have pre-college community outreach and recruitment components in their programs, but the Initiative does not.

The leadership established a smaller group (21 students) from the Initiative that would meet monthly to engage in focus groups. This group also participated in an on-campus faculty-student mentoring program. The name of this 21-member subgroup is the Male Scholars Program (henceforth called the Program—the Program director also serves as the Initiative faculty director). The leadership of the Initiative hoped to use the Program's activities to track and measure the students' experiences for the benefit of the students and for research purposes. The Program director also intended for these 21 students to eventually engage in a community service project, which would include three mini-projects: an after-school program, a fifth-grade lunch program, and a high school mentoring program. The leadership was unable to launch this community service aspect due to insufficient resources, but the underlying purpose was for these young men to put into practice the mentoring they received on campus with school children in their respective community of residence.

In order to determine if the expectations of the various stakeholders were congruent with the goals of the Initiative, the overarching question the evaluators sought to address in this formative evaluation was whether some stakeholders viewed the Initiative the same or differently from others (House, 1978, p. 12). The initial perception of the evaluators was that the Initiative and the Program were separate but associated efforts, thus they outlined two objectives they intended to achieve in this evaluation:

1. To understand the Initiative and the Program from the different vantage points of their stakeholders.
2. To determine if the student participants, administrators, and volunteer mentors were satisfied thus far with the Initiative and the Program.

To achieve the first objective, the evaluators needed to understand how the stakeholders perceived the goals of the Initiative and the Program. Through this understanding and the articulation of different vantage points, the evaluators believed it may help clarify how to describe these efforts as entities for their own benefit and in relation to similar efforts. The second objective would help the evaluators to take into account and to allow for the reporting of whether or not the student participants' initial expectations were met, administrators felt that the efforts were moving in the intended direction, and if volunteer mentors felt that they were part of a worthwhile effort at the university.

The evaluators employed a transactional evaluation model (House, 1978; Patton, 2002; Rippey, 1973; Stake, 1973) in order to understand the various viewpoints of stakeholders invested in the Initiative and the Program. The evaluators believed that this approach would offer a thorough account of a program that was in a developmental stage. They also felt that it would help them to understand and present the stakeholders' perspectives on the goal(s) of the Initiative and its programmatic activities. And finally, using the transactional evaluation model would help the evaluators determine whether Program participants were satisfied or dissatisfied with the given endeavors. This is because using the transactional model allows the evaluators to give voice to the individuals (House, 1978; Patton, 2002) affected by these ongoing national issues on the PSEU campus examined in the present research.

In line with the two outlined objectives, the researchers of this formative evaluation aimed: to understand the different perspectives of the student participants, administrators, and volunteer mentors of the Initiative and its associated Program; and to determine if these stakeholders were satisfied with the efforts' progress and direction at the point of the evaluation. The researchers hoped that by answering such questions and meeting these objectives, they could help demonstrate to the stakeholders if the efforts were effectively supporting the African American and Latino student participants in their higher education endeavors, or at least making progress toward fulfillment of this purpose. For the PSEU administrators of this Initiative, progress toward this purpose could be a contribution toward addressing this national issue.

## 1.1. Literature Review

### *1.1.1. On-Campus Diversity*

The population of the United States may be described as racially and ethnically diverse, but higher education institutions do not reflect this diverse composition. For example, statistics drawn from the PSEU for fall 2012 for the African American Latino undergraduate student population mirrored that of the state in which the PSEU is situated, yet those numbers did not reflect the groups' representations on the campus (PSEU Public Relations Representative, 2012). In addition to promoting inclusion, racial diversity on higher education campuses has been shown to serve educational purposes and benefits (Gurin, Dey, Hurtado, & Gurin, 2002). Not only does diversity provide a feeling of community and fellowship for minority students on predominantly white institution (PWI) campuses, it also enhances the overall campus experience for all races and ethnicities (Spitzberg & Thorndike, 1992). The feeling of community and fellowship encouraged the PSEU to create the Initiative, with the hope of enriching the campus experience of the student participants.

### *1.1.2. Retention*

Colleges and universities might be providing greater access to higher education for minority students, but they still grapple with student retention (Seidman, 2005). Among the myriad explanations cited for this difficulty is students' unpreparedness prior to admission (Redden, 2009). Redden (2009) also adds that the lack of a supportive atmosphere, especially for minorities, undermines enrollment efforts. Kim and Conrad (2006) have found that one of the best approaches for improving retention rates of African American men is encouraging these

students to become involved in "academically engaging interactions" such as assisting faculty members with research projects (p. 420). Lascher (2008) suggests that, "there is strong reason to think that the major factors [of retention] operate similarly across ethnic groups" (p. 12). The PSEU believed that the Initiative could create the needed supportive atmosphere by, for example, linking minority students with mentors with similar cultural backgrounds. It perceived faculty-student interactions as very important in helping the students to stay in school to finish.

### *1.1.3. Experiences of Minority Students at Predominantly White Institutions*

The negative school experiences of African American and Latino male students at many educational levels, such as high school (Ford, 2012; Roderick, 2003) and college (McGee & Martin, 2011; Strayhorn, 2010), have been empirically investigated. Lewis, Oliver and Burris (2011) suggest that the alarming rate of attrition and low enrollment among minorities could be due to reasons such as…

> … disproportionately high school dropout rates for black [sic] male students; stressful economic circumstances for black families; the influence of popular culture, particularly the influence of hip-hop music and culture on black males; the high incidence of black males entering the military; the availability of employment after high schools [sic]; disproportionately higher levels of incarceration among black males; high rates of homicide among black males; and fear, distrust, and unfamiliarity with college environments. (p. 4)

Lewis et al. (2011) also outline student discussions on feelings of a lack of belonging, worthlessness, perceived racism, and lack of support from faculty, administrators, and students. One student explained his experience with racism…

> Everyone has their story of randomly being called a nigger on campus. And I was talking to some guys the other day at the [Drexel Center] and they were like, "Well, it wouldn't be a normal semester for me if I didn't get called 'nigger' at least once." I'm like, "Really?" And, I mean, it recently happened to me and I'm just at the point where…. It happened last year—whatever—I couldn't worry about it. I guess in the initial state of it, you're like, "OK, well, I wonder what…" because it was so random. It's like I wonder what they're thinking about me; I wonder how you really feel. I mean, after a while, it starts to go away; but at this point it's still rampant in my… I mean it's hard to get past that. You know you have to go to class, and you go to class, you do your best and everything, but still in your head, it's like, it's, it's hard to get past that. (Lewis et al., 2011, p. 14)

Other research studies capture the experiences of individuals from differing racial and ethnic backgrounds who are students at a PWI. Jones, Castellanos and Cole (2002) interviewed students from several minority groups (African American, Latino, American Indian, and Asian American) and a number of them reported experiences of overt racism and alienation. The students believed that school administrators were unable to fulfill their promise of taking proactive steps to correct the situation. The PSEU in this research believed such negative experiences are some of the leading forces that impact enrollment, retention, and graduation of minority students (Jackson, 2012; Lewis, Oliver, & Burris, 2011; Strayhorn, 2010), and that neglecting to make efforts to promote belongingness and self-worth among racial minorities is to let the status quo remain. The administrators wanted to use the Initiative to bring some changes to the situation.

### *1.1.4. Mentoring*

To a large extent, recruiting students to occupy classes is not as much of a problem as the ability to retain them until graduation. First generation and low-income college students, especially, are at risk of dropping out. This partly sparked the leadership's desire to have a mentoring program to improve the educational and social experience of students, and to help keep them enrolled until completion. The Initiative's leadership believed that mentoring relationships, especially

those that develop naturally, may be a productive way of addressing the college adjustment issues that at-risk students face (Thile & Matt, 1995).

Santos and Reigadas (2002) attest that "a quality faculty-student mentoring relationship is likely to engender positive self-perceptions in at-risk students as well as feelings of self-efficacy, personal control, respect for oneself, and a sense of being valued and respected by significant others" (p. 42). It is also theorized that when mentors and mentees share common viewpoints by way of similar ethnic or cultural backgrounds, it creates a sense of homogeneity in the relationship and may enhance supportiveness—a quality that is critical in the life of young adults (Tinto, 1987). Thompson (1995) adds that this dimension of the mentor-mentee relationship, in which participants in the mentoring program share common attributes such as occupational goals, religious values, and cultural background, offers more positive perspectives on life and fosters emotional support (p. 44).

The Initiative's leadership felt that mentors could provide advice and guidance to the young men on how to chart their life's course as students and as graduates who would leave to pursue profitable future endeavors. They hoped that the associated Program would serve as a platform for community building and networking among its participants, and would assure the students that they had advocates on campus—people on whom they could rely for better academic and career decisions. The leadership anticipated that the programmatic activities would help the students develop quality leadership skills that would position them socially as equally accepted contributors of national development, with equal rights and responsibilities (Interview with the Program leadership, 2013).

## 2. METHOD

The evaluators first decided to employ the term "African American" to describe the portion of the research population who were of African descent. When reviewing the administrators' interview transcripts, however, the evaluators found that one of the adminstrators primarily employed the term "Black" when referring to the students of African descent who took part in the Initiative. The other administrator used the term "African American" to describe this part of the research population. Although individuals who identify themselves either as African American or as Black may participate in this Initiative, the evaluators chose to employ the term "African American" as it is more representative of the Program participants and, specifically, of the respondents who were interviewed. For this article, the evaluators have also de-identified all respondents and present just the views of the respondents about the Initiative and the Program.

### 2.1. Sampling

Given the young age of the Initiative and its associated Program, three years and four months respectively, and that both were still in a developmental stage, the researchers conducted a formative evaluation (Division of Research, Evaluation and Communication, and National Science Foundation, 2002, p. 8). The researchers used interview protocols with relevant stakeholders to understand more about the Initiative and the Program. The interview data was then treated using qualitative data analysis such as coding themes (e.g., Coffey & Atkinson, 1996) both within each and across all stakeholder interviews. The decision to use interview protocols for data collection and a qualitative approach to data analysis stemmed from the researchers' employment of the transactional model (House, 1978; Patton, 2002; Rippey, 1973; Stake, 1973), discussed further below. This model is based upon gathering the different perspectives of the various stakeholders involved in the evaluated program (in this case the Initiative and its associated Program). In light of the nature of this model, the researchers chose for the evaluation design to use maximal variation sampling, which is a purposeful sampling strategy (Creswell & Clark, 2011, pp. 173-174; Teddlie & Yu, 2007).

According to Creswell and Clark (2011), purposeful sampling is when "researchers intentionally select…participants who have experienced the central phenomenon…explored in the study" (p. 173). In maximal variation sampling, "diverse individuals are chosen who are expected to hold different perspectives on the central phenomenon" (p. 174). The evaluators chose the purposive sampling approach, for which fewer than 30 respondents are typically selected for the "narrative data" gathering process (Teddlie & Yu, 2007, p. 84). The evaluators interviewed the only two administrators of the Initiative and Program, and had originally planned to interview the 42 Program participants (21 student mentees and 21 volunteer mentors). Time, however, did not permit this as the evaluators had hoped. Hence, due to the time limitations, they were only able to interview four students and two mentors.

### 2.2. Instrumentation and Data Collection

The evaluators developed and used three different interview protocols for the various stakeholders, stemming from the employment of the transactional evaluation model (House, 1978). They used notes taken from separate meetings with an OID representative and with the Initiative's leadership to understand more about the Initiative's establishment, as well as its past and present programmatic activities. The evaluators gathered additional background information by observing the interactions between the administrators and the students. To allow access to information on the Program's activities, the Initiative's leadership added the evaluators to the Program mailing list. The evaluators attended the Program's monthly meeting as "observers as participants" (Glesne, 2006, p. 50) in order to have first-hand experience with the student activities. Glesne (2006) supports this notion explaining that, "the researcher remains primarily an observer but has some interaction with study participants" (p. 50). This interaction helped evaluators to build rapport with Initiative members.

Attending the initial meetings with administrators and two monthly Program group meetings with the students allowed the evaluators to obtain data about issues of importance to the students, the topics discussed at focus groups, the students' personal and professional goals, and their future plans (e.g., summer internships). The evaluators used this data, along with related literature, to develop the interview protocols. The evaluators interviewed respondents to gain insight into their differing perspectives on the Initiative and its associated Program. The interview data helped the researchers to meet their two evaluation objectives, stated above, as participants freely expressed their perceptions about the Initiative and the Program's activities. The researchers were allowed to audio-record each interview, and to take notes to supplement the recordings, which were then transcribed for thematic analysis.

Not only did the evaluators seek to understand the Initiative's goal(s) and stakeholders' satisfaction or dissatisfaction with it, they also wanted to understand these aspects with regard to the Program, given its intimate connection with the Initiative. The evaluators discovered some connection after establishing their evaluation objectives in the process of conducting their early research into the background of the Initiative. They, however, maintained these objectives and decided to seek additional information during the interviews to clarify the connection between the two efforts. Thus, they present findings related to both the Initiative and the Program in this report.

### 2.3. Analysis and Interpretation

Data analysis began with transcription of the interviews. The evaluators then employed thematic analysis, which first involves coding the data and then grouping those codes (Coffey & Atkinson, 1996). Mishler (2003) explains that the theoretical framework and methodology underlying a study affects how researchers choose to present the text in their transcriptions and in the conclusions drawn from their findings. Indeed, when the evaluators transcribed and coded the text of their respective interviews, they made certain decisions based upon the specific issues

and theoretical foundations at work in the evaluation. As a result, the evaluators recognize that their particular perspectives from researching this topic influenced the codes they established as a group and individually, how they coded the text, and the quotes they chose to present in their findings.

To begin the data analysis process, the evaluators established initial codes, or categories, for use in the analysis of each interview based on the two evaluation objectives, the overall research question, and the literature review (Coffey & Atkinson, 1996). Since they needed to understand the stakeholders' perspectives on the goal(s) of and satisfaction or dissatisfaction with the Initiative and its associated Program, the evaluators created codes that reflected these objectives. The initial code of "goal" accounted for text sections that included words such as goals, purpose, and intentions. The initial codes of "satisfaction" and "dissatisfaction" accounted for text sections that included words such as expectations, direction, value, and future involvement.

The evaluators coded for their respective interviews using the initial codes, while also allowing for additional codes to emerge from the interview data (i.e., in vivo codes) (Strauss, 1987, as cited in Coffey & Atkinson, 1996, p. 32). To begin identifying themes across the interviews, each evaluator presented the codes used in his or her given interview, both initial and in vivo, to determine if any of the other interviews had similar categories. The evaluators, therefore, maintained the integrity of the individual stakeholders' viewpoints by allowing unique codes to arise from each interview's content. They then compared each interview's distinctive codes to see if any of them conceptually aligned. If they did, the evaluators combined them to produce a new category that was then added to the list of existing categories.

Each of the interviews was then color-coded and segments of text from each interview were placed under the relevant coding category or categories. The evaluators finally grouped the categories and their matching textual data techniques such as developing codes to identify themes (Coffey & Atkinson, 1996). From these themes, the evaluators fulfilled the two evaluation objectives by identifying how each of the interviewed stakeholders understood the goal(s) of the Initiative and the Program (Objective #1), and whether or not they were satisfied with the two endeavors thus far (Objective #2). Thematic analysis of the in vivo codes from each interview helped the evaluators to address the overall evaluation question of whether stakeholders viewed the Initiative (and its associated Program) differently from other stakeholders (House, 1978). At times, the evaluators found sections of text that did not fit within their existing coding categories. Hence, they labeled them as "outliers" and considered them valuable data worthy of additional analysis (Coffey & Atkinson, 1996, p. 47).

## 3. FINDINGS

Quotations from the stakeholders describing their viewpoints on the Initiative and Program goal(s) can be found in Tables 1 and 2, respectively. Stakeholder quotations relaying satisfaction or dissatisfaction with the Initiative and Program can then be found in Tables 3 and 4, respectively. Upon analyzing the interview data, the evaluators found that the stakeholders' views were similar, yet different in the ways in which they were expressed.

Although the evaluators entered the interview with Administrator #1 [Initiative and Program Director] with the idea that the goals of these two efforts would be somewhat related, they left with the knowledge that the purposes were basically one and the same.

**Table 1.** Stakeholders' Perspectives on Initiative's Goal (Evaluation Objective #1)

| Administrator #1 | Administrator #2 | Mentor | Student |
|---|---|---|---|
| "Improve the experience the Black and Latino males have on this campus, educational experience, social experience" | "To turn out educated men who are making a contribution to the betterment of society" | "The goal was to provide these young men forums and situations and with individuals that they can talk to, that they can converse with..." | "To like hook us up with like mentors and leaders" |
| "Improve that experience for them to increase retention rates of Black and Latino males on campus and to increase graduation rates of Black and Latino males" | "An acceptance of Black and Latino males for who they are...Black and Latino males will come to be respected... That there will no longer be the demand that they should explain why they're here..." | "We're trying to give them an opportunity to talk about some of the issues that they face and trying to give them sounding boards in us to where we can say, 'Yeah, we get it. We had the same thing happen to us 5, 10, 15, 20, 30,' however old you are, 'years ago'" | "To address problems on campus specific to PSEU's campus and to try to like see how we can work on them" |
| "To prepare them to leave here to lead successful lives, (i.e., going to graduate school and finding success in professional school, careers, families)" | | | "To just promote diversity on campus" |
| "Help them to find success once they leave and to live up to what their potential is" | | | "To connect us with the resources since we are the minorities on campus, so that we don't feel excluded in anything" |
| | | | "And yeah, like the graduation factor. Having somebody be accountable for you, so that you don't just come here and fail out of college. It's like an intervention tool" |
| | | | "And to just unite Blacks and Latinos because our populations on campus are so small" |

Reflecting on the involvement with the Initiative or Program, Administrator #1 described the experience as being "very similar to my initial expectations… you know, at this point we don't have any support from the university by way of staff or help or finances or anything like that, and so I guess my expectations were pretty in line with what the reality would become." This administrator expected it to be "possible for these young men to be successful," but foresaw a difficult situation because of a lack support from the PSEU. He or she was, however, pleased with the level of commitment from the student participants, volunteer mentors, and other stakeholders involved with the Program, and that the young men were establishing good relationships with their mentors.

