IJATE

Volume 5 Issue 2 2018

International Journal of
Assessment Tools in Education

International Journal of Assessment Tools in Education

International Journal of
Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

IJATE

e-ISSN 2148-7456

http://dergipark.gov.tr/ijate
http://www.ijate.net/index.php/ijate/index

Volume 5 Issue 2 May, 2018

Dr. Izzet KARA
Editor in Chief
International Journal of Assessment Tools in Education (IJATE)

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 : 3 Issues per year (January, May, September)
Online ISSN: 2148-7456
Website : http://www.ijate.net/index.php/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 (IJATE) is a peer-reviewed online journal.
The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE), have be published three times a year (one volume per year, three issues per year - January, May and September). IJATE welcomes the submission of manuscripts that meets the general criteria of significance and scientific excellence.

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

• Türk Egitim İndeksi,

• 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)

ISSN: 2148-7456

ISSN: 2148-7456

Table of Contents

Research Article

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 201–222

DOI: 10.21449/ijate.345150

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

Development of a Scale to Evaluate Virtual Learning Environment Satisfaction

Nazire Burcin Hamutoglu [1], **Orhan Gemikonakli** [2]
Merve Savasci [1], **Gozde Sezen Gultekin** [1]

[1]Sakarya University, Hendek Kampüsü, Başpınar Mah., Muammer Sencel Cad., No: 23, 54300, Sakarya, Turkey
[2]Middlesex University, The Burroughs, London NW4 4BT, United Kingdom

Abstract: Recent advances in information and communication technologies (ICT) have resulted in improvements in the delivery of education. It is a well-known fact that learning technologies currently have a pivotal role in education. Amongst them, Virtual Learning Environments (VLEs) are widely used in education. The role of VLEs in improving quality and interaction in education as well as enabling better achievement through the use of a wealth of activities in teaching and learning is widely reported in the literature. However, there is a gap regarding the development of measurement instruments, especially in the Turkish context. Therefore, this study reports the development of a scale to evaluate students' satisfaction with respect to the use of VLEs in educational settings to address this gap. The dimensions of the scale are contribution (CONT), satisfaction (SAT), and communication (COM), and the scale is formed of 13 items. The sample consists of students enrolled in the Department of Computer Education and Instructional Technologies, studying on blended and face-to-face learning programs. First, the reliability of the instrument was calculated by Cronbach Alpha coefficient and test-retest reliability correlation coefficient. The Cronbach Alpha coefficients were found to be 0.87, 0.83, and 0.81 for CONT, SAT, and COM sub-dimensions respectively. The overall reliability of the scale was 0.92. EFA and CFA were conducted on the data collected from two different sample groups (206 and 186 students for EFA and CFA respectively) for the validity analyses of the scale. Results confirm that the scale is valid and reliable. While the t-test analysis shows no significant difference between gender groups, ANOVA revealed significant differences when year of study is considered.

ARTICLE HISTORY

Received: 19 October 2017
Revised: 24 January 2018
Accepted: 02 February 2018

KEYWORDS

Virtual learning environment,
Scale,
Reliability,
Validity

1. INTRODUCTION

The advances in ICT and the diffusion of the Internet have resulted in the transformation of both the construct and the functioning of educational environments virtually over the last two or three decades. Instructional technologies have witnessed a great change throughout the years, and borders of time and space are crossed by means of electronic learning systems (Raaij & Schepers, 2008), also known as virtual learning environments.

A VLE can be described as “a web-based communications platform, that allows students, without limitation of time and place, to access different learning tools, such as program information, course content, teacher assistance, discussion boards, document sharing systems,

CONTACT: Orhan Gemikonakli ✉ o.gemikonakli@mdx.ac.uk Middlesex University, The Burroughs, London NW4 4BT, United Kingdom

ISSN-e: 2148-7456 /© IJATE 2018

and learning resources" (Raaij & Schepers, 2008, p. 839). The emergence of VLEs gave new impetus to delivering subject content to learners and they are remarkably becoming part and parcel of teaching and learning process (Pituch & Lee, 2006; Raaij & Schepers, 2008).

Incorporation of VLEs into education has changed the way teaching and learning activities are implemented. Especially the interest of Higher Education Institutions (HEIs) in the deployment of VLEs has reached to new heights. Throughout the world, some HEIs currently offer certain forms of VLEs or Learning Management Systems (LMSs) such as Blackboard, Desire2-Learn, or open-source VLEs like Moodle (Rienties, Giesbers, Lygo-Baker, Ma, & Rees, 2016). The management of educational content, monitoring teaching and learning activities, empowering individuals' learning can now all be performed in an integrated environment, and the aim of VLEs is to facilitate e-learning and provide a systematic and well planned approach to teaching and learning activities (McGill & Hobbs, 2008). With VLEs, some of the twenty-first century problems of learning and teaching can also be addressed and solved.

1.1. Review of Literature

A review of the relevant literature shows that both empirical and theoretical research on VLEs focus on several issues such as perceived usefulness of VLEs (Sun et al., 2008; Lang, Dolmans, Muijtjens, & van der Vieuten, 2006; Yilmaz, Karaman, Karakus, & Goktas, 2014), students'attitudes (Liaw, 2008; Ogba, Saul, and Coates, 2012; Sumak, Hericko, Pusnik, & Polancic, 2011; Usta, Uysal, & Okur, 2016), perceptions of VLEs (Love & Fry, 2006), and success and motivation in blended learning environments (Unsal, 2012). The literature provides comprehensive information regarding VLEs' use in teaching and learning processes, and presents the reasons for incorporating them into education. There is abundant research reporting the role of VLEs in improving the quality and interaction in education (Hettiarachchi & Wickramasinghe, 2016). Moreover, a considerable number of studies demonstrate that learning performance is affected positively by VLEs (McGill & Hobbs, 2008; Stricker, Weibel, Wissmath, 2011) when compared to traditional instruction (Chou & Liu, 2005; Zhang, Zhao, Zhou, & Nunamaker, 2004). Empirical evidence from the literature also suggests that VLEs have numerous benefits such as their effect on independent learning (Barker & Gossman, 2013), motivation to learn (Barker & Gossman, 2013; Forteza, Oltra, & Coy, 2015), interaction and communication among learners (Hettiarachchi & Wickramasinghe, 2016; Vuopala, Hyvönen, & Järvelä, 2016), and on student satisfaction (Forteza, Oltra, & Coy, 2015).

Besides these studies, a growing body of literature on VLEs presents data with respect to potential gender differences regarding electronic learning, distance education and VLEs (e.g. Ching & Hsu, 2015; Cutmore, Hine, Maberly, Langford, & Hawgood, 2000; Goulão, 2013; Gunn, McSporran, Macleod, & French, 2003; Horvat, Dobrota, Krsmanovic, & Cudanov, 2015; Lowes, Lin, & Kinghorn, 2016; Perkowski, 2013; Yukselturk & Bulut, 2009). Gender based differences might have an effect on the way the learners perceive VLEs, or their achievement or motivation might be affected.

In addition to potential differences among different sexes, year of study is another factor that might affect use of VLEs. It is expected that students at higher grades are more mature and experienced. Moreover, awareness of information on the Internet and age are also considered as important factors affecting learners' performance in VLEs (Lee, Hong, & Ling, 2001). Therefore, when the fact that "the success of any virtual learning environment depends on the adequate skills and attitudes of learners" (Lee, Hong, & Ling, 2001, p. 231) is taken into consideration, it might be necessary to investigate the role of year of study. Moreover, as stated by Martins and Kellermanns (2004), "awareness of the capabilities of the system, …, and prior experience with computer and Web use are positively related to perceived ease of use of the system, which in turn is positively related to student acceptance of the system." (p. 7).

As it can be seen, the incorporation of VLEs has received considerable attention from the researchers, teachers, and practitioners in the field, due to the benefits attributed to them. Nevertheless, since it is not quite possible to handle all the dimensions of VLEs, in this paper, we chose three dimensions of VLEs, which are considered amongst the critical factors in the implementation of VLEs. Therefore, in this paper, the following dimensions will be embraced: content, student satisfaction, and communication.

1.1.1. Satisfaction

Successful online teaching-learning processes, that are successful implementation of VLEs, hinge on satisfaction or dissatisfaction of users to a large extent. In a VLE, the critical factors affecting users' satisfaction can be categorized into six dimensions, which are learner, instructor, course, technology, design, and environment (Sun, Tsai, Finger, Chen, & Yeh, 2008, p. 1184). From a different point of view, Chua and Montalbo (2014) put forward four factors for users' satisfaction such as learner interface, learning community, content, and usefulness. Additionally, Wang (2003) developed a model for measuring e-learner satisfaction on asynchronous electronic learner systems including a fifth factor: *personalization.* Asoodar, Vaezi, and Izanloo (2016) developed six dimensions such as learner, instructor, course, technology, design, and the environment to improve the satisfaction of learners.

Links have been reported in the literature between VLE use and satisfaction (De Lange, Suwardy, & Mavondo, 2003; McGill & Hobbs, 2008). There are also studies demonstrating that the use of VLEs contributes to students' satisfaction when compared to students receiving traditional instruction (Chou & Liu, 2005; Koskela, Kiltti, Vilpola, & Tervonen, 2005). Hew and Kadir (2016) state that the use of VLEs would enhance student approaches to learning and may promote students' achievement by feedback, extra support, cooperative revision, and so forth. However, it should be noted that successful deployment of VLEs in HEIs depends considerably on user acceptance (Raaij & Schepers, 2008) and their satisfaction. While satisfaction is considered to have a significant relationship with online events continuance (Cheng, Wang, Huang, & Zarifis, 2016), individuals' level of satisfaction of the use of VLEs impacts the future use of those technologies (Al-Khalifa, 2009; Bell & Farrier 2008; Cheng, 2011; Lin, 2012; Sumak et al. 2011; Zafra et al. 2011). It should also be noted that when VLEs are selected appropriately for content, they support learners by providing content, and independent learning, hence increasing learners' satisfaction.