Administrator #1 also admonished that,

> If you want a program to be successful you have to make it somebody's job. Even if it's not their full-time job…and that takes money. That takes staff…. [A]nd so we have no investment in it right now." With part-time or full-time staff to help with coordination, this administrator believes the Initiative would "be further along in terms of building out the Program, building out structures, building out curriculum, so that those things we can produce and put on the shelf and be able to use for years to come.

But, the administrator acknowledges that, "the support for this kind of work needs to come from somewhere. Because when you [look] at institutional research and you look at where Black men are, and where Latino men are, I mean the need is clear as day." This administrator was of the view that, "There are funds available for this type of work. There are a lot of grants out there that fund this type of work…the problem is, it has to be somebody's job, and somebody has to have the time to write grants and go out there and go after the money."

**Table 2.** Stakeholders' Perspectives on Program's Goal (Cont. Evaluation Objective #1)

| Administrator #1 | Administrator #2 | Mentor | Student |
|---|---|---|---|
| Insinuates during the interview that the goals of the Program and the Initiative are the same. | "Even though it's a program to help build the success of our underrepresented young men, it is, in fact, a research project, because hopefully what we learn will be useful to somebody else"<br><br>"When we incrementally grow the resources [described in Findings] ... [the Program's goals will be] to provide these young men with the kinds of experiences that I have described. To see them leave PSEU with the confidence and the knowledge about their place in and their interaction with the world. And that they would take that and encourage more men, generation after generation, every young boy that they meet, all the men involved in their lives, that they would just pay it forward" | "I think the goal for the program is just to make the...I know early on, [an administrator] motivated it by talking a little bit about graduation rates and trying to increase the retention of Black males on campus. I think, though, that the way that the Program can be really valuable is just that if we see ourselves as trying to make the experience of being a Black male here at PSEU just a little bit easier for them" | "I'll just say the goal is just to have us graduate from PSEU and just give us like a safe haven for us, you know"<br><br>"Just to promote our scholarship, our awareness of what's going on in the African, Black community"<br><br>"Just help, just promote us to do positive things on campus. The older guys are just looking out for us 'cuz it will help us in the future" |

Administrator #2 stated that the motivation for starting and staying involved with the Initiative stemmed from years of experience working in higher learning institutions in different parts of the country, and seeing the same issues with the success of African American and Latino male students at all of these various institutions. This administrator shared:

> Being in America where – let's face it, I'm not naïve about it, and I hope you're not – where racism is still alive and well, … racial understanding is yet to reach the depths that they should, in order for us, together, to move this country forward. It sounds terribly idealistic, and I am an idealistic person, but I'm also very realistic and practical. I know that there are things you've got to come to grips with on the ground before you can get those dreams in the sky.

Administrator #2's initial involvement the Initiative stemmed from the desire to put something in place at the PSEU that would help African American and Latino students to achieve, even when there are barriers preventing them from doing so. He or she also served as a mentor to one of the young men in the Program at the time of this evaluation. As for future involvement with the Initiative, Administrator #2 planned to continue to "serve as a mentor for as many of the young men who appreciate that." He or she also had several expectations upon developing this Initiative, but anticipated that the endeavor would yield gradual results. Administrator #2 explained:

> They've [Initiative and Program leadership] only just begun. These things take time. Few people appreciate that growing diversity and strengthening inclusion is a process, and people can be so impatient. And . . . this situation has been decades in the making – how do you expect somebody to fix it overnight?

Furthermore, the administrator shared that what is being done within the Initiative is "endless work" and that "there can never be an end date."

**Table 3.** Stakeholders' Satisfaction/Dissatisfaction with Initiative (Evaluation Objective #2)

| Administrator #1 | Administrator #2 | Mentor | Student |
|---|---|---|---|
| Satisfaction: "I think the community building is very important among [the students] themselves and I see evidence of that.... [the] relationships that I have developed with them are important. And so, it's now more common than [in] the beginning of the year for me to get a phone call or email for somebody to stop by and talk about something or an idea, and that's important" | Satisfaction: "It's not about satisfaction. Like I said, there is no end date. You can never be satisfied, because there's always something to be done, something that needs to be done, something that should be done. There's always something that needs to be undone, because we're dealing with humans and we're not perfect"<br><br>Dissatisfaction: "And we're always grappling with our beliefs, and having the courage (or not) to interrogate those beliefs (or not), so it's a dynamic situation. So no, I'm not satisfied. I'll never be satisfied" | Satisfaction: "No problem… I'm pretty satisfied"<br><br>Satisfaction: "I think the idea is great. . . I think that the effort that Dr.____is putting into it and that the other administrators are putting into it is a genuine thing and it's something that is needed"<br><br>Dissatisfaction: "If there was any dissatisfaction on my part, it would be on my own doing. It's not having been able to make it to a couple of the meetings, the full meetings that they've had with the group . . . but that's all, that's on me. That has nothing to do with the Initiative itself" | Satisfaction: "It has allowed me to meet some people and has given me something to be proud that I am a part of on PSEU's campus as far as something that affects the diversity initiative that the school promotes. It makes me feel like I'm being an active participant in that"<br><br>Satisfaction: "And it makes me feel like my school is providing an opportunity to help me, I guess"<br><br>Satisfaction: "Just have somebody that is similar to me that I can identify with so that I don't feel singled out on campus or anything. And it helps me connect with faculty that are like African American who are in higher positions on PSEU's campus and you know feel more comfortable during my stay here so, yeah" |

As for ideas about improving the Program, Administrator #2 shared that everything could always be done better, but what is most crucial is "engaging it at the proper time and under the proper circumstances." He or she stated that for now "it works" in regard to the Program's activities and the leadership's efforts at the time of the evaluation. But in the future, Administrator #2 hoped to recruit the assistance of a psychologist or a psychiatrist to help the young men involved in the Initiative to deal with the issue of self-efficacy. Tutors could also help the students with concept mastery, and engage them in discussions about literature and art as well as the connections between the two. He or she stressed that the young men could also benefit from assistance in planning trips abroad as a way to encourage them to connect with young men worldwide and to learn about the issues they face.

From the volunteer mentor's perspective, the Program could reach out to other students who may not be as motivated to succeed in school. This would address the issue of retention of African American and Latino men. He explained his thoughts as follows:

> I think the mentees had to apply and it was a pretty selective process [into the Program]. Part of me wonders if those are the students that needed help, anyway, you know? I wonder if there would be some kind of way that we could touch those students that didn't sign up because those are probably the ones who were more likely not going to maybe have the skills that they need, the prerequisite skills to come into a big campus like this and perform academically. That's not necessarily the program's fault. I understand from the perspective of the program you're kind of taking those very talented young men [and] giving them extra resources, so that they can go back and lead others.

Aside from these few recommendations, this mentor was both positive and encouraging about his experience with the Initiative and the Program. He praised the work the administrators were

doing to support this program, and also highlighted the intelligence and the potential of the student participants. He said he was pleased to be a part of this effort.

**Table 4.** Stakeholders' Satisfaction/Dissatisfaction with Program (Cont. Evaluation Objective #2)

| Administrator #1 | Administrator #2 | Mentor | Student |
|---|---|---|---|
| Satisfaction: "…the relationships they are building with their mentors, those are important because they're getting the same messages over and over again"<br><br>Satisfaction: "And, they know that they have people that they can go to with any problem. So, I think we're helping"<br><br>Dissatisfaction: "Would I go so far to say [what] we're doing is now increasing retention and graduation rates? No, I can't say that"<br><br>Dissatisfaction: "Absolutely not. To be honest, I am less than satisfied. I'm really ashamed of what we have here. I mean, the only thing we've been able to build with, specifically, the Program is what I pull out of my extra time, you know after dealing with teaching and writing and other responsibilities. That's not the way to build a program or run a program. So, I'm not satisfied. I'm much more ashamed than satisfied with what we've done" | Satisfaction: "Well, again, under the circumstances, I think that we're going in the best direction that we can"<br><br>Dissatisfaction: "Because we don't have endless money to put into this, to bring in some of the kinds of people that I talked about, to have the sessions that I talked about" | Satisfaction: "Yeah, I'm pretty satisfied with it. And so . . . you know, you just kind of get to know them a little bit. Get to know their [students'] personality. I would anticipate, hopefully, in the next couple of meetings and things like that, as things start to progress a little bit, you know, we may get into other issues that's he's having, academically or whatever" | Satisfaction: "But I would say the group discussion is the most important. And just the goal setting, and helping us keep in mind our goals"<br><br>Satisfaction: "Really satisfied. I liked the induction ceremony. That was a really good experience"<br><br>Satisfaction: "Oh, with just the discussions we have and just the accountability factor of having other people that want you to succeed and the opportunity like if you, like it's a program where like from the beginning I felt that we had ownership of it"<br><br>Satisfaction: "So far, yeah, except for the service part [which hadn't started yet] like I said before. . . I think we guys are getting something out of it, you know. It may not be like life changing on that scale yet, but we are making connections with our mentors. So, I think so far it is [has been positive]" |

A student described his overall experience with the Initiative and the Program as positive. He alluded to the fact that he enjoyed the autonomy that the Program gave the students, as he explained…

I had ideas over winter break that I bounced off [Administrator #1] and [Graduate Assistant]. And they were actually like, "Great idea man, we'll do it." Like one of the discussion topics [sic] at our first meeting was the Django movie and I had had an idea to do that. And so, he let me do the background research to kind of help facilitate the discussion. So, I was like "yeah!" It was a way for me to get more involved in it, and not just sit there and have them talk to me for like two hours. So, I just like the ownership that they let us take over what we learn and what we do in the Program.

He did make a couple of recommendations, such as holding meetings every week on different days, instead of only once a month on Mondays. He also suggested going on various fun outings, like to movies and horse races.

## 4. DISCUSSION and RECOMMENDATIONS

Even considering the limitations of time constraint and small sample size, the study provides a strong contribution to the literature. The utility of this evaluation is in the in-depth examination of the programmatic development process, qualitative data analysis, and application of evaluation methods. Additional utility can be derived from the subject matter of this evaluation as it illustrates procedures PWIs have taken to address diversity and retention issues on campus.

Although the evaluators were unable to each individually code all the interviews to help ensure inter-coder reliability, the team did prescribe agreed-upon codes and resulting coding categories based upon their experience with the evaluation and their review of the literature. Furthermore, at least two evaluators attended each transcribed interview, supporting validity as they were familiar with the content of each of the four transcribed interviews. This inter-rater check also served to strengthen reliability. At the onset, there was some confusion on the evaluators' part as to whether the Initiative or the Program was the unit of analysis. This concern quickly dissolved, as the team understood that while these endeavors were separate in name, their intended goals were tightly linked. The Initiative served as the umbrella under which the Program and the ongoing dialogue session resided.

Ideally, the evaluators would have field-tested the instruments with stakeholder groups involved in similar programming at other institutions prior to employing them in this evaluation. Furthermore, time constraints prevented evaluators from following up with interviewees to obtain additional information and clarification of initial responses; to hold a focus group session for student participants; and to conduct more interviews, which would have provided a broader collection of viewpoints and descriptions of the Initiative and its Program.

The overarching question for the researchers was to find out how various stakeholders see the Initiative from their vantage points. The stakeholders' responses indicated that the Program was born out of the Initiative, and exists to help achieve the overall goals of the Initiative. Thus, future researchers should endeavor to see the two seemingly different efforts as such. Meaning, activities of the Program should be assessed in light of how they impact the goals of the Initiative. Furthermore, even though they expressed it in different ways, the interviewed stakeholders recognized the Initiative as an effort to improve the campus experiences of minority students. To quote from the tables above, a student participant said,

> … it was gonna be something that, I don't know, I just really have a genuine interest in and I just wanted to get more, you know, connected with other African American male college students just, you know, to create more relationships and bonds on campus, just because I wasn't really involved in anything that directly connected me to like a diversity type thing. And just being at a school where African Americans are the minority, it's just another opportunity to do something positive, I guess, with people I can more readily identify with.

Another said, "And yeah, like the graduation factor. Having somebody be accountable for you, so that you don't just come here and fail out of college. It's like an intervention tool." The student added the Initiative was to help "... connect us with the resources since we are the minorities on campus, so that we don't feel excluded in anything." A mentor stated that one of the goals of the Initiative "... was to provide these young men forums and situations and with individuals that they can talk to, that they can converse with..." about the issues they face as they strive to graduate. The stakeholders' views were in line with the Initiative's goals of improving racial diversity, campus experiences of minority students, their graduation rates, among others, indicating that administrators had been effective in communicating the goals of the Initiative to its stakeholders.

One of the objevtices of the researchers was to determine if the student participants, administrators, and volunteer mentors were satisfied thus far with the Initiative. In response to this question, a student participant said,

> So I was afraid that this might be something that would be a little bit too much as far as the requirements for like, hours mentoring and different things. But then as I got into it, I figured out that like … the coordinators were more trying to see what we wanted to do and what we wanted to talk about instead of just giving us a bunch of work and time commitments to do, so I really like that aspect of it. They're really just trying to be like a guidance and more of a help to us than pretty much being like army sergeants to boss us around… I really enjoyed the program so far.

Another student participant stated, "[The Program] has allowed me to meet some people and has given me something to be proud that I am a part of on PSEU's campus as far as something that affects the diversity initiative that the school promotes. It makes me feel like I'm being an active participant in that." The student participant added, "Just to have somebody that is similar to me that I can identify with so that I don't feel singled out on campus or anything. And it helps me connect with faculty that are like African Americans who are in higher positions on PSEU's campus and you know, feel more comfortable during my stay here so, yeah." One mentor said, "I think the idea is great. . . I think that the effort that ... administrators are putting into it is a genuine thing and it's something that is needed." The mentor also said, so far "No problem … I'm pretty satisfied." Such responses gave indications that the initial general goal of the Initiative, which focused efforts on providing a place for this particular group of students to discuss their campus experiences, was being achieved. The mentors and student participants were satisfied thus far with the Initiative's trajectory.

At the time of this evaluation, the Initiative, through its Program activities, had already proven beneficial for students involved; however, a lack of resources impeded the realization of all the goals set forth by the Initiative's leadership. From the analysis of the interview data, it appeared that the PSEU neither fully appreciated nor supplied the support required by the Initiative to continue and grow the efforts that are so urgently needed to encourage and retain the African American and Latino men involved in these endeavors. The Initiative is in need of resources—program coordinators and money. There is a need for a designated individual (preferably a paid position) to carry out the ongoing and planned programming activities, instead of faculty members shouldering this responsibility in addition to their regular duties. Indeed, greater PSEU involvement, such as increased promotion and funding of the Initiative's activities, could help bolster the credibility and acceptance of the overall Initiative. Witnessing such PSEU involvement has the potential to encourage more people to participate in the Initiative, which could in turn help improve retention among minority students at the university.

In sum, the evaluators set out to evaluate the Initiative and the Program, which had been modeled, although a modified version, after the BMIs of a nearby PSEU and that of another institution on the east coast. The responses of the stakeholders indicated that the two efforts did not have different goals, rather the Program was an offshoot of the Initiative and its activities were meant to help achieve the goals of the Initiative. The overarching goal of the Initiative was to improve the campus experiences of students of color at the PSEU, and the interviewed stakeholders believed the Initiative was moving in the right direction and achieving its goals. The negative takeaway was the bemoanment of the failure of the institution to recognize the need to make resources available for the Initiative to thrive. Too often in education, proper funding of initiatives is the reason goals are not met or programs do not reach their full potential. This Initiative is an example of an action response to the difficulties African American and Latino male populations are facing at higher education institutions across the country and, based

on the preliminary evaluation, should be supported by the institution and considered by others dealing with similar issues.

## ORCID

Kelly D. Bradley http://orcid.org/0000-0002-4682-8212

## 5. REFERENCES

Coffey, A., & Atkinson, P. (1996). Concepts and coding. In *Making sense of qualitative data: Complementing research strategies*, pp. 26-53. SAGE Publications: Thousand Oaks, CA.

CUNY Black Male Initiative. (2005). http://www1.cuny.edu/sites/bmi/wp-content/uploads/sites/24/page-assets/about/CUNY_BMI_Overview.pdf

Ford, D. Y. (2012). Culturally different students in special education: Looking backward to move forward. *Exceptional Children*, *78*(4), 391-405.

Glesne, C. (2006). Being there: Developing understanding through participant observation. In *Becoming qualitative evaluators: An introduction* (3rd Ed.), pp. 49-78. Pearson Educational: New York, NY.

Gurin, P., Dey, E., Hurtado, S., & Gurin, G. (2002). Diversity and education: Theory and impact on educational outcomes. *Harvard Educational Review*, *72*(3), 330-366.

House, E. R. (1978). Assumptions underlying evaluation models. *Educational Researcher, 7*(3), 4-12.

Initiative Leadership at PSEU. (2013-01-22). Electronic mailing list message. Male Scholars Mentoring Program.

Jackson, B. A. (2012). Bonds of brotherhood: Emotional and social support among college Black men. *The ANNALS of the American Academy of Political and Social Science*, *642*(1), 61-71. doi: 10.1177/0002716212438204.

Jones, L., Castellanos, J., & Cole, D. (2002). Examining the ethnic minority student experience at predominantly White institutions: A case study. *Journal of Hispanic Higher Education, 1*(1), 19-39.

Kim, M. M., & Conrad, C. F. (2006). The impact of historically Black colleges and universities on the academic success of African-American students. *Research in Higher Education, 47*(4), 399-427.

Lascher, E. L. (2008). Retaining Latino and non-Latino college students: Key similarities and differences. *A report for the Serna Center and the Institute of Higher Education Leadership and Policy*, http://www.csus.edu/sernacenter/assets/publications/retention.pdf. Accessed March 06, 2013.

Lewis, W. D., Oliver, S. T, & Burris, J. L. (2011). A work in progress: The lived experiences of Black male undergraduates at one predominantly White university. *Kentucky Journal of Higher Education Policy and Practice, 1*(1), 1-30.

McGee, E. B., & Martin, D. B. (2011). You would not believe what I have to go through to prove my intellectual value! Stereotype management among academically successful Black mathematics and engineering students. *American Educational Research Journal, 48*(6), 1347–89. doi: 10.3102/0002831211423972.

Mishler, E. G. (2003). Representing discourse: The rhetoric of transcription. In Lincoln Y. & Denzin N. (eds.), *Turning Points in Qualitative Research: Tying Knots in a Handkerchief*, pp. 297-326. Alta Mira Press: Walnut Creek.

Office of Diversity and Inclusion-UCO. (2014). Black Male Initiative and Fellows Program. University of Central Oklahoma. http://sites.uco.edu/student-life/diversity/diversity-initiatives/black-male-initiative.asp

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage Publication.

PSEU public relations representative. Document found on PSEU's website (2012). Accessed March 15, 2013.

Redden, E. (2009, July 14). Reaching Black men. *Inside Higher Education.* Accessed March 6, 2013. http://www.insidehighered.com/news/2009/07/14/blackmale

Rippey, R. M. (1973). The nature of transactional evaluation. In Rippey R. M. (ed.), *Studies in transactional evaluation*, pp. 8-13. McCutchan Publishing Corporation: Berkeley, CA.

Roderick, M. (2003). What's happening to the boys? Early high school experiences and school outcomes among African American male adolescents in Chicago. *Urban Education, 38*(5), 538-607. doi: 10.1177/0042085903256221.

Santos, S. J., & Reigadas, E. T. (2002). Latinos in higher education: An evaluation of a university faculty mentoring program. *Journal of Hispanic Higher Education, 1*(1), 40-50.

Seidman, A. (2005). Minority student retention: Resources for practitioners. In Gaither, G H (ed.), *Minority retention: What works?*, pp. 7-24. Jossey-Bass: San Francisco, CA.

Spitzberg, I. J., & Thorndike, V. V. (1992). *Creating community on college campuses*. State University of New York Press: Albany, NY.

Stake, R. E. (1973). Program evaluation, particularly responsive evaluation. *New Trends in Evaluation*, keynote presentation at a conference at the Institute of Education, Göteborg University, Gothenburg, Sweden, pp. 1 - 15. http://education.illinois.edu/circe/Publications/Responsive_Eval.pdf, accessed April 30, 2014.

Strayhorn, T. L. (2010). When race and gender collide: Social and cultural capital's influence on the academic achievement of African American and Latino males. *The Review of Higher Education, 33*(3), 307-332. doi: 10.1353/rhe.0.0147.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research, 1*(1), 77-100.

Thile, E. L., & Matt, G. E. (1995). The ethnic mentor undergraduate program: A brief description and preliminary findings. *Journal of Multicultural Counseling and Development, 23*(2), 116-126.

Thompson, R. A. (1995). What is social support? In *Preventing child maltreatment through social support: A critical analysis*, pp. 43-65. Sage Publications: Thousand Oaks, CA. Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press: Chicago, IL.

University System of Georgia's African - American Male Initiative. (2002). https://www.usg.edu/assets/aami/files/AAMIAssessmentBrochure.pdf

## Appendix A

Minority Male Student Success Initiative Administrator Interview Protocol:

(Administrators #1 and #2)

1. Could you please describe the establishment of the Minority Male Student Success Initiative?
2. How did you become involved in this Initiative, and what motivated you to become involved?
3. Please describe your expectations regarding your role as an administrator when you started working with this Initiative.
4. Please describe your experience with this Initiative so far. How are your Initiative experiences similar to or different from your initial expectations?
5. Is PSEU working in collaboration with other higher education institutions on this type of initiative? If so, please describe this collaboration.
6. If other institutions do have similar initiatives, can you describe those initiatives? Do they take the same form as the Initiative at PSEU? Can you please describe the institutions that house these initiatives.
7. If similar initiatives do exist at other institutions, are they called by the same or different names?
8. Would you say that you are satisfied or dissatisfied to this point with your experience with this Initiative? Why?
9. Would you say that you are satisfied or dissatisfied to this point with the direction of this Initiative? Why?
10. What aspect of the program do you think is the most successful at this point? What aspect do you think is the least successful?
11. What would you change, if anything, about this Initiative?
12. How do you see your future involvement with this Initiative?
13. How would you describe the goal of the Minority Male Student Success Initiative? Do you think that the Initiative is meeting that goal? Why or why not?
14. Could you please describe the establishment of the Program?
15. Are you involved with this program? Why or why not? If you are involved, please describe how.
16. (If the interviewee is actively involved with the Program) If you are involved with this program, please describe your working relationship with the young men involved in the program.
    a.) Do you interact with them outside of the parameters of the program? If not, why not? If so, how often?

    b.) What do you talk with them about?
17. If you are involved with this program, have you noticed any changes in the group of young men from meeting to meeting? If so, what types of changes have you noticed? To what would you attribute these changes?
18. What aspect of the program do you think is the most successful at this point? What aspect do you think is the least successful?
19. Would you say that you are satisfied or dissatisfied to this point with your experience in this program? Why?
20. Would you say that you are satisfied or dissatisfied to this point with the direction of this program? Why?
21. What would you change, if anything, about this program?
22. How would you describe the goal of the Program? Do you think that the program is meeting that goal? Why or why not?
23. How do you see your future involvement with this program?

## Appendix B

Program Student Member Interview
Protocol:

1. How did you learn about this program?
2. Why did you apply to this program?
3. Please describe your experience with the application process for this program.
4. Please describe your initial expectations when you joined this program.
5. Please describe your experience as a participant in this program. How are your program experiences similar to or different from your initial expectations? What, if anything, have you learned from participation in this program?
6. Please describe your experience with the mentoring relationship to this point.
   a.) How often do you have contact with your mentor? Who usually initiates the contact?

   b.) What is your primary form of communication with your mentor? Do you utilize other forms of communication?

   c.) What are the main topics that you discuss with your mentor? How long do your mentoring meetings usually last?
7. How would you describe the characteristics of a good mentor? How would you describe a mentor's role?
8. How would you describe the goal for your mentoring relationship?
9. How would you define personal success? Academic success? Professional success?
10. How do those around you (family members, friends, acquaintances, classmates, etc.) feel about your attending PSEU?
11. What do you like and dislike about PSEU's campus? Why did you choose to attend PSEU?
12. Would you say that you have any personal and/or academic influences (positive or negative) off campus? If so, who or what are they?
13. Please describe your experience as a Black or Latino male here on campus and in your home community. Is there a difference between these experiences? If so, could you please explain?
14. How would you describe Black/Latino culture? If and how does this influence your academic identity?
15. In what ways, if any, has participation in this program affected your campus experience? Your academic experience? Your experience in your home community?
16. What aspect of the program do you think is the most successful at this point? What aspect do you think is the least successful?
17. What serves as motivation for you in your academic and career development?
18. What effect, if any, do you think this program could have on your future as a professional?
19. Do you spend time with other male students in this program outside of meetings or program activities? If not, why not? If so, what types of things do you do in your time together?
20. Would you describe yourself as satisfied or dissatisfied with this program? Why? What would you change, if anything, about this program?
21. How would you describe the goal of the Program? Do you think that the program is meeting that goal? Why or why not?
22. How would you describe the goal of the Minority Male Student Success Initiative? Do you think that the Initiative is meeting that goal? Why or why not?
23. How do you see your future involvement with the Program? Would you recommend this program to others?
24. What year are you in college?
25. How old are you?

## Appendix C

Program Volunteer Mentor Interview Protocol:

1. How did you learn about this program?
2. How did you become involved in this program, and what motivated you to become involved?
3. Please describe your expectations regarding your role as a mentor when you started working with this program.
4. Please describe your experience with this program so far. How are your program experiences similar to or different from your initial expectations?
5. How would you describe the characteristics of a good mentor? How would you describe a mentor's role?
6. Please describe your mentoring relationship with your mentee.

   a.) Please describe a typical mentoring meeting. How long does it usually last?

   b.) How often do you have contact with him?

   c.) What is your primary form of communication? d.) Do you utilize other forms of communication?
7. What topics do you discuss with your mentee?
8. Have you noticed any changes in your mentee from meeting to meeting? If so, what types of changes have you noticed? To what would you attribute these changes?
9. How would you describe the goal for this mentoring relationship?
10. Would you say that you are satisfied or dissatisfied to this point with your involvement in this program? Why?
11. Would you say that you are satisfied or dissatisfied to this point with the mentoring aspect of this program? Why?
12. What would you change, if anything, about this program?
13. How would you describe the goal of the Program? Do you think that the program is meeting that goal? Why or why not?
14. How would you describe the goal of the Minority Male Student Success Initiative? Do you think that the Initiative is meeting that goal? Why or why not?

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 731–739

DOI: 10.21449/ijate.455177

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# The Relationship between Preservice Teachers' Attitudes towards Statistics and Their Research Anxiety

**Adnan Taşgın** [1,*], **Yasemin Kaya** [1]

[1] Atatürk University, Faculty of Education, Department of Educational Sciences, 25100, Erzurum-Turkey

**ARTICLE HISTORY**

*Received: 27 August 2018*
*Revised: 12 October 2018*
*Accepted: 17 October 2018*

**KEYWORDS**

Attitudes towards statistics,
Research anxiety,
Attitude,
Preservice teacher

**Abstract:** The aim of this study was to investigate the relationship between preservice teachers' attitudes towards statistics and their research anxiety. The population of the study consisted of the students of a public university in Turkey who were studying at the Guidance and Psychological Counseling, and the Preschool Education programs. From this population, 257 participants were sampled for this research based on scale sampling method. In order to determine the attitudes and the anxiety of the participants towards statistics and research, two separate scales were to the participants. The scale of attitudes towards statistics consisted of 33 items, which assessed five subscales. The scale of research anxiety consisted of 12 items measuring a single factor. The data were analyzed through independent samples t-test and correlation analyses. The study results suggested that the female preservice teachers had more positive attitudes towards statistics than the male preservice teachers. On the other hand, the male preservice teachers' research anxiety levels were lower than the female preservice teachers. Moreover, the attitudes and the anxiety levels showed significant differences based on the departments of the preservice teachers in favor of the Guidance and Psychological Counseling department. Finally, a positive correlation found between the research anxiety and the attitudes towards statistics levels of the participants.

## 1. INTRODUCTION

While traditional statistical teaching methodology focuses on formulas, calculations and processes, it is usually criticized because of not using statistics to solve real life problems, and because of not focusing on interpreting, criticizing and evaluating skills of students (Mvududu, 2005; Büyüköztürk, 2000; Barnet, 1999; Köklü, 1996; Köklü, 1994; Hogg, 1992; Watts, 1991; as cited in Doğan, 2009).

Statistics is stated as a central science that is used in educational and social sciences (Ridgway, Nicholson, & McCusker, 2007). It is also emphasized that there is no scientific field without statistics (Griffith, Adams, Gu, Hart & Nichols-Whitehead, 2012). Nowadays, statistics is an

CONTACT: Adnan Taşgın adnantasgin@gmail.com Atatürk University, Faculty of Education, Department of Educational Sciences, 25100, Erzurum-Turkey

**ISSN-e: 2148-7456 /© IJATE 2018**

interdisciplinary science that is constantly evolving with new research and methodologies, and it can be applied to every scientific field. Statisticians bring healthy marketing openings to many areas of their countries such as finance, health, economy, agriculture and education with their researcher identities. They also provide notable contribution to solve existing social problems. Especially in our country, the importance of statistics and statisticians is gradually increasing within the scope of European Union studies (Salihova & Memmedova, 2017).

An examination of the students' attitudes towards statistics is important because it might be relevant to learning process (Griffith et al., 2012). It is a fact that students' attitudes towards statistics are known to have the power to influence their success of statistics course. It is stated that statistics courses have the power to affect many university students' academic career (Parker, Pettijhon, & Keillor, 1999; Potter, 1995). Ramirez, Schau and Emmioğlu (2012) emphasize that positive attitudes towards statistics are important for learning. Moreover, individuals with negative statistical attitudes are likely not to use statistics in the future. Dweck (2002) suggests that positive attitudes have a positive effect on students' behaviors, motivations and achievements. In order to investigate the results of technological development on statistics education, Cebeci and Bek (1999) have founded the first school, which provides online statistical education in Turkey, the Alfa Virtual Statistics School. Doğan (2009) conducted a research to determine whether computer supported statistics education effects students' achievement levels in statistical courses and their attitudes towards statistics. The study resulted that the use of computers in the statistics course (using internet, visual materials, statistical software) increased the positive attitudes towards statistics and the achievement level in the statistics course.

Anxiety is another variable of this research that is defined as a condition that can result in unpleasant emotional situations accompanied by physiological symptoms (Feist, 1990). It is also expressed as a sense of tension, which emerges with sadness, worrying thoughts, concern, and the thoughts of something bad is going to happen (TDK, 2017). Today, anxiety is a reaction of human being to various destructive and disruptive situations directed towards the individual's own being or the things he or she identifies (Canbaz, Sünter, Aker, & Peksen, 2007). Anxiety is a complex feeling of unknown origin (Başaran, 2005).

Research anxiety arises from behaviors such as not doing research unless you have to, feeling bored when you need to do research, feeling uncomfortable with the idea of doing research, and feeling nervous and not trusting yourself when doing research (Çokluk-Bökeoglu & Yılmaz, 2005). It is also stated that individuals' academic potential beliefs are influential on their academic achievement (Patall, Awad & Cestone, 2014).

The previous research findings about the students' attitudes towards research and the students' research anxiety concluded that the college students had negative attitudes towards research (Büyüköztürk, 1997) and their research anxiety levels were found moderate (Çokluk-Bökeoglu & Yılmaz, 2005). A study showed that the graduate students' research anxiety levels differentiated across universities, faculties, their grade level, and whether the students took statistics and measurement courses or not (Saracaloğlu, Varol & Evin, 2003). The graduate students showed high research anxiety levels, and their research competence levels decreased as their research anxiety levels increased (Tekin, 2007). Different studies agreed that students experienced high and intense levels of anxiety in statistics and research methods courses (Onwuegbuzie & Wilson, 2003; Sanders, 2001; Trimarco, 1997), and taking a research methods course reduced their research anxiety (Unrau & Beck, 2004). On the other hand, Saracaloğlu (2008) conducted a research over graduate students and contrarily found that their level of academic motivation was adequate, research anxiety was low and attitudes towards research was positive.

The objective of our study is to investigate the relationship between the attitudes of preservice teachers towards statistics and their research anxiety. Within this scope, the following research questions have been developed:

Q1: Are the attitudes of preservice teachers towards statistics differentiated across gender?
Q2: Are the attitudes of preservice teachers towards statistics differentiated across their departments?
Q3: Is there a significant relationship between the attitudes of preservice teachers towards statistics and their research anxiety?

## 2. METHOD

### 2.1. Research Model

The methodology of this study was survey research. Survey research design enables researchers to depict population tendencies, attitudes or opinions qualitatively or quantitatively through the studies applied on the sampled units that were obtained from the population itself (Creswell, 2013).

### 2.2. Population and Sample

The study population existed of the undergraduate students that are studying in the Preschool Education and the Guidance and Psychological Counseling departments at a state university in Turkey. The sample consisted of 257 preservice teachers determined according to the criterion sampling method. The sample in the study was chosen according to a certain criterion which was to take a statistics course in the semester when the study was conducted. The distribution of the sample according to the variables is given in Table 1.

**Table 1.** *Distribution of Sample According to Gender and Department*

| **Variables** | | | **N** | **%** |
|---|---|---|---|---|
| Gender | Female | | 186 | 72.4 |
| | Male | | 71 | 27.6 |
| | | *Total* | *257* | *100.0* |
| Department | Preschool Education | | 107 | 41.6 |
| | Guidance and Psychological Counseling | | 150 | 58.4 |
| | | *Total* | *257* | *100.0* |

### 2.3. Instruments

In the course of the study, the data were obtained by applying two different scales to determine preservice teacher attitudes towards statistics and their research anxiety.