Earlier studies focused on a range of issues regarding satisfaction. To exemplify, Naveh, Tubin and Pliskin (2010) investigate the relationship between students' satisfaction and achievements when LMSs are used in teaching and learning. Lee, Srinivasan, Trail, Lewis, and Lopez (2011) examine the relationship between satisfaction, outcome, and student perception of support, and Zhu (2012) similarly investigates differences of satisfaction in different cultures. Ku, Tseng, and Akarasriworn (2013) state the importance of interaction on satisfaction. Shubina (2016) compares users' satisfaction on three different Massive Open Online Course (MOOC) platforms. There are also some other studies using instruments based on satisfaction with process and satisfaction with outcome variables (Briggs, Reinig, & de Vreede, 2008, 2014; Cheng et al., 2016; Reinig, Briggs, & de Vreede, 2009). Furthermore, the self-evaluation of students' satisfaction regarding the use of VLEs (e.g. Cassidy, 2016) is investigated in some studies.

All in all, students' satisfaction is considered as a critical element in learning environments in terms of effectiveness of the learning processes, especially of virtual learning environments.

1.1.2. Communication

In addition to their contribution to learner/user satisfaction, VLEs also promote effective communication among students (Barker & Gossman, 2013) as well as between students and teachers (Martins & Kelllermanns, 2004; Raaij & Schepers, 2008). Since borders of time and space are crossed by means of VLEs (Raaij & Schepers, 2008), the opportunities for communication are enhanced. That is, it can be stated that with VLEs, "the potential to improve communication and mutual support between students" (Leese, 2009, p. 70) is enhanced.

Numerous studies in the literature demonstrate that virtual learning environments enrich interaction and therefore communication that students have with one another in addition to the interaction between students and their instructors (Hettiarachchi & Wickramasinghe, 2016). That is, VLEs are considered to facilitate communication (Barker & Gossman, 2013).

1.1.3. Contribution

The contribution of VLEs is manifold. Several previous studies have presented results pertaining to the contribution of VLEs to the quality in education (Hettiarachchi & Wickramasinghe, 2016), students' motivation (Beluce & Oliveria, 2015; Forteza, Oltra, & Coy, 2015) and satisfaction (Forteza, Oltra, & Coy, 2015), learning performance (McGill & Hobbs, 2008; Stricker, Weibel, Wissmath, 2011), interaction and/or communication among students, and between students and teachers (Barker & Gossman, 2013; Hettiarachchi & Wickramasinghe, 2016; Leese, 2009; Martins & Kelllermanns, 2004; Raaij & Schepers, 2008), and so forth.

1.2. The Aim of the Study

In order to establish the impact of VLEs on student satisfaction of teaching and learning in higher education, this study aims to develop a valid and reliable instrument to measure the impact of VLEs on learning, focusing on satisfaction. An effective way of understanding the effectiveness of VLEs on students' learning is through the evaluation of feedback collected from students. The collection of student feedback can best be made through a scale developed in their mother tongue and subjected to reliability and validity tests prior to its use. Furthermore, Vaz, de Bittencourt, Vaz, and Júnior (2015) contend the importance of student feedback in further improving VLEs through enhancing and developing new solutions and strategies. It is believed that measuring satisfaction of the use of VLEs would enable administrators and developers to identify the strengths and weaknesses of the systems concerned, and use these findings to further improve these systems to meet students' needs and expectations. This is further emphasized in other studies in the literature (e.g. Eom, Wen, & Ashill, 2006; Kember & Ginns, 2012; Zerihun, Beishuizen, & Os, 2012). Finally, the report of Universities and Colleges Information System Association (2016) indicates the importance of technology enhanced learning and highlights the challenges faced by participating HEIs.

All in all, for successful deployment of VLEs, it is essential that effective instruments are developed to evaluate user satisfaction. This paper presents such a valid and reliable instrument that was developed.

1.3. The Significance of the Study

An in depth review of the literature points at several scales such as satisfaction scale toward online courses (Kolburan-Gecer & Deveci-Topal, 2015), preparedness and expectancy scale for e-learning process (Gulbahar, 2012), satisfaction scale for learning management systems (Naveh, Tubin, & Pliskin, 2010), and perception of satisfaction toward learning management systems (Horvat, Dobrota, Krsmanovic, & Cudanov, 2015).

When studies conducted in Turkish context are carefully researched, and to the best of our knowledge, there is no measurement tool to determine the satisfaction level of students in the use of VLEs. Moreover, "to measure how students and teachers are going to accept and use a specific e-learning technology or service, an appropriate instrument is needed" (Sumak, Polancic, & Hericko, 2010). This study was thereby motivated by the gap in the literature and is considered significant for determining the satisfaction of students towards VLEs in the Turkish culture of education. Thus, learners' views towards existing systems can support the learning-teaching processes by helping institutions to improve themselves, as well as to see their strengths and weaknesses.

2. METHOD

The purpose of this study is to develop a scale on VLEs to evaluate the satisfaction of the users through gathering the opinions of Sakarya University students regarding the learning platform that they use.

2.1. Sample

The sample of this study is formed of university students (*N*= 433) studying at Sakarya University, Faculty of Education, Department of Computer and Instructional Technology Education (CITE) and Science Teaching Departments, during the 2013-2014 academic year. The participants are drawn from four different groups: The first group used for analyzing EFA consists of 206 students (f=158, 76.7%; m= 48, 23.3%) studying at CITE face-to-face learning program. The second group, from which confirmatory factor analysis results are obtained, consists of 186 students (f=77, 41.4%; m= 109, 58.6%) studying at CITE on a blended learning program. The third group consists of 10 students (f=5, 50%; m=5, 50%) studying at CITE, both on face-to-face and blended learning programs, used for pilot study. Finally, the fourth group consists of 31 students, of whom 11 (34%) are female and 20 (66%) were male, studying at the Department of Science Teaching, used for test-retest analysis in terms of internal consistency. The demographics of the participants are shown in Table 1:

Table 1. Demographics of participants

Participants	Variable		N	M	SD	Skewness	Kurtosis
Participants employed for CFA	Gender	F	77	49.75	7.23	.025	-.049
		M	109	49.69	7.09		
	Year	2	72	48.06	6.28		
		3	66	48.92	7.11		
		4	48	53.27	7.23		
Participants employed for EFA	Gender	F	158	55.09	9.54	-.426	.750
		M	48	56.23	10.28		
	Year	2	88	55.41	10.03		
		3	53	54.55	10.49		
		4	65	55.94	8.64		

F: Female, M: Male

The reason behind employing students enrolled in Sakarya University was the fact that there are two types of programs in CITE Department, which involves face-to-face and blended learning environments. In both programs, Sakarya Universitesi Bilgi Sistemi (SABIS), an institution wide VLE - a course management system - from which students access lecture notes, follow course procedures, etc. is used. While students enrolled in blended learning programs use the system more actively, students studying on face-to-face programs use the system mostly for checking their grades. In blended learning programs, since only 30 percent of courses are delivered face-to-face, 70 percent of instruction is delivered via a virtual learning system. That is, a face-to-face environment complements the virtual learning environment. Learning-teaching materials are sent to the students asynchronously (e.g. as a document, a video, a PowerPoint presentation, etc.) via the system. Besides, students in blended learning programs can also take an exam on the system.

2.2. Procedure

The study was conducted in two phases: the development of the scale, and administering and analyzing the results obtained from the scale.

2.2.1. The development of the scale

First of all, in the process of the development of a scale for evaluating the satisfaction of the students on VLEs, a theoretical basis was created by reviewing the literature. Following this step, a pool consisting of 20 items was created based on this theoretical basis. Expert opinions involving 3 field experts - one assessment and evaluation expert, one language expert and one Psychological Counselling and Guidance expert- were then elicited regarding the item pool. Following the expert opinions, some revisions were made and 2 items were omitted from the scale in light of the expert opinions, and the scale was administered for the pilot study.

The instrument was constructed and validated with the participation of pre-service teachers from Sakarya University. For the pilot study, a group of 10 people was employed in order to analyze the comprehensibility of the items. The participants were invited for a focus-group interview and the items which were not clear or comprehensible for the participants were revised. Following this step, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were conducted to establish the construct validity of the scale. A total of 3 and 2 items were omitted as a result of EFA and CFA respectively. As a result, a total of 7 items were excluded from the scale, leaving 13 items after conducting the pilot study and establishing the validity. The reliability level of the scale was examined by Cronbach alpha internal consistency and test-retest methods.

2.3. Data Collection Instrument

2.3.1. The VLE Scale

Developed within the scope of the research purpose, the VLE scale has a three-factor structure consisting of three dimensions – satisfaction (SAT), contribution (CONT), and communication (COM) - and comprises 13 items which were finalized following the validation study undertaken with the participation of pre-service teachers (see Appendix A). The scale is a 5-point Likert scale in which the options range from 1 to 5 (1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, and 5 = Strongly agree). The scores taken from the scale vary from 13 to 65 at this interval. There are no reverse-scored items in the scale and students’ satisfaction increases with higher scores received.

2.4. Data Analysis

Statistical analyses were conducted with SPSS 20 (Statistical Package for Social Sciences) and LISREL 8.7 (Linear Structural Relations) software programs.

3. FINDINGS

3.1. Validity of the Scale

To establish the validity of the scale, face, content, construct, convergent, and discriminant validity were explored.

3.1.1. Face and Content Validity

First of all, face and content validity were explored through expert opinions. Three field specialists from the field of Computer and Instructional Technology Education, one specialist from the field of Measurement and Evaluation, and one Turkish language expert were consulted for appearance and coverage.

3.1.2. Construct Validity

To investigate the construct validity of the scale, EFA and CFA were conducted. Furthermore, convergent and divergent validity were established.