#### *2.3.1. Attitudes towards Statistics Scale*

The scale developed by Yaşar (2014) consists of 33 items and 5 subscales. The subscales of the scale are named as the Relation of Statistics and Professional Life (7 items), Statistical Anxiety-Fear (9 items), Enjoying Statistics (6 items), the Importance of Statistics (6 items), and Statistical Difficulty Perception. The scale is a 9-point Likert-type scale with a level of agreement varying from "I do not agree" (1) to "I definitely agree" (9). Receiving a high score from the scale is an indicator of a positive attitude towards statistics, but it is opposite for a low score. The calculated Cronbach-Alpha internal consistency coefficient for the data obtained from this research sample is $\alpha = .92$.

#### *2.3.2. Research Anxiety Scale*

The "Research Anxiety Scale" developed by Büyüköztürk (1997) is a five-point Likert-type scale consisting of 12 items that are loading on a single factor. The scale has five positive (items

1, 5, 6, 7, 9, 10, and 12) and seven negative (items 2, 3, 4, 8, and 11) items. Negative items are encoded in reverse order. Receiving a high score on the scale indicates low research anxiety, while the low score indicates high research anxiety. The Cronbach-Alpha internal consistency coefficient calculated for the data obtained from this research sample is α = .93.

### 2.4. Data Analysis

In the analysis process, first a missing values analysis was undertaken. Little's MCAR test indicated that the data were missing at random, $\chi^2$ = 201.318, $p$ =.860. Therefore, the missing data were resolved replacing the missing values with the related variable mean. Then, it has been checked whether the data met the assumptions of normality (Shapiro-Wilk, $p$> .05) and equality of variances, which are among the assumptions of parametric tests. After testing the normality of the data, the independent samples t-test was conducted to compare the scores of the two different participant groups. The relationship between two variables was examined through Pearson's correlation coefficient.

## 3. RESULTS

Table 2 shows the results of the independent samples t-test that has been conducted to examine whether the preservice teachers' attitudes towards statistics and research anxiety differ across gender.

**Table 2.** *t-test Results Comparing Males and Females on Attitudes Towards Statistics and Research Anxiety*

| Scale | Gender | N | $\overline{X}$ | sd | *t* | df | p-value |
|---|---|---|---|---|---|---|---|
| Attitudes Toward Statistics | Female | 186 | 188.51 | 39.04 | 2.461 | 255 | .015 |
| | Male | 71 | 174.35 | 46.58 | | | |
| Research anxiety | Female | 186 | 41.16 | 11.54 | -2.027 | 255 | .044 |
| | Male | 71 | 44.39 | 11.13 | | | |

According to the results presented in Table 2, a statistically significant difference was found between the mean attitudes towards statistics scores of the female ( $\overline{X}$ = 188.51) and the male ( $\overline{X}$ = 174,35) preservice teachers, in favor of female individuals, [$t_{(255)}$= 2.461, $p$<.05]. On the other hand, the mean research anxiety scores of the female ( $\overline{X}$ = 41.16) and the male ( $\overline{X}$ = 44.39) preservice teachers were found significantly different in favor of male teachers, [$t_{(255)}$= -2.027, $p$<.05].

The results of the independent samples t-tests that are conducted to examine whether the attitudes of the preservice teachers towards statistics and their research anxiety levels differ across departments are given in Table 3.

Table 3 shows that a statistically significant difference was found between the mean attitudes towards statistics scores of the guidance and psychological counseling ( $\overline{X}$ = 173.21) and the preschool education ( $\overline{X}$ = 192.72) students, in favor of the preschool education students, [$t_{(255)}$= -3.797, $p$<.05]. On the other hand, the mean research anxiety scores of the guidance and psychological counseling ( $\overline{X}$ = 45.15) and the preschool education ( $\overline{X}$ = 37.70) students were found significantly different in favor of the guidance and psychological counseling students [$t_{(255)}$ = -5.3999, p <.05].

**Table 3.** *t-test Results Comparing Departments on Attitudes Towards Statistics and Research Anxiety*

| Scale | Department | N | $\overline{X}$ | sd | *t* | df | Sig. Dif. |
|---|---|---|---|---|---|---|---|
| Attitudes Toward Statistics | Preschool Education | 107 | 173.21 | 44.31 | -3.797 | 255 | .000 |
| | Guidance and Psychological Counseling | 150 | 192.72 | 37.74 | | | |
| Research anxiety | Preschool Education | 107 | 37.70 | 12.85 | -5.399 | 255 | .000 |
| | Guidance and Psychological Counseling | 150 | 45.15 | 9.29 | | | |

Pearson correlation coefficient was estimated to determine the relationship between the attitudes of preservice teachers towards statistics and their research anxiety is calculated. The obtained result is presented in Table 4 below.

**Table 4.** *Correlation Between Attitudes Towards Statistics and Research Anxiety*

| | | Attitudes Toward Statistics | Research anxiety |
|---|---|---|---|
| Attitudes Toward Statistics | Pearson Correlation | | .328 |
| | p | | .000 |
| Research anxiety | Pearson Correlation | .328 | |
| | p | .000 | |

The finding of the relationship between the attitudes of preservice teachers towards statistics and their research anxiety levels are given in Table 4. The result shows that there is a significant positive correlation between the attitudes of the preservice teachers towards statistics and the anxiety about research [$r$ = .328, $n$ = 257, $p$ <.01].

## 4. DISCUSSION

In this study, the relationship between the attitudes of preservice teachers towards statistics and their concerns about research was investigated. The results obtained from the research are explained below.

According to the analyses results that were examining the relationship between the attitudes of preservice teachers towards statistics and their research anxiety across gender, it was concluded that the female preservice teachers had more positive attitudes towards statistics that the male preservice teachers. As the literature is examined, while some research results indicate that the attitudes of female students towards statistics are more positive (Mahmud & Zainol, 2008), some other studies prove that male students' attitudes towards statistics are more positive (Baloğlu, 2003; Tempelaar & Nijhuis, 2007; Zeidner, 1991). More interestingly, some researchers claim no difference across gender in terms of the students' statistical attitudes (Cherian & Glencross, 1997; Martins, Nascimento, & Estrada, 2011; Mji, 2009; Tomazic & Katz, 1988). It can be argued that the variability in these research results can be caused by sample differences. In terms of research anxieties, male preservice teachers were found to have less research anxiety than female preservice teachers, yet there are earlier studies that contradict this finding (Aslan & Karagül, 2016; Büyüköztürk, 1999; Çokluk-Bökeoğlu &Yılmaz, 2005; Saracaloğlu, Varol & Evin-Ercan, 2005; Saracaloğlu, 2008; Trimarco, 1997; Yılmaz & Çokluk, 2010).

The analysis examining the differentiation of both the attitudes of preservice teachers towards statistics and their research anxiety levels across the department variable showed a significant difference for both variables in favor of the guidance and psychological counseling students. In other words, the preservice teachers from the guidance and psychological counseling program showed more positive attitudes towards statistics and were less anxious about doing research compare to those from the preschool education program. This difference might be due to the fact that the major of guidance and psychological counseling is more concentrated on psychology and research, and the students of this program are required to take related courses more intensely. Also, as a result of the requirements of the national university entrance exam, the students who become qualified for admission to the Guidance and Psychological Counseling programs most likely to have positive attitudes towards mathematics, which might be a cause of difference as well. This interpretation is supported by literature which provides evidence of the relationship between attitudes towards statistics, research anxiety, and mathematical background, (Chiesi & Primi, 2010; Lalonde & Gardner, 1993; Nasser, 2004),

It is proven that there is a significant positive relationship between the attitudes of the preservice teachers towards statistics and their research anxiety. This result can be interpreted as the degree to which the preservice teachers having a positive attitude towards statistics will lead them lower research anxiety levels. This finding is supported by literature (Baloğlu, Kocak & Zelhart, 2007; Khavenson, Orel & Tryakshina, 2012; Rosli, Maat, & Rosli, 2017). Similarly, related studies are suggesting that as the positive attitude toward statistics increases, the anxiety level towards statistics reduces. The results of this study are supported by the results of the existing research, which claims to prove that statistical courses and their course exams cause a high degree of anxiety on the students and consequently students' academic performances decrease (Carmona, Martínez & Sánchez, 2005; Chiesi & Primi, 2010; Macher, Paechter, Papousek, & Ruggeri, 2012; Tremblay, Gardner, & Heipel, 2000). In addition, some research results have proven that the students' statistical attitudes and anxiety levels are also related to their mathematical background (Chiesi & Primi, 2010; Lalonde & Gardner, 1993; Nasser, 2004). Sesé, Jiménez, Montaño and Palmer (2015) stated that the students' statistical performances were affected by their math background and statics related anxiety. Zhang et al. (2012) also stated that the attitudes of the students towards statistics could be influenced by age, statistical knowledge, research experience, and mathematical background. The findings of Onwuegbuzie (2000) that claims that roughly 75-80% of the graduate students are concerned about statistics also reveal the importance of graduating from college with a positive attitude towards statistics.

Based on the results of this study, future comparative research is suggested on the statistical attitudes of the preservice teachers and the reasons of the research anxiety differentiations across gender and department variables. In addition, the variables that might influence the attitude towards the statistic should be identified, and the tendencies of the students regarding these variables can be examined.

## ORCID

**Adnan Taşgın** http://orcid.org/0000-0002-3704-861X

**Yasemin Kaya** http://orcid.org/0000-0003-3503-0122

## 5. REFERENCES

Aslan, C., & Karagül, S. (2016). Anxiety levels of graduate students in Turkish language education programme about carrying out scientific research. *Journal of Mehmet Akif Ersoy University Faculty of Education, 38*, 201-217.

Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences, 34,* 855–865. doi:10.1016/S0191-8869(02)00076-4

Baloğlu, M., Kocak, R., & Zelhart, P. F. (2007). The relationship between statistics anxiety and attitudes toward statistics. *Ankara University Journal of Faculty of Educational Sciences, 40*(2), 23–39. doi:10.1501/Egifak_0000000179

Basaran, İ. E. (2005). *Educational psychology.* Ankara: Nobel.

Büyüköztürk, S. (1997). Development of research anxiety scale. *Educational Management in Theory and Practice, 3*(12), 453–464.

Büyüköztürk Ş. (1999). The relationship between anxiety and research in terms of gender, research experience and course achievement. *Education and Science, 23*(112), 29-34.

Canbaz, S., Sürter, A. T., Aker, S., & Peksen, Y. (2007). The anxiety level of the interns in a medical faculty and affecting factors. *Journal of General Medicine, 17*(1), 15–19.

Carmona, J., Martínez, R. J., & Sánchez, M. (2005). Mathematical background and attitudes toward statistics in a sample of Spanish college students. *Psychological Reports, 97,* 53-62. doi:10.2466/pr0.97.1.53-62

Cebeci, Z., & Bek, Y. (1999, May). *Online statistics education: The Alfa Virtual Statistics School.* Paper presented at the meeting of the first Statistics Conference, Turkish Statistical Society and Statistics Association, Belek, Antalya.

Cherian V. I., & Glencross, M. J. (1997) Sex, socioeconomic status, and attitude toward applied statistics among postgraduate education students. *Journal of the Royal Anthropological Institute, 3,* 1385-1386. doi:10.2466/pr0.1997.80.3c.1385

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal, 9*(1), 6-26.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. (4th ed.). London: SAGE.

Çokluk-Bökeoglu, Ö. & Yılmaz, K. (2005). The relationship between attitudes of university students towards critical thinking and research anxieties. *Educational Administration: Theory and Practice, 11*(41), 47-67.

Doğan, N. (2009). The effect of computer-assisted statistics instruction on achievement and attitudes toward statistics. *Education and Science, 34*(154), 3-16.

Dweck, C. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in Surprising Ways). In J. Aronson (Eds.), *Improving academic achievement: Impact of psychological factors on education* (pp. 37–60). USA: Elsevier Science.

Feist, J. (1990). *Theories of personality*. Worth, TX: Holt, Rinehart, and Winston.

Griffith, J. D., Adams, L. T., Gu, L.L., Hart, C. L., & Nichols-Whitehead, P. (2012). Students' attitudes toward statistics across the disciplines: A mixed-methods approach. *Statistics Education Research Journal, 11*(2), 45-56.

Khavenson, T., Orel, E., & Tryakshina, M. (2012). Adaptation of survey of attitudes towards statistics (SATS 36) for Russian sample. *Procedia - Social and Behavioral Sciences, 46,* 2126–2129. doi: 10.1016/j.sbspro.2012.05.440

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioral Science, 25,* 108-125. doi: 10.1037/h0078792

Macher, D., Puechter, M., Papousek,I., & Ruggeri, K. (2012). Statistics anxiety, trait anxiety, learning behavior, and academic performance. *European Journal of Psychology of Education, 27,* 483-498. doi: 10.1007/s10212-011-0090-5

Mahmud, Z., & Zainol, M.S. (2008). Examining postgraduate students' perceived competency in statistical data analysis and their attitudes toward statistics. *International Journal of Education and Information Technologies, 1*, 79–86.

Martins, J., Nascimento, M., & Estrada, A. (2011). *Attitudes of teachers towards statistics: A preliminary study with Portuguese teachers.* Paper presented at the meeting of Seventh Congress of the European Society for Research in Mathematics Education (CERME 7). Rzeszow, Poland: University of Rzeszow and ESRM. Retrieved from http://www.cerme7.univ.rzeszow.pl/WG/5/CERME_Martins-Nascimento-Estrada.pdf

Mji, A. (2009). Differences in university students' attitudes and anxiety about statistics. *Psychological Reports, 104*, 737-744. doi:10.2466/PR0.104.3.737-744

Nasser, F. (2004). Structural model of the effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education, 12*. Retrieved from: http://www.amstat.org/publications/jse/v12n1/nasser.html

Onwuegbuzie A. J. (2000). Attitudes toward statistics assessments. *Assessment & Evaluation in Higher Education, 25*(4), 321–339. doi: 10.1080/713611437

Onwuegbuzie, A. J. & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments-a comprehensive review of the literature. *Teaching in Higher Education, 8*(2), 195-209. doi: 10.1080/1356251032000052447

Parker, R. S., Pettijohn, C. E., & Keillor, B. D. (1999). The nature and role of statistics in the business school curriculum. *Journal of Education for Business, 751*, 51-54. doi:10.1080/08832329909598990

Patall E. A., Awad, G. H., & Cestone, C. M. (2014). Academic Potential Beliefs and Feelings: Conceptual Development and Relations with Academic Outcomes. *Self and Identity, 13*(1), 58-80. doi: 10.1080/15298868.2012.738705

Potter, A. M. (1995). Statistics for sociologists: Teaching techniques that work. *Teaching Sociology, 23*(3), 259- 63.

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal, 11*(2), 57–71.

Ridgway, J., Nicholson, J., & McCusker, S. (2007). Teaching statistics – Despite its applications. *Teaching Statistics, 29*(2), 44–48. doi: 10.1111/j.1467-9639.2007.00276.x

Rosli, M. K., Maat, S. M., & Rosli, R. (2017). Students' attitude and anxiety towards statistics: A descriptive analysis. *Research on Education and Psychology (REP), 1*(1), 47-56.

Salihova, S., & Memmedova, V. (2017). Students attitudes toward statistics lesson: Validity and reliability study. *Journal of Academic View, 59*, 116-127.

Sanders, L. (2001). Reducing students' anxiety in research methods. *Focus on Teacher Education, 1*(3), 3-9.

Saracaloğlu, A.S. (2008). The relationship between post graduate students' academic motivation level, research anxiety and attitudes with their research competence. *Yüzüncü Yıl University, Journal of Faculty of Education, 5*(2), 179-208.

Saracaloğlu, A. S., Varol, R. & Evin, İ. (2003, October). *The relationship between the scientific research anxiety, the attitudes towards research and statistics, and the research competencies of graduate students.* Paper presented at the meeting of the Contribution to Science in Education: Graduate Education Symposium, Dokuz Eylul University Institute of Educational Science, Izmir.

Saracaloğlu, A. S., Varol, S. R., & Evin-Ercan, İ. (2005). The relationship between post graduate students' research competence and scientific research anxiety and their attitudes

towards statistics and research. *Journal of Buca Faculty of Education, 17,* 187-199. Retrieved from http://dergipark.gov.tr/deubefd/issue/25433/268360

Sesé, A., Jiménez, R., Montaño, J.J., & Palmer, A. (2015). Can attitudes towards statistics and statistics anxiety explain students' performance? *Revista de Psicodidáctica, 20*(2), 285-304.

Tekin, M. (2007, October). *Examination of the level of anxiety and proficiency of graduate students across various variables.* Paper presented at the meeting of the third Graduate Education Symposium, Anadolu University Institute of Educational Sciences, Eskişehir.

Tempelaar, D. T., & Nijhuis, J. F. (2007). Commonalities in attitudes and beliefs toward different academic subjects. In M. K. McCuddy, H. van den Bosch, J. W. B. Martz, A. V. Matveev, & K. O. Morse (Eds.), *Educational innovation in economics and business X: The challenges of educating people to lead in a challenging world* (pp. 225–250). Berlin: Springer.

Tomazic, T. J., & Katz, B. M. (1988, August). *Statistics anxiety in introductory applied statistics.* Paper presented at the meeting of the American Statistical Association, New Orleans, LA.

Tremblay, P. F., Gardner, R. C., & Heipel, G. (2000). A model of the relationships among measures of affect, aptitude, and performance in introductory statistics. *Canadian Journal of Behavioral Science, 32,* 40-48. doi: 10.1037/h0087099

Trimarco, K. A. (1997). *The effects of a graduate learning experience on anxiety, achievement, and expectations in research and statistics* (Report No. TM028328). Paper presented at the Annual Meeting of the Northeastern Educational Research Association. (ERIC Document Reproduction Service No. ED 419 022)

Türk Dil Kurumu (TDK). (2017). *Türkçe sözlük [Turkish dictionary].* Ankara, Türkiye.

Unrau, Y.A., & Beck, A.R. (2004). Increasing research self-efficacy among students in professional academic programs. *Innovative Higher Education, 28*(3), 187 - 204. doi:10.1023/B:IHIE.0000015107.51904.95

Yaşar, M. (2014). Attitudes toward statistics scale: Validity and reliability study. *Pamukkale University Journal of Education, 36*(36), 59 - 75. Retrieved from http://dergipark.gov.tr/pauefd/issue/32869/374831

Yılmaz, K., & Çokluk, Ö. (2010). Research anxiety levels of faculty of arts and science graduates. *Journal of Abant İzzet Baysal University, 10*(1), 1-9.

Zhang, Y., Shang, L., Wang, R., Zhao, Q., Li, C., Xu, Y., & Su, H. (2012). Attitudes toward statistics in medical postgraduates: Measuring, evaluating and monitoring. *BMC Medical Education, 12,* 117. doi:10.1186/1472-6920-12-117

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology, 61,* 319-328.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 740–753

DOI: 10.21449/ijate.478277

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Facebook as a Peer-Assessment Platform: A Case Study in Art Teacher Education Context

**Yasemin Ersöz** [1,*], **Süleyman Nihat Şad** [1]

[1] Inonu University, Faculty of Education, Curriculum and Instruction Department, Malatya, Turkey

**ARTICLE HISTORY**

*Received: 04 May 2018*
*Revised: 12 October 2018*
*Accepted: 26 October 2018*

**KEYWORDS**

Facebook,
Peer assessment,
Interactive learning environments

**Abstract:** This research intended to answer the question "Is Peer-Assessment on Facebook useful in visual art education?" in an intrinsic case study. Participants were a group of prospective visual-art teachers, who regularly share and comment on the photographs of their paintings in a special group they created on Facebook. Ten volunteering prospective visual-art teachers were involved in the study during 2013-2014 academic year. Focus group interviews were conducted with the participants to collect data in addition to online digital documents, including photographs of students' paintings and comments on them. In general, participants stated that Facebook-based peer assessment is beneficial, since it helps them notice their deficiencies, look at their works from a different perspective and improve their artistic skills. Thanks to the productive feedback, their motivation and self-confidence are boosted. It was also found that peer-assessment on Facebook has the advantage of ubiquity, allowing more peer involvement, easy and objective criticism, and sustainable learning opportunities in the long-run. The participants also emphasized some disadvantages of the practice of peer assessment on Facebook such as subjective feedback and poor quality of digitalized visuals.

## 1. INTRODUCTION

"*Art educators will need assistance to develop a vision for planning and interpretation of ICT in their curricula*" (Wilks, Cutcher & Wilks, 2012, p. 64).

The transition from teacher-centered instructional approaches to a learner-centered approach implies a change in the conventional measurement and assessment procedures (Uysal, 2008). Unlike teacher-centered conventional assessment approaches, contemporary constructivist approaches advocate that both learning process and outcomes should be evaluated through more learner-based methods. Changing perspectives about assessment has brought about a shift from assessment of learning approach to an assessment for learning approach (Ploegh, Tillema, & Segers, 2009). This shift also suggests that self- and peer- assessment is an integral component of assessment process (Sluijsmans, Dochy, & Moerkerke, 1998). Self- and peer-assessment

CONTACT: Yasemin Ersöz ✉ yaseminersozphd@gmail.com Inonu University, Faculty of Education, Curriculum and Instruction Department, Malatya, Turkey

**ISSN-e: 2148-7456 /© IJATE 2018**

help students understand their own learning better (Wen & Tsai, 2006), become active, autonomous, responsible, and reflective (Sambell & McDowell, 1998). Primarily used for formative purposes (Sluijsman et al., 1998), peer-assessment is a pedagogical and practical evaluation method especially in higher education (Wheater, Langan & Dunleavy, 2005). It provides opportunities for authentic assessment, student autonomy, and collaboration (Loureiro, Pombo & Moreira, 2012).

As one of the most successful Web 2.0 applications (Ozmen, Akuzum, Sunkur & Baysal, 2011) Facebook has become an effective learning or, more specifically, a peer assessment tool in various subjects (Keles & Demirel, 2011; Shih 2011; Suthiwartnarueput & Wasanasomsithi, 2012). Art education is one of the disciplines that theorists encourage art educators to embrace ICT to create new pedagogies for learning and teaching (Wilks et al., 2012). Thus, there is a need to discover and investigate the pedagogical ways of using popular technologies like Web 2.0 for pedagogical purposes. In this regard, present research intended to introduce a case study on a group of visual-art students studying at an education college, who regularly share and comment on the photographs of their paintings in a special group created on Facebook. More specifically, it was aimed to investigate the effect of using Facebook as a platform for peer assessment in visual-art teacher education.

## 1.1. Literature Review

### *1.1.1. Peer Assessment*

One effective strategy used as a part of learning from peers is peer-assessment [PA]. PA has been implemented in schools both as a learning and assessment tool (Lu & Zhang 2012; Ploegh et al. 2009). However, the assessment nature of PA is not summative (determining success), but rather formative (supporting learning) (Patton, 2012; Sluijsman et al., 1998; Topping, 1998). Moreover, even though much assessment in higher education is summative (Topping Smith, Swanson & Elliot, 2000), pedagogical and practical arguments acknowledge that PA, as a learner-centered approach, supports learning and is adopted extensively in higher education context (Bay 2011; Liang & Tsai 2010; Tillema, Leenknecht & Segers, 2011; Weaver & Esposto, 2012; Wen & Tsai, 2006; Wheater et al., 2005).

Topping (1998 p.250) defines PA as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of peers of similar status." PA refers to "those activities of learners in which they judge and evaluate...[products or work]…of their peers with similar learning backgrounds (Sung, Chang, Chiou & Hou 2005, p. 188). PA involves good amount of cognitive investment thanks to increased time on task, thinking, comparing, contrasting, and communicating (Topping, 1998). Thus, teachers are recommended to encourage their students to give especially cognitive feedback to their peers (Lu & Zhang 2012).

### *1.1.2.* Web 2.0 echnologies and Learning

Today it is evident that young people embrace the digital world as intuitive learners of ICT, and "many have developed their skills unprompted and independent of what they are doing at school" (Wilks et al., 2012, p. 64). Thanks to Web 2.0 technologies, blended learning, and peer assessment, students are able to contact with each other, question each other, comment on each other's works, and exchange opinions, which actually refer to the principles of social constructivist theory (Shih, 2011). Web 2.0 is increasingly used for distributed and collaborative learning (Poldoja et al., 2012). Integrating social media with blended learning in higher education seems to be a usable way of supporting student learning (Shih, 2011). Online peer collaborative teaching methods have gradually been expanded, applied, and explored by teachers and researchers (Wu, Hou & Hwang, 2012). Web-based portfolios and assessments have gained popularity due to the accessibility of network and the limitations of their paper-

based counterparts (Chang et al., 2011). Web-technology can provide students with opportunities to interact with their peers without time and space, more diversity and flexibility in designing peer-assessment procedures based on students' demands and needs (Sung et al., 2005).

PA is a mainstay in most online student question-generation systems (Yu, 2011). Popularity of web-based PA increases gradually due to the increasing interest in web-based learning because of the rapid development of internet technologies (Topping, 1998). Liu and Lee (2013) suggest that students can make constructive modifications to their work with the help of feedback from their online peers. Lin, Liu and Yuan (2001, p. 422) list the advantages of online PA procedure: the anonymity, which facilitates willingness to critique; teachers' ability to monitor students' progress at any time; and a decrease in the time and expense of photocopying. Similarly, Wen and Tsai (2006) argue that using the Internet as a tool for implementing PA can provide learners with anonymous environments to express their thoughts and ideas about other students' work freely.

### *1.1.3.* Facebook as a Peer Assessment tool

Today social networks such as Facebook, YouTube, and Twitter are among the most popular especially for Generation Y (Shih, 2011). Facebook is nominated as technically most successful Web 2.0 technology by the authorities and reported to be the most popular one among the educators in US (Ozmen et al., 2011). Facebook enables the users to communicate or interact with each other, create and/or join groups, and share a variety of contents (e.g. text, photo, video, links) both private and public. In particular, thanks to its sharing and commenting function, Facebook allows people to almost instantly discuss and share a large spectrum of information and knowledge, turning it into an online discussion board (Shih, 2011). As an alternative learning tool, Facebook can provide students convenient and attractive means to engage in discussions with the teacher and other users (Suthiwartnarueput & Wasanasomsithi, 2012).

Considering the availability of facilitated communication and interaction features, online systems have been designed by the researchers to implement peer-assessment like the NetPeas designed by Lin et al. (2001); the Web-SPA by Sung et al. (2005), the iLAP by Lu and Zhang (2012) and the DigiMina by Poldoja et al. (2012). Moreover, thanks to providing some diversity and flexibility in designing peer-assessment procedures (Sung et al., 2005), some well-established popular online platforms are also used for similar purposes. For example, Wu et al. (2012) utilized a synchronous communication tool -MSN Messenger- to conduct peer assessment discussion activities. Shih (2011) used Facebook to incorporate peer assessment in college-level English writing classes, which motivated the students to participate in the study and help them enjoyed the learning process thanks to the convenience and popularity of the Facebook platform. Keles and Demirel (2011) used Facebook as a platform for the "Computer-assisted physics" course, where students shared and commented on their assignments for six weeks.

Though digital technologies have a potential, educators need to use them in meaningful, relevant, and worthy ways while creating and criticizing the visual arts (Wilks et al., 2012). We need fresh thinking and new approaches to find out new pedagogical ways to integrate ICT into art classrooms (Wilks et al., 2012). For example, in a rare study on visual art education Lin, Yang, Hung and Wang (2006) found use of a web-based portfolio system and peer-assessment helped students' learning in visual-art education at elementary school. Since there are few studies investigating students' perceptions of online PA specifically, there seems to be a need to examine students' views about online PA (Wen & Tsai, 2006) and PA in visual art teacher education more specifically.

The goal of this study was to investigate the use of Facebook as a platform for peer assessment in visual-art education context. To this end, we analyzed a specific group of visual-art education students studying at an education college, who regularly shared and commented on the photographs of their paintings in a special group created on Facebook. Thus, this research intended to explore how a small group of visual art education students used Facebook as a peer-assessment platform, as well as to better understand the advantages and disadvantages of this particular practice in terms of visual art teacher education.

## 2. METHOD

### 2.1. Design

The present study employed a qualitative intrinsic case study design. Case study designs allow researchers "to retain the holistic and meaningful characteristics of real-life events" (Yin, 2003, p.2). A case may involve "one individual, several individuals, a group, an entire program, or an activity" (Creswell, 2007, p. 74). Furthermore, an intrinsic case study focuses "the case itself (e.g., evaluating a program, or studying a student having difficulty…) because the case presents an unusual or unique situation" (as cited in Creswell, 2007, p. 74). In the present study, the intrinsic case under investigation was the practice of using Facebook as a peer assessment tool by a small group of visual art education students during 2013-2014 academic year. It was considered unusual or unique because this rather informal yet pedagogical and innovative practice was completely initiated and sustained by the learners.

### 2.2. Context and Participants

The investigated case of using Facebook as a peer assessment tool was first noticed by the first researcher during her fieldwork as an instructor. She works at the educational faculty of a medium-scale university located at the eastern part of Turkey. During the 2013-2014 academic year, she was assigned to teach second-year Visual Art Education students a course entitled Instructional Technologies and Material Design. When she introduced the subject digital learning, some of the students mentioned about their specific Facebook group, where they regularly share and comment on the photographs of their paintings. This specific Facebook group was administered by a senior art teaching student the same department. Altough most of the participants in the group came from the same department, any peers interested in the group were welcome.

**Table 1.** Profiles of participating prospective art teachers

| Participants (nicknames) | Profiles |
|---|---|
| Kamil | male, senior art teaching student, the founder and the administer of Facebook group |
| Fatih | male, senior art teaching student |
| Ali | male, senior art teaching student |
| Musa | male, senior art teaching student |
| Erdem | male, senior art teaching student |
| Yasin | male, senior art teaching student |
| Tugce | female, senior art teaching student |
| Hakan | male, second year art teaching student |
| Baris | male, second year art teaching student |
| Suna | female, second year art teaching student |

Being informed about such a specific practice, both researchers decided to study this intrinsic case because it was "unusual and [had] merit in and of itself" (Creswell, 2012, p. 465). Next, researchers asked the group members to volunteer in an academic research. The procedure

rather followed a chain sampling method, where "information-rich key informants or critical cases" (Patton, 2002, p.237) were accumulated by asking each student for the next participant. Eventually, a total of ten voluntary visual-art education students, mostly seniors, who actively involved in the peer assessment process on Facebook accepted to get involved into the study. Detailed information about the participants can be traced in Table 1.

### 2.3. Data collection and analysis

The case study design is characterized with use of multiple sources of data collected through observations, interviews, documents, audiovisual materials, pictures, scrapbooks, e-mails, archival records or physical artifacts (Creswell, 2007, 2012; Yin, 2003). Thus, in order to collect multiple data about whether and how this specific practice of peer-assessment via Facebook worked for visual-art students, face-to-face focus group interviews were conducted with the members of Facebook group in two sessions, in addition to collecting online digital documents, including photographs of students' art and comments on them. During the focus group interviews, voluntary students were asked to explain and comment on their Facebook group and the practice of peer-assessment. The interviews, which lasted about three hours in total, were recorded by a voice-recorder. Obtained data through interviews and documents were analyzed using content analysis method through NVivo10 software program. Throughout the analysis the coding process was compared and contrasted by both researchers. When there was a disagreement between both analysts, briefs were arranged to reach complete consensus. Also, both the process and products of the analysis were exposed to the audit of the external consultants to examine whether the findings, interpretations, and conclusions are credible. Finally, as part of member checking, the participants' views were asked to confirm the findings and interpretations (Creswell, 2007)

## 3. RESULTS

The analysis of the multiple data revealed that the use of Facebook as a Web 2.0 technology for peer assessment purposes is an informal but pedagogical and innovative practice with several advantages and disadvantages to be handled for better results. As it is apparent in Figure 1, the peer assessment at Facebook in art education context has both positive and negative aspects.

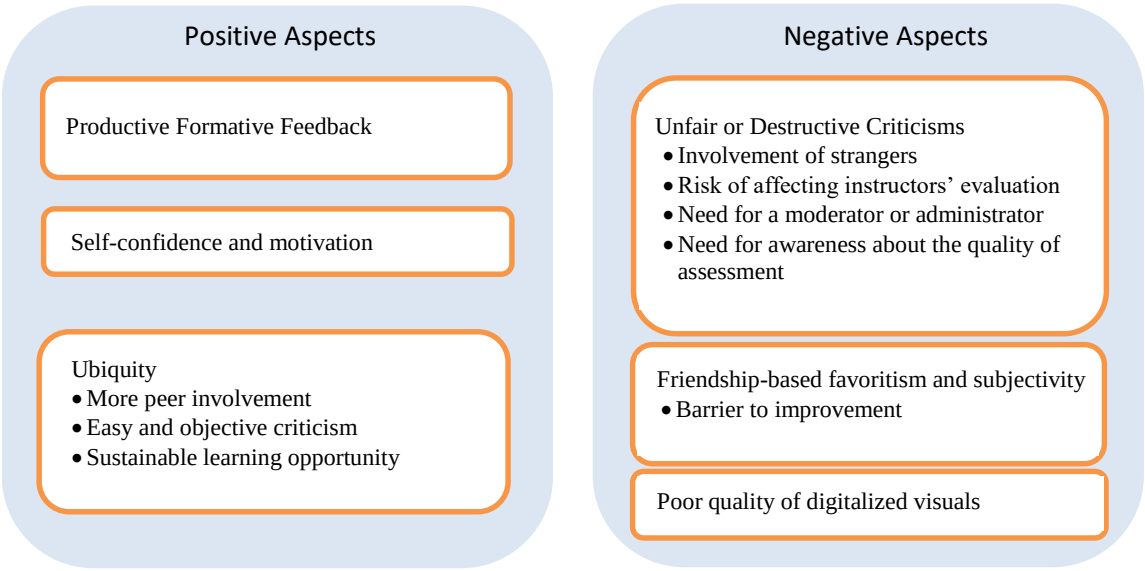



**Figure 1.** Views of participants concerning peer-assessment at Facebook

### 3.1. Productive Formative Feedback

Participants emphasized that publishing a piece of painting on Facebook means exposing it to the artistic appreciation and/or criticism of many potential artists. Receiving practical and formative feedback from others on Facebook helps them get aware of some problematic, defectful, or immature aspects of their work. Below is an example of how a senior, Musa, commented on the proportional deficiencies of the drawing, with two *likes* indicating others also agreed with the feedback.