3.1.2.1. Exploratory Factor Analysis

The construct validity of the scale was evaluated via EFA. Before performing an exploratory factor analysis, it is necessary to determine whether the data set is suitable for factor analysis. The process for this is to perform Kaiser-Meyer-Oklin (KMO) (Kaiser, 1974) and Barlett Sphericity (Bartlett, 1954) tests. Therefore, before conducting EFA, KMO measure of sampling adequacy and Barlett Sphericity tests were conducted. The KMO ranges from 0 to 1, and the KMO values above 0.5 are acceptable (Field, 2009). However, it is accepted that the KMO "values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are superior" (Field, 2009, p. 679, as cited in Loewen & Gonulal, 2015). The results exhibited a KMO measure of sampling adequacy of 0.92 (KMO= 0.92), a value greater than 0.70, indicating that the sample size was adequate for factor analysis (Bryman & Cramer, 1999). Bartlett test of sphericity was 1805.933 (p<.001, SD=105), indicating that a factor analysis was appropriate (Bryman & Cramer, 1999). According to these results, it can be stated that the data were fit for the factor analysis. As the scree plot of Figure 1 depicts, eigenvalue-greater-than-one showed itself as a good choice in determining the optimal number of factors to retain EFA, which in case of this study is 15, and with the basic components analysis prioritized, the varimax (25) rotation was performed. The results of validity analysis demonstrated that the VLE scale had a 3 factor structure.

When the items to be included in the instrument were determined as a result of the EFA for the construct validity of the scale, it was noted that the eigenvalues of the factors constituting the scale items were 1 and above, and the factor loadings were 0.30 and above. In addition, it was also noted that the materials are included in a single factor or that there is at least a 0.10 difference between the factor loadings of the items (Buyukozturk, 2012).

Then, 3 items that were not suitable for these criteria were omitted from the scale. In addition, a rotation was performed on the factors. The results obtained from EFA indicate that there is a three-dimensional structure of the scale. These dimensions are called "Satisfaction" (SAT), "Contribution" (CONT), and "Communication" (COM). The self-scattering diagram regarding the three-dimensional structure larger than the eigenvalue of 1 is presented in Figure 1 below whereas factor loadings and variance rates explained by the scale are presented in Table 2.

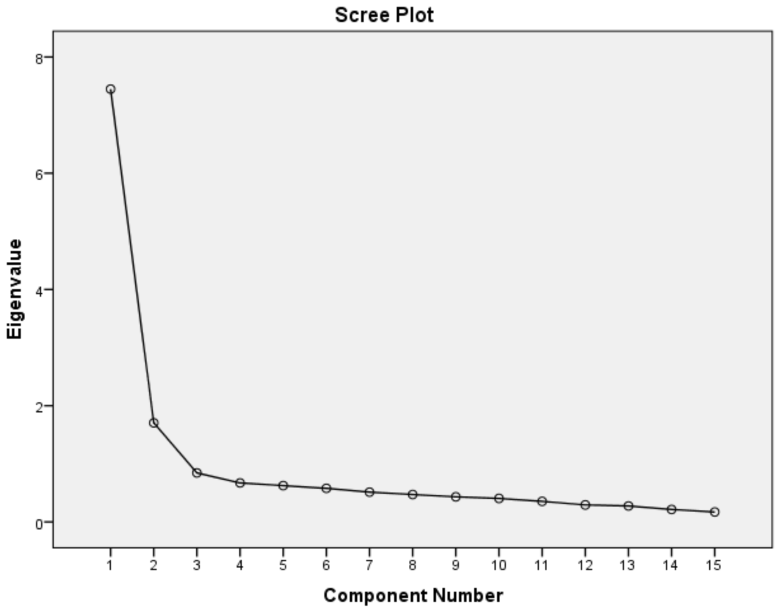


Figure 1. The graph of the eigenvalue-component number of the scale

There are a total of six items in the first factor- contribution. One of these items “I would like to use VLEs in my other courses as well” is the sample item of this factor. The factor loadings of these items on this factor vary between 0.56-0.74. This factor which explains 21.98% of the total variance of the scale is categorized as “CONT”. The second factor - satisfaction - in the scale consists of a total of five items. One of these items “I am content with the VLE used in the course” is the sample item of this factor. The factor loadings of these items on the second factor vary between 0.37 - 0.75. This factor which explains 23.59% of the total variance of the scale is named as “SAT”. The third factor – communication – in the scale consists of a total of four items. One of these items “I would recommend the use of forums for other courses as well” is a sample item of this factor. The factor loadings of these items on the third factor vary between 0.61 - 0.84. This factor which explains 21.06% of the total variance of the scale is named as “COM”.

Overall, the scale indicates a three-factor structure. The factor loadings of the 15 items in the scale on the factors vary between 0.37-0.84. Three factors in the scale explain 66.64% of the total variance. After EFA, the scale overall consists of 15 items and three factors. These values indicate that the scale explains participants’ opinion of the learning platform well. According to EFA results, the CONT sub-scale consists of 6 items and explains 21.98% of the total variance. The factor loadings of the items in the CONT sub-scale range from 0.560 to 0.704. The SAT subscale consists of 5 items and accounts for 23.59% of the total variance. The factor loadings of the items in the SAT subscale range from 0.367 to 0.747. The COM sub-scale consists of 4 items and accounts for 21.06% of the total variance. The factor loadings of the two items in the COM sub-scale are 0.605 and 0.840. The findings show that not only the scale can be used as it is but also the three-factor structure of the scale can be evaluated as three separate scales.

Table 2. The factor structure and factor loads of the scale

Factor	Item No	Items	Common Factor Variance	Factor load
CONT	2	Diğer derslerde de VLE kullanmak isterim. *[I would like to use VLEs in my other courses as well.]*	.609	.677
	3	Öğretim materyallerinin diğer derslerde VLE üzerinden sunulmasını isterim. *[I would like the materials of the other courses to be presented via VLE]*	.675	.614
	4	Öğrenme & öğretmen materyallerinin VLE üzerinden sunulması ders sürecine katkı sağlar. *[Presenting the learning and instruction materials via VLE contributes to the course process]*	.704	.628
	5	Diğer derslerde duyuruların VLE üzerinden yapılmasını öneririm. *[I would recommend the announcements in other courses to be made via VLE]*	.635	.700
	7	VLE üzerinden gönderilen mesaj yayınları öğrenme &öğretme sürecine katkı sağlar. *[Messages that are sent via VLE contribute to the learning and teaching process.]*	.638	.405
	12	Bana gore bütün derslerin VLE üzerinden sunulması gerekir. *[To me, all the courses should be offered via VLE]*	.560	.654
	Explained variance %			**21.98**
SAT	6	Derste kullanılan VLE'den memnunum. *[I am content with the VLE used in the course]*	.701	.747
	8	VLE üzerinden ders kapsamında sunulan öğrenme & öğretmen materyallerinden memnunum. *[I am content with the learning & teaching materials presented within the course via VLE]*	.683	.743
	9	VLE üzerinden yayınlanan mesaj ve duyurulardan memnunum. *[I am content with the messages and announcements that are broadcasted via VLE]*	.636	.709
	10	Dersin VLE üzerinden sunumundan memnunum. *[I am content with the presentation of the course via VLE]*	.614	.581
	18	Derste VLE üzerinde kullanılan anketlerden memnunum. *[I am content with the questionnaires employed on VLE in the course]*	.767	.367
	Explained variance %			**23.59**
COM	11	Diğer dersler için de VLE üzerinden forum kullanılmasını öneririm. *[I would recommend the use of forums via VLE for other courses as well]*	.605	.382
	14	Diğer dersler için de VLE üzerinden anket kullanılmasını öneririm. *[I would recommend the use of questionnaires via VLE for other courses as well]*	.653	.750
	16	VLE üzerinden daha fazla forum kullanılmasını isterdim. *[I would like to use more forums via VLE]*	.677	.732
	17	VLE üzerinden daha fazla anket kullanılmasını isterdim. *[I would like to use more questionnaires via VLE]*	.840	.898
	Explained variance %			**21.06**
	Total explained variance %			**66.64**

3.1.2.2. Confirmatory Factor Analysis (CFA)

Followed by the EFA, a confirmatory factor analysis (CFA) was conducted to verify and determine the factor structure of the scale, and the following fit indices were selected: 1) Chi-Square goodness of fit test, 2) Goodness of Fit Index (GFI), 3) Adjusted Goodness of Fit Index (AGFI) 4) Comparative Fit Index (CFI), 5) Normed Fit Index (NFI), 6) the Root Mean Square Error of Approximation (RMSEA) and 7) Standardized Root Mean Square Residual (SRMR).

In general, for indices GFI, CFI, and NFI 0.90 and 0.95 onwards represent acceptable and superior fit respectively (Bentler, 1980; Bentler & Bonett, 1980; Marsh, Hau, Artelt, Baumert & Peschar, 2006). For AGFI, a value of 0.85 indicates acceptable and a value of 0.90 indicates superior fit (Schermelleh-Engel & Moosbrugger, 2003). For RMSEA, 0.08 indicates acceptable fit and 0.05 indicates superior fit (Brown & Cudeck, 1993; Byrne & Campbell, 1999). For SRMR, the 0.05 value is considered as superior fit and the 0.10 value as acceptable fit (Schermelleh-Engel & Moosbrugger, 2003).

The structure reduced to 15 items following EFA and formed of three factors was then tested by CFA. CFA analysis was performed as first and second-level CFA (BD-CFA and ID-CFA). Factor loads for the three-dimensional model obtained from the first-order CFA are shown in Figure 2.