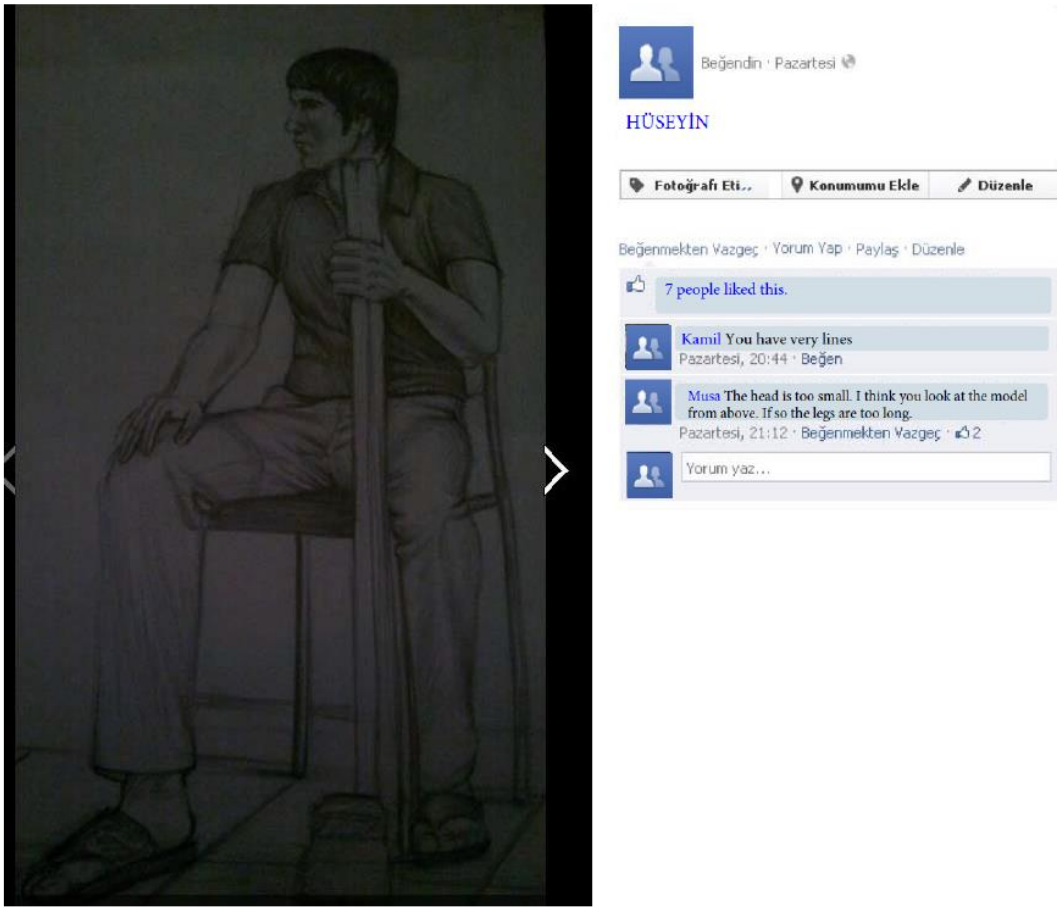



**Figure 2.** A screenshot of example peer assessment of the painting on Facebook

Participants believed such productive feedback is beneficial for the improvement of their artistic skills. Senior art students Fatih, Ali, Musa and Tugce expressed how they formatively benefited from the Facebook sharings as below:

> *Fatih: "Eventually, a painting expresses what it is supposed to express to the extent that the target audience feel, not that you want to express. Therefore, receiving feedback on Facebook tell me to what extend I was able to express what I wanted to."*
>
> *Ali: "Comments on my painting help me notice my defects on the work and contribute to me one hundred percent."*
>
> *Musa: "Criticisms from my friends on my works on Facebook give me idea about how to continue with my work. These criticisms perhaps make the most contribution to my self-assessment.*
>
> *Tugce: "I have benefited a lot from the comments of those people who can look at my work from a different perspective. In case of negative criticisms on my painting, I check over it again, contemplate on whatever impairs the integrity of the painting, and try to concentrate to fix those points."*

### 3.2. Self-confidence and Motivation

In addition to its positive formative function, Facebook-based peer assessment also boosted students' self-confidence and motivation. Participants thought, as expressed by Musa and Baris below, positive peer assessment increased their self-confidence in their artistic skills. Besides, it motivated them to improve their current work or produce new ones. Surprisingly, while some participants (e.g. Kamil and Ali) stated they were motivated especially from positive feedback, some others (e.g. Fatih, and Hakan) asserted that negative feedback was also a source of motivation for them.

> *Musa: "Especially when my friends appreciate my pencil drawings on the net, this gives me confidence."*
> *Baris: "Though it may seem insignificant, actually when our paintings are appreciated by people, we feel happy and confident.*
> *Kamil: "Especially when the comments are positive, we put more importance on our work and we enjoy the work much more."*
> *Ali: "When my peers' feedbacks are positive, this gives me positive reinforcement for my next works."*

As it can be seen in Figure 3, the painting received many *likes* and simple positive comments were made, not necessarily including complex technical evaluations. The peers' dialog seems to be encouraging and motivating others to share their paintings as well.

The source of motivation in case of negative feedback is characterized with some ambition or determination to do better for some students:

> *Fatih: "Actually I like the negative comments more. When negative comments are received, I can say 'Ok, these are my deficiencies".*
> *Hakan: "However, I think negative comments are more advantageous. They help us do better job."*

### 3.3. Ubiquity

Compared to the other advantages, i.e. formative feedback, self-confidence and motivation, which can be alternatively achieved through face-to-face encounters, *ubiquity* seems to be a unique advantage of the Facebook-based peer assessment. Referring to the ability to *happen whenever and wherever needed* (Peng et al., 2009), ubiquitous nature of Facebook-based peer assessment enables the students to share easily their works with more peers or see works of others including those from upper classes in the same department (see Hakan and Baris below) or peers from departments of different faculties at different universities (see Kamil below). Students especially highlight the role of mobile communication tools like smartphones and internet in facilitating the ubiguious nature of Facebook-based peer assessment. Since ubiqutious way of peer-assessment through Facebook has the potential to involve more peers to the process, previously mentioned advantages including formative feedback and self-confidence and motivation are enjoyed more compared to face-to-face peer assessment. Also the distant nature of assessment encourages peers to *criticize more easily and objectively* compared to face-to-face assessment (see Ali below):

> *Hakan: "We are in second year. We feel happy when we have feedback from fellows in upper classes. This helps improve us. Facebook is necessary. We are not always together with our friends. We can communicate more easily via Facebook."*


**Figure 3.** A screenshot of example peer assessment of the painting on Facebook

> *Baris: "It takes longer time if ask some of our friends in person to comment on our works. However, Facebook is fine… for a fact everybody can see and comment on your work. "*
>
> *Kamil: "People from different universities join our group. We can see each other's works. With their participation, we gain different point of views. Internet and smartphones have become an inevitable part of lives and thanks to them I can communicate with peers studying at different universities. We can have access to everybody via the Facebook pages we have developed."*
>
> *Ali: "When asked in face, we cannot always tell each other the negative aspects of a work, thinking that it would be unkind. We can give negative more easily on Facebook. When it is not face-to-face, comments are more realistic."*

Still an advantage of ubiquitous nature of Facebook-based feedback, *sustainable learning opportunity* stands out to be a strong advantage as highlighted by the some students (e.g. Hakan). Sustainability of learning refers to the ability to follow the lesson asynchronously even after the lesson or the term has finished. Participants generally believed that this practice should be a formal or informal part of the lesson, because the time constraints of the lesson is removed as they can reciprocally share their works on Facebook and have comments from the instructor and other peers anytime. Even after the students successfully complete the course, they have the chance to follow the course content and sustainably learn new things.

> *Hakan: "I think it is fine. For example, we can share and have comments on our ongoing works from the instructor and peers at any time on the Facebook group created. Otherwise, we would have to go to the instructor's room, which would be a waste of time both for the instructor and us."*

Although less in number compared to positive ones, some negative aspects were also attributed to Facebook-based peer assessment practice by the students. These negative aspects were categorized under the main themes of *unfair or destructive criticisms*, *friendship-based favoritism and subjectivity*, and *poor quality of digitalized visuals.*

### 3.4. Unfair or Destructive Criticisms

As reported earlier, some participants asserted negative feedback was a source of motivation for them. Nevertheless, some other students (see Suna's comments below) complained about unfair or destructive criticisms about their works exhibited on Facebook. It was clear that these negative criticisms either unfair or destructive offended and discouraged the students:

> *Suna: "If comments are negative, they should be like a remark helping the painter notice mistakes without offending… However, there are unfair and ungrounded criticisms beside the justified ones. There are comments not made to criticize the work but to insult the artist as if the commentator had some personal issues against the artist. And this demoralizes that person. No matter how much you enjoyed painting the assigned homework or appreciated the final work, you suddenly lose your affection for it."*

Unfair or destructive criticism may sound to be a common potential problem in case of face-to-face peer assessment. However, what seems unique to Facebook-based peer assessment is the involvement of strangers. Suna justifies this by saying "*we can change the privacy settings in our personal Facebook accounts, but in group's Facebook account people we don't know may also see our works and some of them just write comments to demoralize on purpose.*"

Fatih and Suna also voiced their concerns about the possibility that these negative criticisms may affect the formal evaluation of the instructor about the assigned work. Thus, most participants were in favor of having a moderator or administrator in charge of monitoring and pre-checking the sharings and comments on the group's Facebook account for appropriacy (e.g. Tugce and Yasin):

> Tugce: "*I believe there should be someone administrating the account. Otherwise everything can get out of hand and very rude comments can appear."*
>
> Yasin: "*Definitely there must be a moderator. He/she can intervene in case things go wrong. Nothing undesired happened so far in my personal account, but we have witnessed many harsh arguments."*

Participants also believed personally that they should be aware whether the criticism comes from a knowledgeable person and should be taken into consideration or not. Some indicators mentioned regarding the quality of assessment were the richness and justification of criticisms, and the quality of the paintings of the person himself.

### 3.5. Friendship-based Favoritism and Subjectivity

Among the negative aspects, the most frequently cited one was the favoritism in peer judgments while assessing someone's work due to friendship. Participants emphasized that generally friends hesitate to criticize and tend to comment more positively their friends' works. Participants also assert that these friendship-based subjective comments pose a major barrier preventing the opportunity for improvement.


**Figure 4.** A screenshot of example peer assessment of the painting on Facebook

> *Kamil: "It happens that peers make positive comments to their close friends. I uploaded the painting of a friend of mine* (see Figure 4). *The work, a pencil drawing, was not very good actually. Yet the comments were all positive. There were only suggestions about minor corrections. Later, the owner of the painting commented on his own painting saying 'awful work'. Despite all the positive comments he made a self-criticism".*

### 3.6. Poor Quality of Digitalized Visuals

Finally some participants admitted that unless the photos of the paintings are taken with high resolution, there is a loss of quality in visuals, as it is exemplified in Figure 2 where the dim light makes it difficult to clearly see the picture. That prevents the peers to clearly see the details and objectively assess the quality of the painting:

> Tugce: "*I think it is better to assess the painting in the workshop, because on Facebook the details cannot be seen. Sometimes, the quality of the photograph may be very poor."*

## 4. DISCUSSION

This intrinsic case study investigated the merit of Facebook-based peer assessment in terms of art teacher education in higher education context. The results suggested that participants were generally positive about the online peer assessment practices they informally and reciprocally performed in their art education classes. Considering that young people have embraced the digital world as intuitive learners of ICT (Wilks et al., 2012), it is inevitable that they explore innovative ways of integrating learning and ICT, including social media tools. Likewise, integrating education with online platforms is a current trend at all levels (Cooke, 2017; Hocevar, 2013; Lau, 2017; Wen & Tsai, 2006). More specifically, using online peer assessment is becoming a popular practice for different subjects at various levels (see Loureiro et al., 2012; Thomas, Martin & Pleasants, 2011). For example, in an initial attempt Liang and Tsai (2010) utilized online peer assessment to facilitate students' learning by science writing. Liu and Lee (2013) investigated the influence of peer observation and feedback on learning Statistics in Education and Psychology, which resulted in improved learning and positive impressions on the part of students. Shih (2011) investigated the effect of incorporating Facebook and peer assessment with English writing class, which proved to be interesting, motivating, cooperative and effective.

The main reason for the positive views towards using Facebook as a peer assessment tool was the formative feedback students received which worked productively in improving their paintings. The formative and productive nature of peer assessment is the most frequently cited pedagogical strength of it. Previous research revealed that students find peer-assessment or feedback valuable because it helps them see their mistakes and shortcomings from different perspectives (Bay 2011), make valuable modifications to their work (Liu & Lee 2013), improve their assignments (Rubin & Turner 2012), prevent some errors and provide hints for making progress (Lin et al., 2001), and increase their awareness of their strengths and weaknesses (Koc 2011). More specifically, recent studies on online peer assessment practices proved that it is favorable and beneficial on the part of students (Liang & Tsai 2010; Liu & Lee 2013; Loureiro et al., 2012).

This productive feedback, no matter negative or positive, motivated or increased students' self-confidence to produce better works. This kind of increase in motivation and self-confidence is known to characterize peer-assessment, which in turn brings better academic achievement (Topping, 1998).

Among others, ubiquity was found to be the unique advantage of the Facebook-based peer assessment as it refers to the ability to happen whenever and wherever needed (Peng, Su, Chou & Tsai 2009). Ubiquitous learning environments enable "access (to) learning content from anywhere at any time, and to communicate with colleagues or lecturers synchronously and asynchronously much more frequently" (Hummel & Hlavacs, 2003, p.1). Thanks to the ubiquitous nature of Facebook-based peer assessment practice, more peers can involve assessment process reciprocally, make criticism more easily and objectively, and enjoy the opportunity to earn sustainably. This is actually the most important reason that web-based portfolios and assessments have gained popularity: due to the accessibility of network and the limitations of their paper-based counterparts (Chang et al., 2011). In similar studies, Facebook or similar web-technologies were also reported to facilitate peer interaction during peer assessment (Shih 2011), provide students with more opportunities of peer interaction beyond the constraints from time and locations (Sung et al., 2005); peers other than the class community take part in peer assessment process (Keles & Demirel, 2011). In lifelong learning perspective, the use of Facebook in art education for peer assessment can be said to be a contemporary ubiqutious practice as against its classroom-base face-to-face alternative, since both knowledge and skills cannot be confined to the school-based training only (Uysal, 2008).

Contrary to the pros about the Facebook-based peer assessment practice in art education, there were also cons mentioned. One major disturbing aspect was the perceived unfair or destructive criticisms especially coming from the strangers, which makes the assessed students feel offended or discouraged. Similarly, Koc (2011) found that students can feel under pressure and anxious about not being assessed objectively online. Participants were further stressed that these negative criticisms might affect the instructors' evaluation. Although peer-assessment is more pedagogically used for formative purposes (Sluijsman et al., 1998) and its use for scoring purposes is not welcome especially by students (Lin et al., 2001; Patton, 2012; Wen & Tsai, 2006), this still may be the case. However, peer assessment should not only be used for scoring (Bay, 2011). If used for scoring purposes, PA cannot be reliable or valid because some students may tend to give extremely low or high scores (Lin et al., 2001). Moreover, using peer-assessment for summative purposes may spoil the rapport in class. Thus, it should be made clear in such online platforms that formative peer responses or criticisms are not done to measure academic achievement but to improve the work. Also a recommendation to prevent this unfair or destructive critics was to assign a moderator or administrator. In Shih's (2011) work on online peer assessment, for example, instructor himself served as a facilitator and monitor evaluating and commenting on students' work and responses (Shih, 2011). For better

results, it is also suggested to set-up, monitor or manage the peer assessment processes (Topping, 1998; Wheater et al., 2005). Finally, a last recommendation against unfair and destructive criticisms was to raise awareness among the peers about distinguishing between high quality and low quality criticisms. As a matter of fact, assessing peers may not have adequate knowledge to evaluate others' works (Lin et al., 2001). Thus, before assessing peers should be trained about the rules, principles and criteria (Bay, 2011; Cihanoğlu, 2008).

On the other end of the continuum, there were ungrounded positive assessments by close friends. Students complained about being complimented, because it is a barrier to improvement. Friendship-based subjectivity of peer assessment process has been widely criticized in previous research (Lin et al., 2001; Uysal, 2008). To prevent this it can be suggested that extensive training should be given (Sluijsmans et al., 1998), and teachers or whoever is in charge should ask students to be specific in their feedback, particularly with regard to the problems in assessees' work, and to provide suggestions (Lu & Law, 2012).

Last but not the least, most critical criticism about using peer-assessment in art education on Facebook was about the poor quality of digitalized visuals. While the merit of producing art with digital technologies is disputable (Wilks et al., 2012), reproducing it, even in poor quality, for assessment purposes cannot be acceptable. Then it becomes a matter of validity and reliability against usability. That is, despite the availability of a highly usable (ubiqutious) online platform, assessors cannot assess the real merit of the painting because they cannot see the details. Thus, it is highly recommended that peers photograph their works using high resolution settings.

## 5. CONCLUSIONS

We attempted to answer whether peer-assessment on Facebook work in visual art education through an intrinsic case study. For the investigated group, Facebook-integrated peer-assessment practice yielded productive and innovative results in terms of visual-art education. Most of the students had positive opinions regarding peer observation and comments via Facebook. In general, participants stated that such peer assessment is beneficial, since it helps them notice their deficiencies, look at their work from a different perspective and improve their artistic skills; thanks to this productive feedback their motivation and self-confidence is boosted. It was also concluded that peer-assessment on Facebook has the advantage of ubiquity, allowing more peer involvement, easy and objective criticism, and sustainable learning opportunities. Yet, there are some drawbacks regarding this practice. The participants also emphasized the disadvantages of the practice of peer assessment on Facebook, which included the subjective comments (either favorable or unfavorable) biased according to the degree of friendship, and destructive comments demotivating and discouraging the students from sharing their paintings. A rather technical drawback of Facebook assessment via photographed painting was the deterioration in the visual quality of the painting, which allegedly affected the accuracy of assessment.

**Acknowledgement**

The findings of the research were presented as a paper at International Congress on Education for the Future: Issues and Challenges (ICEFIC 2015) held on 13-15 May 2015, in Ankara University, Turkey

**ORCID**

Yasemin Ersöz http://orcid.org/0000-0001-7386-2961
Süleyman Nihat Şad http://orcid.org/0000-0002-3169-2375

## 6. REFERENCES


Bay, E. (2011). The Opinions of Prospective Teachers about Peer Assessment. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi*, *5*(2), 909 -925.

Chang, C., Tseng, K., Chou, P., & Chen, Y. (2011). Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education*, *57*, 1306–1316.

Cooke, S. (2017). Social teaching: Student perspectives on the inclusion of social media in higher education. *Education and Information Technologies*, *22*(1), 255-269.

Creswell, J.W.(2007). *Qualitative inquiry and research design: Choosing among five approaches*. Sage.

Creswell, J.W. (2012). *Educational research: Planning, conducting, and evaluating quantitative* (4th edition). Baston: Pearson Education, Inc.

Hocevar, K.P. (2013). *What is social about social media users? How social media efficacy impacts information evaluation online*. (Unpublished master thesis) University of California, Santa Barbara.

Hummel, K.A., & Hlavacs, H. (2003). Anytime, anywhere learning behavior using a web-based platform for a university lecture. In *Proceedings of the SSGRR 2003 Winter Conference, L'Aquila, Italy*.

Koc, C. (2011). The Views of Prospective Class Teachers about Peer Assessment in Teaching Practice, *Educational Sciences: Theory & Practice, 11*(4), 1979-1989.

Lau, W. W. (2017). Effects of social media usage and social media multitasking on the academic performance of university students. *Computers in Human Behavior*, *68*, 286-291.

Liang, J., & Tsai C. (2010). Learning through science writing via online peer assessment in a college biology course. *Internet and Higher Education*, *13*, 242–247.

Lin, K.C., Yang, S.H., Hung, J. C., & Wang, D.M. (2006). Web-based appreciation and peer-assessment for visual-art education. *International Journal of Distance Education Technologies*, *4*(4), 5-14.

Lin, S., Liu, E., & Yuan, S.M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*, 420-432.

Liu, E., & Lee, C. (2013). Using peer feedback to improve learning via online peer assessment. *TOJET: The Turkish Online Journal of Educational Technology*, *12*(1), 187-199.

Loureiro, M.J., Pombo, L. & Moreira, A. (2012). The quality of peer assessment in a wiki-based online context: a qualitative study. *Educational Media International*, *49*(2), 139-149.

Lu, J., & Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instr Sc*i, *40*, 257–275.

Lu, J., & Zhang, Z. (2012). Understanding the effectiveness of online peer assessment: a path model. *J. Educational Computing Research, 46*(3), 313-333.

Ozmen, F., Akuzum, C., Sunkur, M. & Baysal, N. (2011). Sosyal Ağ Sitelerinin Eğitsel Ortamlardaki İşlevselliği [Functionality of Social Networks in Educational Settings], *6th International Advanced Technologies Symposium*, 16-18 May 2011, Elazığ, Turkey.

Patton, M.Q. (2002) Qualitative research and evaluation methods (3rded.). Thousand Oaks, California: SAGE Publications.

Patton, C. (2012). 'Some kind of weird, evil experiment': student perceptions of peer assessment, *Assessment & Evaluation in Higher Education*, *37*(6), 719-731.

Peng, H., Su, Y.-J., Chou, C., & Tsai C.-C. (2009). Ubiquitous knowledge construction: mobile learning re-defined and a conceptual framework. *Innovations in Education and Teaching International*, *46*(2), 171–183.

Ploegh, K., Tillema, H.H., & Segers, S. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*, 102–109.

Poldoja, H., Valjataga, T., Laanpere, M., & Tammets, K. (2012). Web-based self- and peer-assessment of teachers' digital competencies, *World Wide Web, 17*, 255–269.

Rubin, R.F., & Turner, T. (2012). Student performance on and attitudes toward peer assessments on Advanced Pharmacy Practice Experience assignments. *Currents in Pharmacy Teaching and Learning, 4*, 113–121.

Sambell, K., & McDowell, L. (1998). The Construction of the Hidden Curriculum: messages and meanings in the assessment of student learning. *Assessment & Evaluation in Higher Education*, *23*(4), 391-402.

Shih, R.C. (2011). Can Web 2.0 technology assist college students in learning English writing? Integrating Facebook and peer assessment with blended learning, *Australasian Journal of Educational Technology*, *27*(5), 829-845.

Sluijsmans, D., Dochy, F. and Moerkerke, G. (1998). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research*,1, 293–319.

Sung, Y., Chang, K., Chiou, S,. & Hou, H. (2005). The design and application of a web-based self- and peer-assessment system. *Computers&Education, 45*, 187–202.

Suthiwartnarueput, T., & Wasanasomsithi, P. (2012). Effects of using Facebook as a medium for discussions of English grammar and writing of low-intermediate EFL students. *Electronic Journal of Foreign Language Teaching*, *9*(2), 194-214.

Thomas, G., Martin, D. & Pleasants, K. (2011). Using self- and peer-assessment to enhance students' future-learning in higher education. *Journal of University Teaching & Learning Practice, 8*(1), 1-17.

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning: A review of research studies. *Studies in Educational Evaluation, 37*, 25–34.

Topping, K.J., Smith, E.F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing between Postgraduate Students. *Assessment &Evaluation in Higher Education*, *25*(2), 149-169.

Topping, K. J. (1998). Peer Assessment between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249-276.

Uysal K. (2008). *Involving students in the assessment process: Peer assessment and self assessment* (Unpublished Master Thesis). Abant İzzet Baysal University, Bolu, Turkey.

Weaver, D., & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment & Evaluation in Higher Education*, *37*(7), 805-816.

Wen, M.L., & Tsai C.C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education*, *51*, 27–44.

Wheater, C.P., Langan, M. & Dunleavy, P.J. (2005). Students assessing student: case studies on peer assessment. *Planet*, *15*, 13-15.

Wilks, J., Cutcher, A., & Wilks, S. (2012). Digital technology in the visual arts classroom: an [un]easy partnership. *Studies in Art Education, 54(1)*, 54-65.

Wu, S., Hou, H., & Hwang, W. (2012). Exploring Students' Cognitive Dimensions and Behavioral Patterns during a Synchronous Peer Assessment Discussion Activity Using Instant Messaging. *The Asia-Pacific Education Researcher, 21(3)*, 442-453.

Yin, R.K. (2003) *Case study research: Design and methods*. Sage publications.

Yu, F.Y. (2011). Multiple peer-assessment modes to augment online student question-generation processes. *Computers&Education, 56*, 484–494.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 4, 754–762

DOI: 10.21449/ijate.481162

Published at http://www.ijate.net http://dergipark.gov.tr/ijate *Research Article*

# Interpersonal Emotion Regulation Scale (IERS): Adaptation and Psychometric Properties in a Turkish Sample

**Asude Malkoç [1,*], Meltem Aslan Gördesli [1], Reyhan Arslan [1], Ferah Çekici [1], Zeynep Aydın Sünbül [1]**

[1] Istanbul Medipol University, Faculty of Education, Department of Psychological Counseling and Guidance, Kavacık - Beykoz 34810, Istanbul

**ARTICLE HISTORY**

*Received: 29 September 2018*
*Revised: 05 December 2018*
*Accepted: 11 December 2018*

**KEYWORDS**

Emotion regulation,
Interpersonal emotion regulation,
Validity,
Reliability

**Abstract:** This study aims to test the validity and reliability of the Turkish adaption of Hofmann, Carpenter and Curtis' (2016) Interpersonal Emotion Regulation Scale (IERS). The original scale is comprised of four sub-dimensions; namely, enhancing positive affect, perspective taking, soothing, and social modeling. The study was carried out with 326 students from various departments of Medipol University. Work on adapting the scale began with an attempt to find linguistic equivalence. After ensuring this linguistically equivalence for the scale's original form, a Confirmatory Factor Analysis was launched to examine its construct validity. The results of this confirmatory factor analysis revealed that the four-factor original structure of the scale was also valid for the Turkish sample and the goodness of fit indices of the scale was within acceptable limits. The Cronbach-Alpha internal consistency coefficient was found as .92 for the overall scale. The scale shows outstanding psychometric characteristics.

## 1. INTRODUCTION

In psychology, understanding of the effects and importance of emotion on personal development, evolution, and psychological health has gained much ground as a subject of research. Emotions play a role in the process of an individual's learning to adapt to their environment and influence their behaviour and decisions (Greenberg, 2011). The subject has been researched in various contexts, including motivation, social interactivity, self-regulation, and mental Health (Berking & Wupperman, 2012; Cote, 2005; Diamond & Aspinwall, 2003; Gross & Munoz, 1995) One of the most important subjects related to emotion, however, is emotion regulation.

In recent years, emotion regulation has become a popular area of psychological research (Hofmann, Carpenter and Curtis, 2016). The concept, first used in the early 1980s (Gaensbauer, 1982), appears to have gained prominence in both developmental and adult psychology (Gross,

CONTACT: Asude Malkoç ✉ amalkoc@medipol.edu.tr İstanbul Medipol Üniversitesi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, PDR Anabilim Dalı, Güney Kampüs, Kavacık Kavşağı - Beykoz 34810 İstanbul

**ISSN-e: 2148-7456 /© IJATE 2018**

1999). This study deals with emotion regulation in terms of how it has been approached in the field of adult psychology. According to Thompson (1994), emotion regulation refers to "one's ability to realize one's goals through the control, observation, evaluation, and adaptation of one's inner and outer emotional responses." This definition defines regulation and adaptation of one's emotional responses not only as a strategy of self-regulation, but also refers to it as a means of interpersonal processes involved in participating in relationships.

According to Gross (1998b), emotion regulation is defined as a process which includes awareness of what emotions one is feeling, when these emotions are being experienced, how they are being experienced, and how they express themselves. According to Gross (1999a), this process includes conscious or unconscious strategies such as modifying feelings, behaviour and physiological responses and retaining balance. Gross' emotion regulation model (2001) is founded on the pillars of reappraisal and suppression in the process of emotion regulation. Reappraisal is a cognitive process. Before providing an emotional response to a situation, an individual develops cognitive strategies to deal with it (Garnefski, Kraaij and Spinhoven, 2001) and reevaluates the situation which tiers these feelings as well as the meaning of the situation itself. This evaluation allows for an emotional shift in the individual (Northoff et al. 2006). Suppression, meanwhile, prevents behaviour in which emotions are expressed. Efforts geared toward preventing and suppressing the expression of emotions causes physiological changes in the individual. As an example, it has been shown that those who embark on the path of suppression in regulating their emotions may appear to have more contracted veins than others (Gross, 2001).

The attention owed to emotion regulation has surged thanks to many extensive on-going studies. These studies have shown that emotion regulation is an important factor in allowing individuals to cope with stressful events (Gross, 1998a, 1999b, 2002) and increases wellbeing (Cicchetti, Ackerman and Izard, 1995; Gross & John, 2003). Furthermore, it has been shown that emotion regulation has a positive link with mindfulness (Goldin & Gross, 2010) and social relations (Lopes, Salovey, Côté, Beers and Petty, 2005), and a negative link to variables such as depression (Garnefski & Kraaij, 2006) and vulnerable psychological states (Kashdana, Barrios, Forsyth and Steger, 2006).

Although important from a theoretical and developmental standpoint, when studied in a more personal light, emotion regulation does not seem to address the roles played in interpersonal relations (Hofmann, Carpenter and Curtis, 2016). The experiencing and expression of emotions plays an important role not only in personal but social processes (Hofmann, 2014). This has lead Higgins and Pittman (2008) to claim that emotion regulation is of critical importance to the individual's socialization processes, whilst Einsberg, Fabes, Guthrie and Reiser (2000) stress that in parallel with, the socialization process is one in which individuals are affected by the reactions of others.

Interpersonal emotion regulation is a concept describing the interpersonal context of people's emotions being regulated by others (Hofmann, Carpenter and Curtis, 2016). More broadly, it is a process regulated by the scope of social relations in which the emotional response to various stressful situations are experienced in tandem with others. As part of this process, fostering social connections in the reduction of an individual's negative emotions (Berscheid, 2003; Coan, Schaefer and Davidson, 2006), sharing emotional experiences with others (Rime, 2007, 2009), and allowing for emotional calm (Tamir & Millgram, 2017) is important.

The most important aspect of interpersonal emotion regulation is the ability to explore how one can benefit from other individuals when regulating one's emotion. Zaki and Williams' (2013) proposed interpersonal emotion regulation model divides response-dependent and response-independent interpersonal emotion regulation and identifies what sets them apart. Furthermore, it focuses on the difference between internal and external interpersonal emotion regulation.

Internal interpersonal emotion regulation begins when the individual experiences an emotion and set about to modify it. If the individual appeals to others for help in achieving this, and the emotion can be regulated thanks to the support of the other, then this is labelled internal and response-dependent emotion regulation. If an individual can regulate their emotions (identifying, evaluating, and experiencing them) without the assistance of another, then this is a response-independent process. As for external regulation, this refers to processes experienced by another (Zaki & Williams, 2013). In other words, if an individual is required to assist another in their emotion regulation processes, he/she identifies the goals for regulating the other person's emotions and act according to this goal.

One of the most important requirements for future studies related to the subject of interpersonal emotion regulation is to have a valid and reliable means of measuring the phenomena. According to literature on the field, Niven, Totterdell, Stride and Holman (2011) were the first to develop the "Emotion Regulation of Others and Self" scale. The results of a construct validity analysis on some of the items included in sub-divisions of the scale, however, showed that it appeared to be insufficient. In response to this, Hofmann, Carpenter and Curtis (2016) developed a new measure known as the Interpersonal Emotion Regulation Scale. The validity and reliability analyses conducted on this showed it included strong psychometric features. Due to the prior lack of a measuring device for gauging interpersonal emotion regulation in Turkey, this study set out to adapt the Interpersonal Emotion Regulation Scale developed by Hofmann, Carpenter and Curtis (2013) into Turkish and examine its psychometric features. It is hoped that the findings of the study will be used in relevant future research and practical studies employing the scale.

## 2. METHOD

### 2.1. Participants

The study was conducted using data attained from 326 students from various departments of Istanbul Medipol University. 223 of the participants were female and 103 were male students. The age of participants varied between 18-53 with the average of 20.95 (SD= 3.22). The students participated in the study voluntarily. Some demographic information regarding the participants is presented in Table 1. A convenient sampling technique was used to determine the sample due to the time and cost benefits of this method to the researcher (Fraenkel, Wallen and Hyun, 2011; Marshall, 1996). The departmental distribution of the sample was 182 students (55.8%) from psychological guidance and counseling program, 59 students (18.1%) from psychology program, 42 students (12.9%) from international trade program, 26 students (8.0%) from math teacher education program, 17 students (5.2%) from business programs.

### 2.2. Translation Work

In order to adapt the Interpersonal Emotion Regulation Scale into Turkish, we first set out to contact one of the researchers responsible for its development, Stefan G. Hoffman, to ask for permission to adapt his scale and access all items. Next, Istanbul Medipol University's Social Sciences Ethics Committee was convened in order to be granted ethical permission. With the ethics committee's permission obtained, the scale was then translated into Turkish by six professionals working independently of one another, including 1 staff member of ELT department and 5 staff members of the Psychological Counseling and Guidance department – all of whom possessed a high command of English. The translated forms were then reviewed by the researchers in terms of comprehensibility and a Turkish version of the scale was drawn up. The form was then translated back into Turkish by two research staff from the English Language Teaching department. After these were compared, the final version of the Turkish form was completed. After this, both English and Turkish language versions were given to a group of 20 students in their second year of their English Language Teaching Studies at two

weeks interval. The Pearson Correlation Coefficient between the two tests was found as r=0.82, p<0.01 which refers to positive and significant correlation. Accordingly, it was agreed that the linguistic equivalence had been achieved. After the translation work had been undertaken and satisfying results yielded, the scale was applied to 326 students from various departments of Istanbul Medipol University.