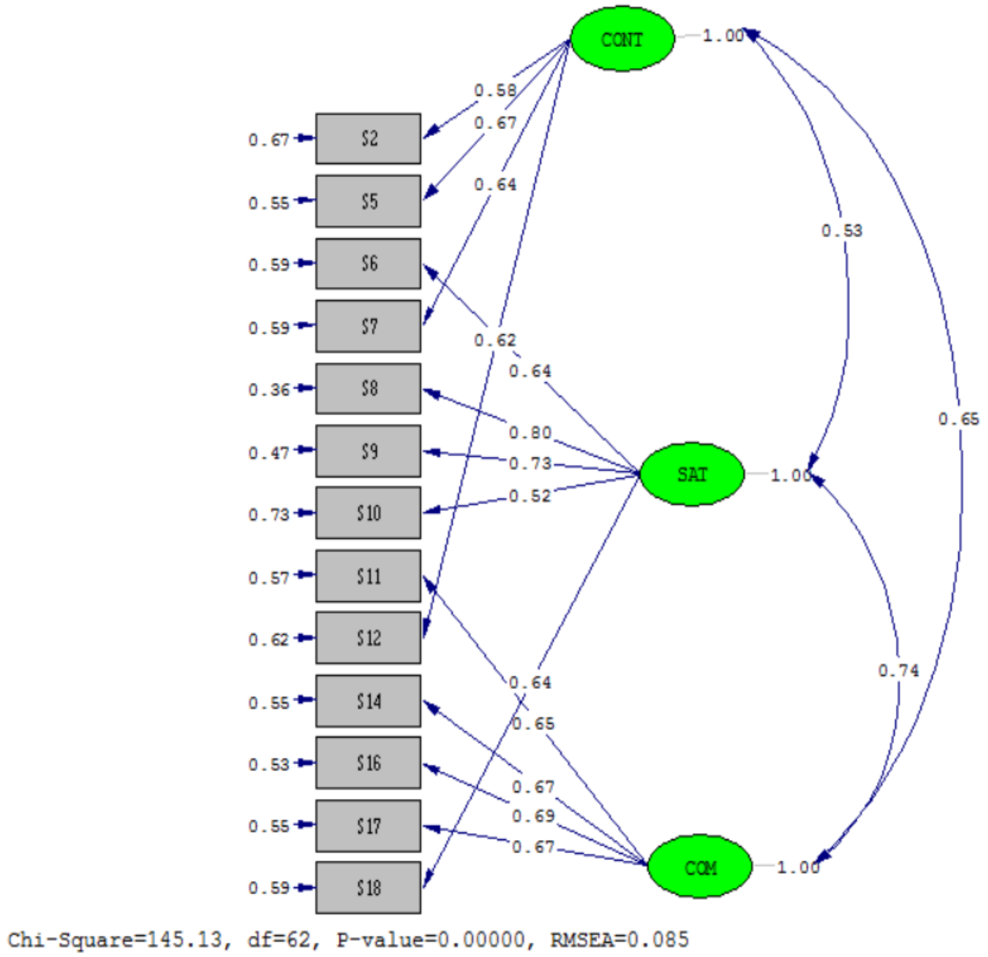


Figure 2. Path diagram and factor loadings obtained from first level CFA regarding the scale

As seen in Figure 2, the factor loadings for the CONT sub-dimension range from 0.58 to 0.67, from 0.52 to 0.80 for the SAT sub-dimension, and from 0.64 to 0.69 for the COM sub-dimension. The fit indices of the three-factor model consisting of 15 items and three sub-dimensions were examined at the first level. The standard solutions and t-values of 2 items serving for the CONT dimension were excluded on the grounds that they were not meaningful for the factor. In the first-level CFA, the items of the CONT factor were 0.58, 0.67, 0.64, and 0.62; The SAT factor was 0.64, 0.80, 0.73, 0.52, and 0.64; and the COM factor had a standard solution of 0.65, 0.67, 0.69 and 0.67, respectively. Since all the factors have a value higher than 0.45, thirteen items were important factors in terms of three factors. In addition, t values of thirteen items and three-factor structure are examined.

The t values of the items of CONT factor were 7.45, 8.84, 8.33 and 8.00; SAT factor were 9.05, 12.09, 10.69 and 7.05, and COM factor were 9.05, 9.39, 9.66 and 9.37, respectively. The calculated t values are greater than 1.96 and at 0.05 level (Jöreskog & Sörbom, 1993; Kline, 2011; Cokluk, Sekercioglu & Buyukozturk, 2012, p. 304) which are significant at the .01 level, and the number of people in the research group is at a sufficient level for factor analysis. When the correction proposal for 13 items was examined as a result of the CFA, items 3, 4, 8, and 9 were corrected. The reason for this correction can be explained as follows: If a change suggested by the correction indices corresponds to a significant decrease in the value of χ^2 of the model, and if this is the declining trend, it can be evaluated that the proposed correction is a critical change in terms of the model (Cokluk, Sekercioglu and Buyukozturk, 2012, p. 312). In addition, if more than one correction is required, these corrections must be made one at a time. The fit index of the model obtained in CFA was examined and it was found that the minimum chi-square value (χ^2 = 145.13, N = 62, p = 0.00) was significant. The fit index values were RMSEA = 0.085, GFI = 0.89, AGFI = 0.84, CFI = 0.94, NFI = 0.91, and SRMR = 0.06. The superior and acceptable fit measures for the fit indices examined show that the three-factor model from the CFA is consistent and that the factor structure identified in the EFA is validated.

In addition to the first-level CFA, second-level CFA was applied to determine the extent to which the CONT, SAT, and COM subscales fit into the scale's implicit variable, which is defined as a superstructure. The analysis produced the same results as the first-level CFA, hence, it can be concluded that in terms of the model-data fit, the two models are identical, and the scale can be measured by a three-factor structure called CONT, SAT, and COM. The factor loadings for the three-dimensional model obtained from the second-level CFA are shown in Figure 3.

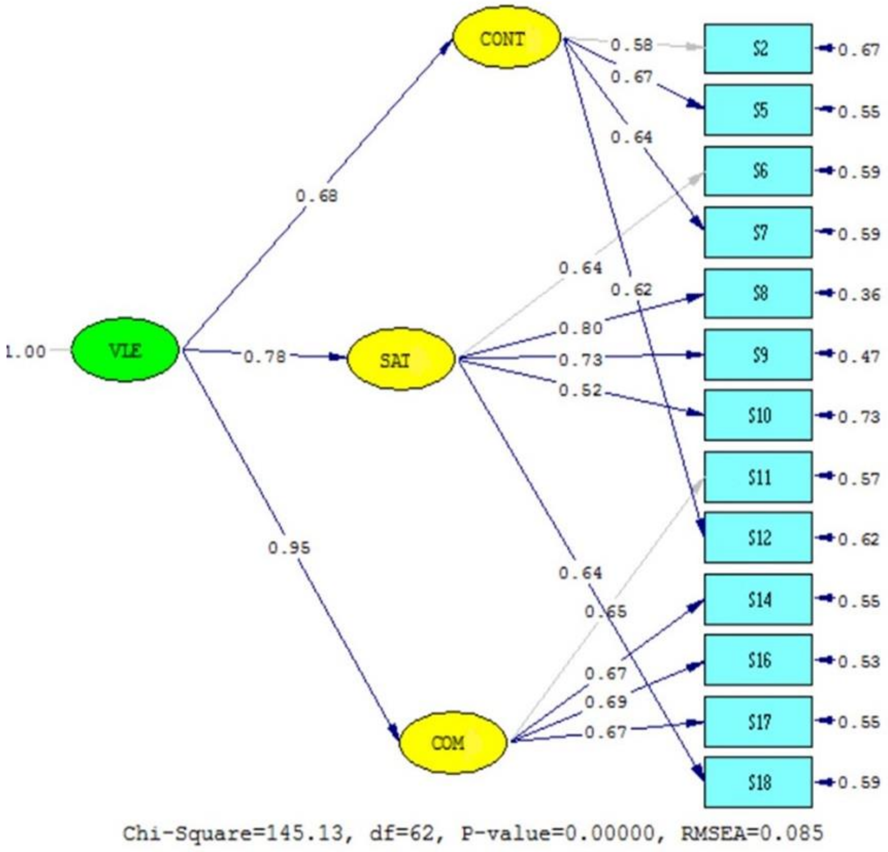


Figure 3. Path diagram and factor loads obtained from second-level CFA regarding VLES

As it can be seen in Figure 4, the factor loads for CONT, SAT and COM, defined as sub-dimensions of the scale's implicit variable, are 0.68, 0.78, and 0.95, respectively. Having a value higher than 0.45 for each and every factor of the scale, it is fair to state that these are important factors for the scale. In addition, the results of t values obtained as a result of examining the three-factor structure of the second-level CFA and the scale demonstrated that all t values were greater than 2.58, so it was statistically significant (p<.01). Therefore, it can be said that the CONT, SAT, and COM subscales are significant predictors of the scale's implicit variable.

In the final step, the R^2 findings were examined. Given the variances in R^2 explained above, the values for the items of the CONT factor were 0.33, 0.45, 0.41, and 0.38; the values for the items of the SAT factor were 0.41, 0.64, 0.53, 0.27, and 0.41; and the values for the items of the COM factors were 0.43, 0.44, 0.47, and 0.45 respectively. When the R^2 values of the factors in the latent variable are considered, they are 0.46, 0.61, and 0.91 respectively. The values of R^2 for the variance are above 20%, indicating that the fit indices are acceptable. The superior and acceptable fit measures for the fit indices examined in the study and the fit indices obtained from the first and second level CFA are presented in Table 3.

Table 3. Obtained fit index values

Fit Indices	Superior fit	Acceptable fit	Fit indices from first level CFA
χ^2 /sd	$0 \leq \chi^2/sd \leq 2$	$0.2 \leq \chi^2/sd \leq 3$	2.34
GFI	.95 ≤ GFI ≤ 1.00	.90 ≤ GFI ≤ .95	.89
AGFI	.90 ≤ AGFI ≤ 1.00	.85 ≤ AGFI ≤ .90	.84
CFI	.95 ≤ CFI ≤ 1.00	.90 ≤ CFI ≤ .95	.94
NFI	.95 ≤ NFI ≤ 1.00	.90 ≤ NFI ≤ .95	.91
RMSEA	.00 ≤ RMSEA ≤ .05	.05 ≤ RMSEA ≤ .08	.085
SRMR	.00 ≤ SRMR ≤ .05	.05 ≤ SRMR ≤ .10	.067

As it can be seen from the findings given in Table 3, the fit indices obtained from the first and second level CFA are very close to each other. Accordingly, it can be said that the compatibility of both models is identical. The fit indices obtained from the first and second level CFA; the construct validity is established. It was then thought that an RMSEA of between 0.08 and 0.10 provides a mediocre fit and below 0.08 shows a good fit (MacCallum et al., 1996). In addition to all these, the total score of the scale and the individual correlation coefficients of the three factors were examined (see Table 4).

Table 4. The factor correlation values of the scale

	CONT	SAT	COM	Total
CONT	-	.42**	.48**	.75**
SAT		-	.60**	.85**
COM			-	.85**
Total				-

**p<.01

The correlation scores between CONT, SAT, and COM factors were 0.75, 0.85, and 0.85, respectively, with a total score from the developed scale, and a significant correlation was found between these scores (p< 0.01). Correlation coefficients of CONT, SAT, and COM factors were 0.42, 0.48 and 0.60, and it was also found that there was a significant correlation between these values (p< 0.01. The findings related to the correlation coefficient indicate that the factors comprising the scale are compatible and related. When the item total correlations are examined, it is seen that the correlation values for all the items in the scale change between

0.49 and 0.68. These values are higher than 0.30, indicating that all items can distinguish individuals at a high level (Buyukozturk, 2012).

The fit indices are observed to be at acceptable levels when the fit indices of the scale are examined. The internal consistency factors (alpha) were calculated for the reliability studies of the scale.

3.1.2.3. Convergent and Discriminant Validity

Convergent and discriminant validity were investigated for the construct validity that measures the 3 factorial structure of the VLE satisfaction scale. With respect to convergent validity, AVE values were examined for each factor [CONT(F1), SAT(F2), COM(F3)], and they were 0.72; 0.73 and 0.76 respectively. Being higher than 0.50, all these values demonstrate convergent validity (Bagozzi & Youjae, 1988), showing evidence of the VLE scale's convergent validity. On the other hand, discriminant validity of the scale was measured by calculating whether the AVE square root of the scale were greater than both the correlation among the structures and the value 0.50 (Fornell & Larcker, 1981), and the results indicated that VLE scale has discriminant validity (see Table 5).

Table 5. The coefficients of discriminant validity

	F1	F2	F3
F1	**0.850**		
F2	0.336	**0.856**	
F3	0.664	0.637	**0.875**

3.2. Reliability of the Scale

The reliability of the scale was calculated by internal consistency (Cronbach α) and test retest methods, for both the first and second group of the study. The results are illustrated in Table 6, and these values for the sub-dimensions and the total score of the scale can be stated as high values for the internal consistency values and the reliability factors of the scale are quite good.

Table 6. Reliability coefficients of the scale calculated by internal consistency method

Sub-scales	EFA		CFA
	Cronbach Alpha	**Test-retest**	**Cronbach Alpha**
CONT	.87	.94	.71
SAT	.83	.87	.78
COM	.81	.95	.76
The scale overall	.92	.94	.86

In the study, the internal consistency coefficient obtained from the first group of 206 students was 0.92 for the scale. Internal consistency coefficients for subscales were 0.87 for the subscale of CONT, 0.83 for the subscale of SAT, and 0.81 for the subscale of COM. The internal consistency coefficient obtained from 186 students in the second group was 0.86 for the scale. In addition, internal consistency coefficients for subscales were calculated as 0.71 for the subscale of CONT, 0.78 for the subscale of SAT, and 0.76 for the subscale of COM. In order to calculate the test retest reliability of the scale, it was administered to 31 students who were enrolled in the Department of Computer and Instructional Technology Education twice with three weeks intervals and the correlations between the two applications were calculated. The reliability coefficients calculated by the test re-test method are 0.94 for the scale, 0.94 for the CONT subscale, 0.87 for the SAT subscale, and 0.95 for the COM subscale. Reliability coefficients of 0.70 and over are considered to be reliable (Buyukozturk, 2012; Pallant, 2005). According to this, it can be stated that the reliability coefficients of the scale and CONT, SAT, and COM subscales are appropriate.

3.3. Analysis of Scores from the Scale

The scale consists of 13 items. A 5-point Likert scale was used with responses ranging from Strongly agree (5), to Strongly disagree (1). There are no reverse-scored items in the scale. As there are 4 items in the CONT sub-dimension, the lowest score that can be taken from this dimension is 4 and the highest score is 20. There are 5 items in the SAT dimension. Therefore, the lowest score that can be taken from this dimension is 5 and the highest score is 25. Similarly, there are 4 items in the COM sub-dimension. For this reason, the lowest score that can be received from this dimension is 4 and the highest score is 20. The scale provides adequate fit indices in both first-level and second-level CFA; the scale can be used as a whole or just for the subscale. The higher the scores in subscales or overall scale indicate higher satisfaction from VLEs. Moreover, obtaining acceptable fit indices for both the first-level and second-level CFA means that it is possible to compute the scores obtained from the subscales of the scale as well as a total score on the scale.

4. CONCLUSION

As shown in Table 7, there are no significant differences between the participants' opinions on the CONT sub-dimension [F(2,183)=2,165, p>.05] in terms of participants' year of study. However, there are significant differences between the participants' opinions on SAT sub-dimension [F(2,183)=8,024, p<.05], COM sub-dimension [F(2,183)=8,457, p<.05] and overall Virtual Learning Environment Satisfaction [F(2,183)=9,008, p<.05] with respect to year of study. To investigate which groups differ from each other, a Scheffe test was performed for each of these dimensions. In the analysis, Scheffe test results revealed that there are significant differences in favor of 4th year students compared to 3rd and 2nd year students. In this case, it can be stated that 4th year students' levels of satisfaction, communication, and overall Virtual Learning Environment Satisfaction are higher than that of 3rd and 2nd year students'.

Table 7. ANOVA results based on year of study

		Sum of squares	Df	Means of squares	F	p	Significant Variation
CONT	Among groups	29.030	2	14.515	2.165	.118	no significance
	Within groups	1226.884	183	6.704			
	Total	1255.914	185				
SAT	Among groups	156.847	2	78.424	8.024	.000	4-3 and 4-2
	Within groups	1788.551	183	9.774			
	Total	1945.398	185				
COM	Among groups	126.612	2	63.306	8.457	.000	4-3 and 4-2
	Within groups	1369.952	183	7.486			
	Total	1496.565	185				
VLE overall	Among groups	843.145	2	421.572	9.008	.000	4-3 and 4-2
	Within groups	8564.753	183	46.802			
	Total	9407.898	185				

As Table 8 shows, there are no significant differences between the participants' opinions on overall Virtual Learning Environment Satisfaction [t(.061)=184, p>.05] and its sub-dimensions contribution [t(-.362)=184, p>.05], satisfaction [t(.667)=184, p>.05] and communication [t(-.275)=184, p>.05] in terms of gender variable. In this case, it can be expressed that the female and male participants' opinions on Virtual Learning Environment Satisfaction are similar to each other.

Table 8. The results of t-test based on gender differences

	Gender	N	$\bar{X}$	SS	df	t	p
CONT	Female	77	15.8961	2.57306	-.362	184	.718
	Male	109	16.0367	2.63849			
SAT	Female	77	18.9740	3.04343	.667	184	.505
	Male	109	18.6514	3.38399			
COM	Female	77	14.8831	2.94231	-.275	184	.783
	Male	109	15.0000	2.78554			
VLE overall	Female	77	49.7532	7.23325	.061	184	.951
	Male	109	49.6881	7.09159			

5. DISCUSSION AND CONCLUSION

The results of ANOVA analysis indicate that there are no significant differences between the participants' opinions on the contribution sub-dimension while there are significant differences on satisfaction and communication sub-dimensions, and overall on Virtual Learning Environment satisfaction in terms of year of study. In this case, based on the Scheffe test, it can be stated that 4th year students' scores in satisfaction, communication, and overall Virtual Learning Environment satisfaction are higher than the 3rd and 2nd year students' opinions. The results of t-test analysis reveal that there are no significant differences between the participants' opinions on overall Virtual Learning Environment satisfaction and its sub-dimensions in terms of gender variable. In this case, it can be expressed that the female and male participants' opinions on Virtual Learning Environment Satisfaction are similar to each other. Similarly, Chua and Montalbo (2014) revealed in their study that there was no significant difference between the scores of male and female respondents in all dimensions.

The construct validity of the developed scale was examined with EFA and CFA. The KMO sample consistency coefficient (0.92) and the Barlett Sphericity test value of 1805,933 (p <.001, SD = 105) were superior fit for the data obtained from 206 students to EFA for factor analysis. In EFA, a 3-factor structure is described which accounts for 66.64% of the total variance in the principal components and varimax return results. The CONT subscale is 6, the SAT subscale is 5, and the COM subscale is 4. The CONT subscale describes 21.98% of the total variance, 23.59% of the SAT subscale total variance, and 21.06% of the COM subscale total variance. Factor loads are between 0.560 and 0.704 for the CONT subscale, 0.747 and 0.747 for the SAT subscale, and 0.605 and 0.840 for the COM subscale, respectively.

The data from 186 students were analyzed to confirm the factor structure of the scale developed with CFA. In order to demonstrate the adequacy of the model tested with CFA, the fit indices of the three factor model consisting of 15 items were examined. The standard solutions and t-values of 2 items serving for the CONT dimension were excluded on the grounds that these items were not meaningful for the factor. Moreover, when the correction proposal for thirteen items was examined with CFA, it was concluded that there was a significant decrease in chi-square value between the third and fourth items as well as the ninth and eighth items and that this might be of critical importance for the developed model (Cokluk, Sekercioglu & Buyukozturk, 2012, p. 312). The fit indices (χ^2 = 145.13, N = 62, p = 0.00), RMSEA = 0.085, GFI = 0.89, AGFI = 0.84, CFI = 0.94, NFI = 0.91, and SRMR = 0.06. Factor loads for the three-dimensional model obtained from the first-level CFA range from 0.58 to 0.67 for the CONT sub-dimension, from 0.52 to 0.80 for the SAT sub-dimension, and from 0.64 to 0.69 for the COM sub-dimension. Since all the factors have a value higher than 0.45, thirteen items were important items for the three dimensions considered. In addition, the t values of the items of CONT factor were 7.45, 8.84, 8.33, and 8.00, respectively. The same for SAT factor were 9.05, 12.09, 10.69 and 7.05, respectively; and 9.05, 9.39, 9.66 and 9.37, respectively for the COM factor. The calculated t values are greater than 1.96 and at 0.05 level (Jöreskog & Sörbom, 1993; Kline, 2011; Cokluk, Sekercioglu, & Buyukozturk, 2012, p. 304) is significant at the 0.01 level, and the number of people in the research group is at a sufficient level for factor analysis. With the first-level CFA, the number of people in the research group was at a sufficient level for factor analysis. Furthermore, the superior and acceptable fit measures for the fit indices examined show that the three-factor model from the CFA is acceptable and that the factor structure identified in the EFA is validated.