### 2.3. Instruments

***2.3.1. Interpersonal Emotion Regulation Scale.*** The scale developed by Hofmann, Carpenter and Curtis (2016) measures interpersonal emotion regulation. The scale consists of 20 items and four sub-dimensions set out in a 5-point Likert-type scale. These sub-dimensions refer to enhancing positive affect, perspective taking, soothing, and social modeling. The lowest score to be taken from the scale is 20, whilst the highest score is 100. The results of the confirmatory factor analysis (CFA) showed that the emerged model fit indices support the four-factor structure of the scale (CFI = 0.97, NNFI = 0.97, RMSEA = 0.04). In addition, the internal consistency indicator of Cronbach alpha was .89 for enhancing positive affect, .91 for taking others' perspectives, .94 for soothing and .93 for social modeling sub-dimensions. These results show that the scale possesses strong psychometric validity. The Turkish adaption of the scale and a study of its psychometrics will be included as part of this study.

### 2.4. Data Analysis

The construct validity was conducted in order to determine the validity of the Interpersonal Emotion Regulation Scale. A confirmatory factor analysis (CFA) was conducted to confirm the scale's construct validity. In interpreting the values produced as a result of the CFA in this study, a Chi-Square Goodness of Fit test, RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), GFI (Goodness of Fit Index) and TLI (Tucker-Lewis Index) were used. The reliability of the scale was investigated with an internal consistency coefficient method. In determining the internal consistency, the Cronbach alfa coefficient was used (Kline, 2011). Analysis of data was performed using the SPSS 20 and AMOS 18 programmes.

## 3. FINDINGS

### 3.1. Preliminary Analysis

For reviewing the construct validity of the scale, Confirmatory Factor Analysis was used to check evidence over the proposed four-factor structure of Interpersonal Emotion Regulation Scale (Hofmann, Carpenter & Curtis, 2016). First of all, missing values were screened, and mean substitution was conducted due to the fact that the number of empty cells were less than 5% of the total cells. Then, a number of assumptions for CFA were tested before the main analysis. The sufficient sample size in CFA is suggested to be minimum 200 cases along with 5 or 10 units that is 326 for this study satisfying that requirement (Kline, 2011). In addition, univarite outliers were examined based on the ±3.29 criterion for the z scores yielding no data staying out the criterion ranges. Lastly, screening of Skewness and Kurtosis parameters for normality assumption and bivariate scatterplots for the linearity requirement of CFA disclosed that the data has a normal and linear distribution for the sample (Tabachnick & Fidell, 2006).

### 3.2. Item Analysis

An item analysis was conducted in order to investigate the distinctiveness of each item. Table 2 provides the mean, standard deviation, and item-total correlation of the scores of participants on the Interpersonal Emotion Regulation Scale. The results of the item analysis showed that the item-total correlation ranged between 0.48 and 0.80.

**Table 2.** Descriptive statistics regarding the Interpersonal Emotion Regulation Scale

| Item Number | Mean | Standard Deviation | Item-Total Correlations |
|---|---|---|---|
| 1 | 3.67 | 0.99 | 0.60 |
| 2 | 3.55 | 1.03 | 0.66 |
| 3 | 4.26 | 0.89 | 0.48 |
| 4 | 3.60 | 1.21 | 0.70 |
| 5 | 3.56 | 1.10 | 0.77 |
| 6 | 4.27 | 0.91 | 0.55 |
| 7 | 2.92 | 1.27 | 0.57 |
| 8 | 4.17 | 0.90 | 0.57 |
| 9 | 3.73 | 1.17 | 0.70 |
| 10 | 2.99 | 1.20 | 0.62 |
| 11 | 3.51 | 1.14 | 0.76 |
| 12 | 3.66 | 1.18 | 0.77 |
| 13 | 3.81 | 1.07 | 0.51 |
| 14 | 2.76 | 1.22 | 0.59 |
| 15 | 3.30 | 1.18 | 0.80 |
| 16 | 3.11 | 1.22 | 0.65 |
| 17 | 3.09 | 1.11 | 0.70 |
| 18 | 4.19 | 0.96 | 0.50 |
| 19 | 3.24 | 1.22 | 0.77 |
| 20 | 3.35 | 1.21 | 0.68 |

## 3.3. Construct Validity

### *3.3.1. Model fit indices and standardized parameter estimates for IERS*

Following to obtain satisfactory results for the requirements of CFA, a Maximum likelihood estimation method was used to validate the four-factor framework of IERS by running AMOS 18 program (Byrne, 2001). In the first step, the model fit indicators were checked. The results of these indices are shown in Table 3.

**Table 3.** Model fit indices from measurement models of IERS

| Goodness of Fit Indexes | Measurement Model of IERS | Criterion Ranges |
|---|---|---|
| $\chi 2/df$ | 2.53 | $\chi^2/df < 3$ |
| CFI | .94 | .90 < CFI or close to 1 |
| TLI | .93 | .90 < TLI or close to 1 |
| RMSEA | .07 | .05 < RMSEA <.08 |
| GFI | .90 | .90 < GFI |

As presented in Table 3, the normed chi square indicator of 2.53 is satisfactory due to being lower than criterion value of 3 (Kline, 2011). Likewise, both CFI (.94) and TLI (.93) and GFI (.90) values stay out the acceptable ranges of .90-1.00 (Bentler, 1990; Tucker & Lewis, 1973). Likewise, the RMSEA indice of .05 indicates a satisfactory value by remaining between .05-.08 interval. As a result, it can be stated that the goodness of fit indices emerged meet the model fit requirements for the four-factor IERS.

At the next step, both standardized and unstandardized estimates for the 20 items of four-construct IERS were examined. Results of these estimates along with the standardized errors, t values and the variance explained are exhibited in Table 4.

Given the parameters presented in Table 4, it can be stated that the standardized factor loadings change between .57 and .88 for the individual items of the scale and all values are greater than .30 that is the minimum value for the factor loading (Brown, 2006). The explained variance comes from the items has the range of .33 to .77 that are all statistically significant ($p< .001$).

**Table 4.** Unstandardized and standardized parameter estimates for IERS

| Construct | Item | Unstandardized Factor Loadings | Standardized Factor Loadings | SE | T | $R^2$ |
|---|---|---|---|---|---|---|
| Enhancing Positive Affect | Item 3 | .68 | .77 | .04 | 15.73 | .55 |
| | Item 6 | .77 | .85 | .04 | 18.31 | .77 |
| | Item 8 | .75 | .84 | .04 | 17.86 | .67 |
| | Item 13 | .70 | .65 | .06 | 12.58 | .65 |
| | Item 18 | .69 | .75 | .05 | 15.26 | .37 |
| Perspective Taking | Item 2 | .69 | .67 | .05 | 12.88 | .66 |
| | Item 7 | .73 | .57 | .07 | 10.61 | .44 |
| | Item 10 | .71 | .60 | .06 | 11.09 | .77 |
| | Item 14 | .72 | .59 | .07 | 10.92 | .65 |
| | Item 17 | .77 | .69 | .06 | 13.34 | .62 |
| Soothing | Item 4 | .95 | .79 | .06 | 16.53 | .48 |
| | Item 9 | .95 | .81 | .06 | 17.18 | .35 |
| | Item 12 | 1.03 | .88 | .05 | 19.63 | .35 |
| | Item 16 | .81 | .67 | .06 | 13.13 | .33 |
| | Item 19 | .99 | .81 | .06 | 17.35 | .45 |
| Social Modeling | Item 1 | .60 | .61 | .05 | 11.74 | .56 |
| | Item 5 | .88 | .81 | .05 | 17.25 | .42 |
| | Item 11 | .93 | .82 | .05 | 17.69 | .70 |
| | Item 15 | 1.03 | .88 | .05 | 19.57 | .72 |
| | Item 20 | .90 | .74 | .06 | 15.27 | .59 |

Note: All t values were significant, $p < .001$.

### *3.3.2. Convergent validity*

In order to get more evidence over the construct validity of Interpersonal Emotion Regulation Scale, Interpersonal Competency Scale **(**Buhrmester, Furman, Wittenberg and Reis, 1988) was used as the convergent validity assessment. Results of the correlation analysis showed that IERS and ICS are significantly and positively correlated with each other ($r=.39$, $p< .001$).

### *3.3.3. Internal consistency*

The internal consistency indicator Cronbach Alpha was calculated .92 for the overall scale yielding satisfactory evidence for the reliability of IERS. In addition, Cronbach alpha was found .86 for enhancing positive affect, .80 for taking others' perspectives, .88 for soothing and .87 for social modeling sub-dimensions.

## 4. DISCUSSION and CONCLUSION

In this study, the Interpersonal Emotion Regulation Scale developed by Hofmann, Carpenter and Curtis (2016) was adapted into Turkish. In order to ensure the validity of the scale, a confirmatory factor analysis was then conducted. The results of confirmatory factor analyses showed that the emerged model fit indices support the four -factor structure of the scale. In other words, the unique construct of the scale was validated by the sample of Turkish university students. The results of a reliability analysis conducted using the internal consistency coefficient method proved the scale's reliability. The scale is composed of four sub-dimensions. The first of these relates to enhancing positive emotion, with the second relating to perspective taking, the third soothing, and the last social modeling.

Interpersonal emotion regulation stresses the importance of social relations on the experiencing and expressing of emotions (Hoffmann, 2014). Skill in interpersonal emotion regulation enables one to handle stressful situations in an effective way and increases wellbeing. University is an

important stage in a young person's life and is a period in which one encounters various hardships, both academic and personal. These problems can lead to the experiencing of emotional ups and downs. The ability to temper such fluctuations may enable students to enjoy a more relaxed and secure academic and personal life.

With this in mind, in the first year especially, it seems that aside from orientation projects, it is very important for universities to develop and promote psycho-educational programmes geared towards encouraging students to develop skills in emotional regulation. For example, Kuzucu (2006) states that psycho-educational programme which he developed to increase awareness and expression of emotions have been effective in increasing skills in the expression of emotions. Similarly, those who express emotions are proven to have a high level of psychological and subjective well-being. Thus, it can be said that activities and workshops organized at counseling centers in universities can hold a plethora of real benefits for students.

Upon an examination of the current global literature, it seems that interpersonal emotion regulation has been studied in the context of social support and depression (Marroquin, 2011); performance of athletes (Tamminen & Crocker, 2013); close relations (Debrot, Schoebi, Perrez and Horn, 2013) affective disorder and anxiety disorder (Hofmann, 2014). The scale, adapted for use in the Turkish cultural context, makes it possible to examine a variety of concepts in regard to personal emotion regulation and identify the factors which contribute to its development.

In conclusion, it is safe to say that Interpersonal Emotion Regulation Scale-Turkish Form is a valid and reliable scale in measuring the interpersonal emotion regulation of university students. It must be stated, however, that there are a number of limitations to this study. Primarily, the study group was made up exclusively of university students. The examination of sample groups made up of participants from various age-groups and professions would be beneficial in terms of psychometrics. Another limitation of the study was that test re-test reliability was not calculated. A suggestion for further studies would be to consider the stability and reliability of the instrument as it would yield more generalisable results. Furthermore, sample groups from different universities and students of various faculties would contribute to more in-depth results.

## ORCID

Asude Malkoç https://orcid.org/0000-0002-9073-2752

## 5. REFERENCES

Bentler, M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, 107, 238-246.

Berking, M., & Wupperman, P. (2012). Emotion regulation and mental health: recent findings, current challenges, and future directions. *Current Opinions in Psychiatry*, *25*(2), 128-134.

Berscheid, E. (2003). *The human's greatest strength: Other humans*. In L. G. Aspinwall & U. M. Staudinger (Eds.), A psychology of human strengths: Fundamental questions and future directions for a positive psychology (pp. 37-47). Washington, DC, US: American Psychological Association.

Brown, A. T. (2006). *Confirmatory factor analysis for applied research* (1st ed). New York: Guilford Press.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: basic concepts, applications, and programming*. NJ: Lawrence Erlbaum Associates.

Cicchetti, D., Ackerman, B. P., & Izard, C. E. (1995). Emotions and emotion regulation in developmental psychopathology. *Development and Psychopathology, 7*(1), 1-10.

Coan J. A., Schaefer, H.S., & Davidson, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychological Science*, *17*, 1032–1039.

Cote, S. (2005). A social interaction model of the effects of emotion regulation on work strain. *Academy of Management Review, 30(*3), 509-530.

Debrot, A., Schoebi, D., Perrez, M., & Horn, A.B. (2013). Touch as an interpersonal emotion regulation process in couples' daily lives: The mediating role of psychological intimacy *Personality and Social Psychology Bulletin, 39*(10), 1373–1385.

Diamond, L. M., & Aspinwall, L.G. (2003). Emotion regulation across the life span: an ıntegrative perspective emphasizing self-regulation, positive affect, and dyadic processes. *Motivation and Emotion, 27*(2), 125–156.

Eisenberg, N., Fabes, R. A., Guthrie, I. K., & Reiser, M. (2000). Dispositional emotionality and regulation: Their role in predicting quality of social functioning. *Journal of Personality and Social Psychology, 78*, 136-157.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: McGraw-Hill.

Gaensbauer, T. J. (1982). Regulation of emotional expression in infants from two contrasting caretaking environments. *Journal of the American Academy of Child Psychiatry, 21* (2), 163–171.

Garnefski, N., Kraaij, V., & Spinhoven, P. (2001). Negative life events, cognitive emotion regulation and emotional problems. *Personality and Individual Differences, 30*(8), 1311-1327.

Garnefski, N. & Kraaij, V. (2006). Relationships between cognitive emotion regulation strategies and depressive symptoms: A comparative study of five specific samples. *Personality and Individual Differences*, 40(8), 1659-1669.

Golden, P. R., & Gross, J. J. (2010). Effects of mindfulness-based stress reduction (MBSR) on emotion regulation in social anxiety disorder. *Emotion*, *10*(1), 83-91.

Greenberg, J. (2011). *Behavior in organizations* (10 th ed.). Upper Saddle River, NJ: Prentice Hall. Hare, A. P. Handbook of small group research. New York, NY.

Gross, J. J. & Munoz, R.F. (1995). Emotion Regulation and Mental Health. *Clinical Psychology: Science & Practice, 2(*2), 151-164.

Gross, J. J. (1998a). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology*, *74*, 224–237.

Gross, J. J. (1998b). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, *2*, 271–299.

Gross, J. J. (1999a). *Emotion and emotion regulation*. In L.A. Pervin & O.P. John (Eds.), Handbook of personality: Theory and research (2nd ed., pp. 525–552). New York: Guilford.

Gross, J. J. (1999b). Emotion regulation: Past, present, future. *Cognition & Emotion*, *13*, 551–573.

Gross, J. J. (2001). Emotion Regulation in Adulthood: Timing Is Everything. *Current Directions in Psychological Science, 10*(6), 214–219.

Gross J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39, 281–291.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2), 348-362.

Higgins, E. T., & Pittman, T. S. (2008). Motives of the human animal: Comprehending, managing, and sharing inner states. *Annual Review of Psychology*, 59, 361–385.

Hofmann, S. G. (2014). Interpersonal emotion regulation model of mood and anxiety disorders. *Cognitive Therapy and Research*, *38*, 483–492.

Hofmann, S.G., Carpenter, K., & Kurtiss, J. (2016). Interpersonal Emotion Regulation Questionnaire (IERQ): Scale development and psychometric characteristics. *Cognitive Therapy and Research, 40* ( 3), 341 – 356.

Kashdana, T. B., Barrios, V., Forsyth, J. P., & Steger, M. F. (2006). Experiential avoidance as a generalized psychological vulnerability: Comparisons with coping and emotion regulation strategies. *Behaviour Research and Therapy, 44*(9), 1301-20.

Kline, B. R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kuzucu, Y. (2006). *Duyguları fark etmeye ve ifade etmeye yönelik bir psiko-eğitim programının, üniversite öğrencilerinin duygusal farkındalık düzeylerine, duyguları ifade etme eğilimlerine, psikolojik ve öznel iyi oluşlarına etkisi*. (Doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Lopes, P. N., Salovey, P., Côté, S., Beers, M., & Petty, R. E. (Ed.). (2005). Emotion regulation abilities and the quality of social interaction. *Emotion*, 5(1), 113-118.

Marroquín, B. (2011). 1Interpersonal emotion regulation as a mechanism of social support in depression. *Clinical Psychology Review, 31*(8), 1276-1290.

Marshall, N. M. (1996). Sampling for qualitative research. *Family Practice,* 13(6), 522-525.

Niven, K., Totterdell, P., Stride, C. B., & Holman, D. (2011). Emotion Regulation of Others and Self (EROS): The development and validation of a new individual difference measure. Current Psychology: *A Journal for Diverse Perspectives on Diverse Psychological Issues, 30*(1), 53-73.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain — A meta-analysis of imaging studies on the self. *NeuroImage, 31*, 440–457.

Rime (2007). The social sharing of emotion as an interface between individual and collective processes in the construction of emotional climates. *Journal of Social Issues, 63*(2), 307-322.

Rime (2009). Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review. *Emotion Review 1*(1), 60-85.

Tabachnick. B. G. & Fidell. L. S. (2006). *Using multivariate statistics* (5th ed.). Boston: Allyn andBacon.

Tamminen, K. A., & Crocker, P. R. E. (2013). "I control my own emotions for the sake of the team": Emotional self-regulation and interpersonal emotion regulation among female high-performance curlers. *Psychology of Sport and Exercise, 14*, 737–747.

Tamir, M., & Millgram, Y. (2017). *Motivated Emotion Regulation*. In Advances in Motivation Science, 207–247.

Thompson R. A. (1994). Emotion regulation: A theme in search of definition. *Monographs of the Society for Research in Child Development*, 59, 25–52.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrica*, 38, 1-10.

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion, 13*, 803–810.