The second level CFA was used to determine the extent to which the subscales fit into the scale’s implicit variable, which is defined as a superstructure. RMSEA = 0.085, GFI = 0.89, AGFI = 0.84, CFI = 0.94, NFI = 0.91, and SRMR = 0.067, respectively, as the result of the analysis (χ^2 = 145.13, N = 62, p = 0.00) and these values were sufficient. The factor loads for CONT, SAT, and COM, defined as sub-dimensions of the scale implicit variable, appear to be 0.68, 0.78, and 0.95, respectively. Accordingly, it can be said that it can be measured by a three factor structure called CONT, SAT, and COM. The fit indices obtained from the first and second level CFA confirm the validity of the developed scale. In addition, all t-values were significant at 0.01 level; it was established that the CONT, SAT, and COM subscales were significant predictors of the scale implicit variable.

Jöreskog and Sörbom (1996) state that examining the R^2 values is a strong indicator of the significance of the items and factors of the scale. It turns out that the R^2 values of the items of the adapted scale are above 30% in terms of the explained variance. The 3 factors of the scale showed more than 30% variance of explanatory state fit indices on the scale.

The reliability of the scale was examined by the internal consistency coefficient (Cronbach Alpha) and test-retest methods. The internal consistency coefficient obtained from the data was 0.92 for the scale, 0.87 for the CONT subscale, 0.83 for the SAT subscale, and 0.81 for the COM subscale. 31 students from the Department of Computer and Instructional Technologies participated in the test-retest reliability analysis and the correlation value was obtained as 0.94 for the scale itself, while for the subscales CONT, SAT, and COM the same was 0.94, 0.87, and 0.95 respectively. The findings show that there is a sufficient level of reliability coefficients for all of the scale and its subscales.

Findings from the study provide evidence of the validity and reliability of the scale developed by the researchers. The increasing importance of virtual learning in educational environments today, and the lack of adequate means of measuring VLE satisfaction mean that the developed scale can be an instrument to be used in future research.

In addition to the development of an instrument, this study presents the findings of exploring students' expectations from a VLE system. Such findings will be useful for stakeholders such as instructors, managers and parents to reach key factors that will provide satisfaction in teaching.

Clearly, work presented here may have certain limitations. The first one concerns the sample of the study; the participants used for the development of the scale were drawn from a single sample - a particular university. The reason behind employing these participants was the fact that the participants had to be experienced in using VLEs. However, it worth noting that two different groups of participants were employed during the process of scale development, and scale administration. Secondly, invariance design analysis was not conducted since the developed scale was administered to the participants drawn from the same sample, yet it is highly recommend that test of measurement invariance is conducted if the scale is going to be administered to participants from different contexts. Last but not least, the scale was originally developed in Turkish language and if this scale is going to be administered in a foreign culture, scale adaptation studies should be conducted. We hope that further studies undertake the task of creating a richer item pool, which can be followed by meetings –qualitative in nature- with students. Moreover, it should be noted that science advances cumulatively. Since technology changes constantly, so do the needs. The individuals feel satisfied when their needs are met. Therefore, further research can address the needs of the students on the basis of technological developments and the dimensions of satisfaction may further be developed.

ORCID

Nazire Burcin Hamutoglu https://orcid.org/0000-0003-0941-9070
Orhan Gemikonakli https://orcid.org/0000-0002-0513-1128
Merve Savasci https://orcid.org/0000-0002-4906-3630
Gozde Sezen Gultekin https://orcid.org/0000-0002-2179-4466

6. REFERENCES

Al-Khalifa, H. S. (2009). JUSUR: The Saudi Learning Management System. In Proceedings of 2nd Annual Forum on e-Learning Excellence in the Middle East, Dubai, UAE.

Asoodar, M., Vaezi, S., & Izanloo, B. (2016). Framework to improve e-learner satisfaction and further strengthen e-learning implementation, *Computers in Human Behavior*, *63,* 704-716.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*(1), 74-94.

Barker, J., & Gossman, P. (2013). The learning impact of a virtual learning environment: students' views teacher education advancement network. *Journal University of Cumbria, 5*(2), 19-38.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological), 16*(2), 296-298.

Bell, M., & Farrier, S. (2008). Measuring success in e-learning – a multi-dimensional approach. *The Electronic Journal of e- Learning*, *6*(2), 99-110.

Beluce, A. C., & Oliveira, K. L. D. (2015). Students' motivation for learning in virtual learning environments. *Paidéia (Ribeirão Preto)*, *25*(60), 105-113.

Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31,* 419-456.

Bentler, P.M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bermingham V. (2016). Student feedback. In C. Ashford & J. Guth (Eds.), *The legal academic's handbook* (pp. 99-101). London, the UK: Palgrave Macmillan.

Briggs, R. O., Reinig, B. A., & de Vreede, G. J. (2008). The yield shift theory of satisfaction and its application to the IS/IT domain. *Journal of the Association for Information Systems, 9*(5), 267-293.

Briggs, R. O., Reinig, B. A., & de Vreede, G. J. (2014). An empirical field study of the Yield Shift Theory of satisfaction. *47th Hawaii International Conference on In System Sciences (HICSS),* 492-499.

Brown, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Bryman, A., & Cramer, D. (1999). *Quantitative data analysis with SPSS release 8 for Windows. A guide for social scientists.* London and New York: Taylor & Francis Group.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Yayıncılık.

Cokluk, O., Sekercioglu, G., & Buyukozturk, S. (2012). *Sosyal bilimler icin cok degiskenli SPSS ve LISREL uygulamalari.* Ankara: Pegem Yayincilik.

Byrne, B.M., & Campbell, T.L. *(*1999*).* Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology*, *30,* 555-574*.*

Cassidy, S. (2016). Virtual learning environments as mediating factors in student satisfaction with teaching and learning in higher education. *Journal of Curriculum and Teaching, 5*(1), 113-123.

Cheng, K. W. (2011). The gap between e-learning managers and users on satisfaction of e-learning in the accounting industry. *Journal of Behavioral Studies in Business, 3*, 70-79.

Cheng, X., Wang, X., Huang, J. & Zarifis, A. (2016). An experimental study of satisfaction response: evaluation of online collaborative learning. *International Review of Research in Open and Distributed Learning*, *17*(1), 60-78.

Ching, Y. H., & Hsu, Y. C. (2015). Online Graduate Students' Preferences of Discussion Modality: Does Gender Matter? *Journal of Online Learning and Teaching*, *11*(1), 31.

Chou, S. W., & Liu, C. H. (2005). Learning effectiveness in a Web-based virtual learning environment: A learner control perspective. *Journal of Computer Assisted Learning*, *21*(1), 65-76.

Chua, C., & Montalbo, J. (2014). Assessing students' satisfaction on the use of Virtual Learning Environment (VLE): An input to a campus-wide e-learning design and implementation. *Information and Knowledge Management, 3*(4), 108-116.

Cutmore, T. R., Hine, T. J., Maberly, K. J., Langford, N. M., & Hawgood, G. (2000). Cognitive and gender factors influencing navigation in a virtual environment. *International Journal of Human-Computer Studies*, *53*(2), 223-249.

De Lange, P., Suwardy, T., & Mavondo, F. (2003). Integrating a virtual learning environment into an introductory accounting course: determinants of student motivation. *Accounting Education*, 12(1), 1-14.

Eom, S. B., Wen, H. J., & Ashill, N. (2006). The determinants of students' perceived learning outcomes and satisfaction in university online education: An empirical investigation. *Decision Sciences Journal of Innovative Education*, *4(*2), 215-235.

Field, A. (2009). *Discovering statistics Using SPSS*. London: SAGE Publications Ltd.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, 382-388.

Forteza, F. R., Oltra, A., Miquel, A., & Coy, R. P. (2015). University students and virtual learning environments: motivation, effectiveness and satisfaction. *Social & Economic Revue*, *13(*4), 50-54.

Goulão, M. D. F. (2013). Virtual learning styles: does gender matter?. *Procedia-Social and Behavioral Sciences*, *106*, 3345-3354.

Gulbahar, Y. (2012). Study of developing scales for assessment of the levels of readiness and satisfaction of participants in e-learning environments. *Ankara University Journal of Faculty of Educational Sciences*, *45*(2), 119-137.

Gunn, C., McSporran, M., Macleod, H., & French, S. (2003). Dominant or different? Gender issues in computer supported learning. *Journal of Asynchronous Learning Networks*, *7*(1), 14-30.

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: User's guide*. Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tahtam, R. L. (2006). *Multivariate data analysis*. Upper Saddle River: Prentice Hall.

Hettiarachchi, S., & Wickramasinghe, S. (2016). Impact of virtual learning for improving quality of learning in higher education. *2 nd International Conference on Education and Distance Learning – 1 st July 2016, Colombo, Sri Lanka.*

Hew, T. S., & Syed Abdul Kadir, S. L. (2016). Predicting instructional effectiveness of cloud-based virtual learning environment. *Industrial Management & Data Systems*, *116*(8), 1557-1584.

Horvat, A., Dobrota, M., Krsmanovic, M., & Cudanov, M. (2015). Student perception of Moodle learning management system: a satisfaction and significance analysis. *Interactive Learning Environments*, *23*(4), 515-527.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Kember, D., & Ginns, P. (2012). *Evaluating teaching and learning: A practical handbook for colleges, universities and the scholarship of teaching*. New York, NY: Routledge.

Kline, R.B. (2011). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press.

Kolburan-Gecer, A., & Deveci-Topal, A. (2015). Development of satisfaction scale for e-course: Reliability and validity study. *Journal of Theory and Practice in Education*, *11*(4), 1272-1287.

Koskela, M., Kiltti, P., Vilpola, I., & Tervonen, J. (2005). Suitability of a Virtual Learning Environment for Higher Education. *Electronic Journal of e-Learning*, *3*(1), 23-32.

Ku, H. Y., Tseng, H. W., & Akarasriworn, C. (2013). Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. *Computers in Human Behavior*, *29*(3), 922-929.

Lang, B. A., Dolmans, D.H.J.M., Muijtjens, A.M.M., & van der Vieuten, C.P.N. (2006). Student perceptions of a virtual learning environment for a problem-based learning undergraduate medical curriculum. *Medical Education*, *40*(6), 568-575. http://dx.doi.org/10.1111/j.1365-2929.2006.02484.x

Lee, J., Hong, N. L., & Ling, N. L. (2001). An analysis of students' preparation for the virtual learning environment. *The Internet and Higher Education*, *4*(3), 231-242.

Lee, S. J., Srinivasan, S., Trail, T., Lewis, D., & Lopez, S. (2011). Examining the relationship among student perception of support, course satisfaction, and learning outcomes in online learning. *The Internet and Higher Education, 14*(3), 158-163.

Leese, M. (2009). Out of class—out of mind? The use of a virtual learning environment to encourage student engagement in out of class activities. *British Journal of Educational Technology*, *40*(1), 70-77.

Liaw, S. (2008). Investigating students' perceived satisfaction, behavioural intention, and effectiveness of e-learning: A case study of the Blackboard system. *Computers and Education*, *51*(2), 864–873. http://dx.doi.org/10.1016/j.compedu.2007.09.005

Lin, W. S. (2012). Perceived fit and satisfaction on web learning performance: IS continuance intention and task-technology fit perspectives. *International Journal of Human-Computer Studies,70(*7), 498–507. Special Issue on User Experience (UX) in Virtual Learning Environments, http://dx.doi.org/10.1016/j.ijhcs.2012.01.006

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal component analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182-212). New York, NY: Routledge.

Lowes, S., Lin, P., & Kinghorn, B. R. (2016). Gender differences in online high school courses. *Online Learning*, *20*(4), 100-117.

Love, N., & Fry, N. (2006). Accounting students' perceptions of a virtual learning environments: Springboard or safety net? *Accounting Education*, *15*(2), 151-166. http://dx.doi.org/10.1080/06939280600609201

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149.

Marsh, H.W., Hau, K.T., Artelt, C., Baumert, J., & Peschar, J.L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311-360.

Martins, L. L., & Kellermanns, F. W. (2004). A model of business school students' acceptance of a web-based course management system. *Academy of Management Learning & Education*, *3*(1), 7-26.

McGill, T. J., & Hobbs, V. J. (2008). How students and instructors using a virtual learning environment perceive the fit between technology and task. *Journal of Computer Assisted Learning*, *24*(3), 191-202.

Naveh, G., Tubin, D., & Pliskin, N. (2010). Student LMS use and satisfaction in academic institutions: The organizational perspective. *Internet and Higher Education*, *13*, 127–133.

OECD (2005). E-learning in Tertiary Education: Where do we stand? Paris: Centre for Educational Research and Innovation.

Ogba, I. E., Saul, N., & Coates, N. F. (2012). Predicting students' attitudes towards advertising on a university Virtual Learning Environment (VLE). *Active Learning in Higher Education*, *13*(1), 63–75.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows.* Australia: Australian Copyright.Perkowski, J. (2013). The role of gender in distance learning: A meta-analytic review of gender differences in academic performance and self-efficacy in distance learning. *Journal of Educational Technology Systems*, *41*(3), 267-278.

Pituch, K. A., & Lee, Y.-K. (2006). The influence of system characteristics on e-learning use. *Computers & Education, 47*, 222–244.

Reinig, B. A., Briggs, R. O., & Vreede, G. J. de. (2009). A cross-cultural study of the relationship between perceived changes in likelihood of goal attainment and satisfaction with technologically supported collaboration. *International Journal fore-Collaboration, 5*(2), 61-74.

Rienties, B., Giesbers, B., Lygo-Baker, S., Ma, H. W. S. & Rees, R. (2016). Why some teachers easily learn to use a new virtual learning environment: A technology acceptance perspective. *Interactive Learning Environments, 24(*3), 539-552. DOI:10.1080/10494820.2014.881394.

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Shubina, M. (2016). *Usability evaluation of MOOC platforms. Usability evaluation of massive open online course platforms*. Bachelor’s Thesis Degree Programme in Business Information Technology, Haaga Helia University of Applied Sciences.

Stricker, D., Weibel, D., & Wissmath, B. (2011). Efficient learning using a virtual learning environment in a university class. *Computers & Education, 56*(2), 495-504.

Šumak, B., Polancic, G., & Hericko, M. (2010, February). An empirical study of virtual learning environment adoption using UTAUT. In *Mobile, Hybrid, and On-Line Learning, 2010. ELML'10. Second International Conference on* (pp. 17-22). IEEE.

Šumak, B., Hericko, M., Pusnik, M., & Polancic, G. (2011). Factors affecting acceptance and use of Moodle: An empirical study based on TAM, *Informatica, 35*, 91–100.

Sun, P., Tsai, J. R., Finger, G., Chen, Y., & Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education, 50*(4), 1183−1202.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

UCISA Report (2014). 2014 Survey of technology enhanced learning for higher education in the UK. Retrieved on October 17, 2016 from https://www.ucisa.ac.uk/publications/tel_casestudies2016.

Unsal, H. (2012). Harmanlanmış öğrenmenin başarı ve motivasyona etkisi. *Journal of Turkish Educational Sciences, 10*(1), 1-27.

Usta, I., Uysal, O., & Okur, M. R. (2016). Çevrimiçi öğrenme tutum olcegi: Gelistirilmesi, gecerligi ve guvenirligi. *Journal of International Social Research*, *9*(43), 2215-2222.

Van Raaij, E. M., & Schepers, J. J. (2008). The acceptance and use of a virtual learning environment in China. *Computers & Education*, *50*(3), 838-852.

Vaz, M. S. M. G., de Bittencourt, D. F., Vaz, M. C. S., & Júnior, A. S. (2015). Information technology and communication and educational practices. *Iberoamerican Journal of Applied Computing*, *5*(1), 7-15.

Vuopala, E., Hyvönen, P., Järvelä, S. (2016). Interaction forms in successful collaborative learning in virtual learning environments. *Active Learning in Higher Education, 17*(1), 25–38.

Wang, Y. S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. *Information & Management*, *41*, 75–86.

Yilmaz, R. M., Karaman, A., Karakuş, T., & Goktas, Y. (2014). İlköğretim öğrencilerinin 3 boyutlu sanal öğrenme ortamlarına yönelik tutumları: Second life örneği. *Ege Eğitim Dergisi, 15*(2), 538-555.

Yukselturk, E., & Bulut, S. (2009). Gender differences in self-regulated online learning environment. *Journal of Educational Technology & Society*, *12*(3), 12-22.

Zafra, A., Gibaja, E., Luque, M., Ventura, S. (2011). An Evaluation of the Effectiveness of e-Learning System as Support for Traditional Classes. 7th International Conference on Next Generation Web Services Practices. October 19-21, 2011. Salamanca, Spain.

Zerihun, Z., Beishuizen., J., & Os, W. (2012) Student learning experience as indicator of teaching quality. *Educational Assessment, Evaluation and Accountability, 24*(2), 99-111.

Zhang, D., Zhao, J. L., Zhou, L., & Nunamaker Jr, J. F. (2004). Can e-learning replace classroom learning?. *Communications of the ACM*, *47*(5), 75-79.

Zhu, C. (2012). Student satisfaction, performance and knowledge construction in online collaborative learning. *Journal of Educational Technology and Society*, *15*(1), 127-136.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 223–234

DOI: 10.21449/ijate.402690

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

Analysis of the Relationship between University Students' Problematic Internet Use and Loneliness

Mustafa Tevfik Hebebci [1], **Mack Shelley** [2,*]

[1] Educational Science Institute, Necmettin Erbakan University, Konya, Turkey
[2] Department of Statistics, 1413 Snedecor Hall, Iowa State University, Ames, Iowa 50011-1210

Abstract: The computer is part of the information and communication age, and the Internet today is the most used communication tool. Studies have shown that there is a relationship between problematic Internet use and loneliness. The aim of this study was to investigate the relationship between problematic Internet use sub-scales and loneliness. In this study, data were collected from the college students at an Anatolian University in Turkey. The participants of this study consisted of 392 undergraduates. Of the participants, 43% are male (n = 167) and 57% female (n = 225). The average age for the participants is 22 years old. The Problematic Internet Use Scale and the University of California, Los Angeles (UCLA) Loneliness Scale were used as data collection instruments. In the present study, structural equation modeling (SEM) procedures are used to explore the relationships that exist among the variables. The findings of the study revealed that while university students' social benefit/social comfort of Internet has a direct effect on their excessive Internet use and negative consequences, it is related to the loneliness level indirectly. In addition, it is seen in the research model that with an increase in the negative consequences of the Internet, the loneliness level was raised. Another result from the study is that when university students' excessive Internet use increased, their loneliness level decreased.

ARTICLE HISTORY

Received: 27 November 2017
Revised: 28 February 2018
Accepted: 03 March 2018

KEYWORDS

Problematic internet use, loneliness,
structural equation modeling

1. INTRODUCTION

The computer is part of the information and communication age, and the Internet today is the most used communication tool (Koc & Ferneding, 2013). Quick access to the information provided by the Internet is rapidly increasing the ability of individuals to communicate without time and space limitations. The negative effect of increased Internet use on social interaction is considered to be one of the disadvantages and may be associated with feelings of loneliness (Ceyhan, Ceyhan, & Gurcan, 2007; Demirer, Bozoglan, & Sahin, 2013 Eren, Çelik, & Aktürk, 2014). Some users may be affected by negative aspects of Internet use. Positive returns tend to occur when people use the Internet in accordance with the purpose of the online environment (Boz & Adnan, 2017; Bozoglan, Demirer, & Sahin, 2014; Li, Newman, Li, & Zhang, 2016; Pontes, Caplan, & Griffiths, 2016; Tokunaga & Rains, 2016). For example, the goal of increasing the academic achievement of students around the world can be facilitated by providing access to online information resources (Erdogan, 2016; Karahan & Roehrig, 2016).

CONTACT: Mack Shelley ✉ mshelley@iastate.edu Department of Statistics, 1413 Snedecor Hall, Iowa State University, Ames, Iowa 50011-1210

ISSN-e: 2148-7456 /© IJATE 2018

Through the Internet it is possible to communicate with experts who are far away and to share information as if they were present in the same environment as the Internet user. Many innovations that take place in the world have the opportunity to be followed simultaneously on the Internet (Siciliano et al., 2015; Unsal, Sahin, Celik, Akturk, & Shelley, 2012). However, using the Internet outside of its purpose to share information and resources can bring about negative outcomes (Lam & Wong, 2015; Mazzoni et al., 2016; Škařupová, Ólafsson, & Blinka, 2015). For instance, a person who cannot make productive use of the time spent on the Internet is obligated to spend much effort and time in the online environment. A person who spends lots of time on the Internet environment can be dragged into negative behaviors from unknown people who are present on that platform. Those who cast about online without a clear goal in mind cannot use the Internet in a useful way and can be subject in the virtual environment to problematic consequences of Internet use. The concept of Internet addiction has been used in some studies to address the potentially pathological dimension of Internet use (Ceyhan et al., 2007; Sahin, Kesici, & Thompson, 2010; Tutgun, 2009). According to Morahan-Martin and Schumacher (2000), problematic Internet usage, characterized as extensive use of the Internet that is not under control, can result in serious harm to people's lives. Excessive problematic Internet use is associated with emerging social and academic/vocational difficulties that may be associated with negative cognitive and behavioral symptoms within a multidimensional syndrome (Caplan, 2005; Casale, Caplan, & Fioravanti, 2016). In other words, the situation of problematic Internet use, combined with a person's inability to prevent actualization of the desire to use the Internet, may establish the conditions for an adverse impact on daily life (Douglas et al., 2008; Li, Li, & Newman, 2013; Spada, 2014).

According to 2016 data from Global Digital Statistics, the Internet is used worldwide by 3.42 billion people. College students 18-24 years of age are the heaviest users of the Internet. University students' desire to locate and use academic resources is among the reasons for their use of the Internet, in addition to being able to build social relationships through easy and limitless Internet access that provides opportunities to play, watch movies, listen to music, and establish romantic relationships (Ceyhan, 2010). When university students spend more time on the Internet for these purposes, one consequence may be to weaken their prospects to remain in contact with their ability to socialize through the real-world environment. The online environment, if it is adopted by college students as the single environment for conducting social skills, may lead students to neglect friendship relationships because the time spent online may mean that they cannot spend enough time with their social circles in real life and thereby may drift toward loneliness. People who sustain social interactions only through the online environment seem to have reduced opportunity to improve their social skills and seem to feel that they have become lost in their relationships (Lopez-Fernandez et al., 2014; Odabasioglu et al., 2007).

Loneliness is considered to be a different situation than being alone. The difference lies in the fact that loneliness is the condition in which people want to have social relationships but are not successful in doing so, resulting in an unpleasant emotional state (Batıgün, 2010). The unhappiness associated with loneliness results in a number of problems that may arise in one’s professional life, which, combined with the loss of enjoyment of life, can lead to a cascade of problems such as negative results associated with problematic Internet use. Time spent on the Internet reduces the amount of time that could be spent on family and the social environment (Ang et al., 2012; Odaci & Celik, 2013).

In the literature, studies have shown that people who attempt to get rid of their sense of loneliness spend time on online platforms; young people under the age of 25 are especially more likely to have adopted this behavior (Ozturk & Ozmen, 2011). According to Yildiz and Bolukbas (2005), as the duration of Internet usage grows, users are less likely to enter into real social relationships with people and as a consequence suffer from social isolation. To save themselves from their perceived loneliness beyond the scope of interacting with individuals using the Internet, they actually have pushed more people into physical loneliness by staying away from real-life social situations. It is noted that one of the most basic developmental tasks of adolescence is to establish close relationships with peers of the same or opposite gender (Can, 2004; Odacı & Cikrikci, 2014). Communicating effectively prevents problematic Internet use by adolescents and young adults. Bonetti et al. (2010) found that a high level of loneliness poses problems for adolescents and young adults in later ages. Today, as adolescents and young adults have found that loneliness provides them with the opportunity to spend extra time on the Internet, it can be concluded that problems in the social environment occur as a result of improper Internet usage by individuals to isolate themselves. Individuals who prefer to continue to communicate in the online environment rather than being in contact with each other face inevitable problems of inappropriate Internet usage, not to mention their communication problems with other individuals also confronting solitude. Within the young population, the rate of spread of problematic Internet usage is greatest among college-age youth. In another study conducted with university students, perceived social support and loneliness variables were found to be significant predictors of problematic internet use (Oktan, 2015). Despite there are some studies on loneliness and problematic internet use in the relevant literature (Demirer et al. 2013; Derbyshire et al., 2012; Moreno, Jelenchick, & Christakis, 2013; Ozgur, Demiralay, & Demiralay, 2014), to the best of our knowledge, there is no study that examines the relations between the sub dimensions of problematic internet use and loneliness based on a model. The determining the sub scales associated with problematic internet use and level of perceived loneliness may contribute to literature. Thus, problematic internet usage of university students can give information about their loneliness levels. The research conducted for this study was carried out on university students. The aim of the study was to investigate the relationship between problematic Internet use and loneliness.

2. METHOD

2.1. Participants

In this study, data were collected from the college students at an Anatolian University in Turkey in spring term of 2016-2017. University students participated in the current study voluntarily after receiving the necessary permission for the research. The participants of this study consisted of 392 undergraduates. Of the participants, 43% are male (n = 167) and 57% female (n = 225). The average age for the participants is 22 years old.

2.2. Data Collection Instruments

Problematic Internet Use Scale (PIUS): Ceyhan et al. (2007) developed the PIUS to measure problematic Internet use. The PIUS is a Likert scale consisting of 33 items rated on a five-point metric ranging from “not appropriate at all” to “very appropriate.” High scores on the scale indicate problematic Internet use and addictive tendencies. The PIUS has three subscales derived from factor analysis: negative consequences of the Internet, social benefit/social comfort, and excessive use. Negative consequences of Internet use include items such as: “I neglect my daily routines for spending more time on Internet,” “Internet makes me experience relationship difficulties with my significant others,” “Internet enslaves me,” and “I am late to my courses and my appointments since I cannot give up using Internet.” The social benefit/social comfort of Internet comprises items such as: “Concealing my name on Internet makes me freer” and “I share my loneliness with Internet.” A few examples of the last subscale, excessive use, are: “I cannot understand how time flows when I am online” and “I cannot give up Internet usage although I want to quit it very much.”

University of California, Los Angeles (UCLA) Loneliness Scale: The UCLA Loneliness Scale was developed by Russell, Peplau, and Cutrona (1980) to determine individuals' perceived loneliness levels. The validity and reliability of the Turkish form was established by Demir (1989). This is a Likert-type scale consisting of 20 items, each with four options. Scores for this scale range from 20 to 80. Higher scores indicate a higher level of loneliness. The standardized UCLA Loneliness Scale has a high level of internal consistency (Cronbach's alpha coefficient=0.96) and test-retest reliability (Spearman-Brown coefficient=0.94).

2.3. Data Analysis

In the present study, structural equation modeling (SEM) procedures are used to explore the relationships that exist among the variables. The SEM procedure is used due to its capacity to test casual associations between constructs with multiple measurement items (Joreskog & Sorbom, 1996). For each endogenous (dependent) variable, an equation is estimated by exogenous (independent) or other endogenous variables from another equation. Both the direct and indirect effects of independent variables on the dependent variables are estimated. Data analyses were conducted using SPSS (Statistical Package for the Social Sciences) 17.0 and AMOS (Analysis of Moment Structures) 16.0 software. Before data analysis, the SEM assumptions were checked. For the normality assumption, the skewness and kurtosis values were in an acceptable range for a normal distribution. Considering the literature regarding sample size in SEM studies, it is stated that the participants more than 200 is acceptable (Harrington, 2009; Kline, 2005).

3. FINDINGS

The structural equation analysis was conducted to test the relationships among the constructs negative consequences of the Internet, social benefit/social comfort, excessive use, and loneliness. In Table 1, the ideal and acceptable fit indices and the actual results for our estimated structural equation research model are presented (Celik, Sahin, & Aydin, 2014; Hu & Bentler, 1999; Jöreskog & Sörbom, 1984; Tanaka & Huba, 1985).

Table 1. Criterion references for fit indices of structural equation model

Criterion References	Ideal Fit Indices	Acceptable Fit Indices	Indices for the Estimated Research Model
χ^2/df	≤ 3	$\leq$ 4-5	1.074
Root Mean Square Error of Approximation(RMSEA)	≤ 0.05	0.06-0.08	0.013
Normed Fit Index (NFI)	≥ 0.95	0.94-0.90	0.998
Comparative Fit Index (CFI)	≥ 0.97	≥ 0.95	0.981
Goodness of Fit Index (GFI)	≥ 0.90	0.89-0.85	0.999
Adjusted Goodness of Fit Index (AGFI)	≥ 0.90	0.89-0.85	0.987
Tucker Lewis Index (TLI)	≥ 0.95	0.94-0.90	0.999

As seen from Table 1, the research model fits the data well ($\chi2$ = 1,074, df = 1, p= 0.300; GFI = 0.999; AGFI = 0.987; CFI = 0.981; TLI = 0.999; NFI = 0.998; RMSEA = 0.013). As depicted in Figure 1, the research model includes three exogenous variables (negative consequences of the Internet, excessive use, and loneliness) for the endogenous variable (social benefit/social comfort). Negative consequences and excessive use also are endogenous with respect to social benefit/social comfort. In the figure representing the SEM, only significant paths are included.

Table 2. Decomposition of Total Effects for Research Model

Predictor variable	Dependent Variable	Total Effect^{a}	Direct Effect	Indirect Effect	Standard Error	Critical Ratio
social benefit/social comfort	negative consequences	0.78	0.78	-	0.05	25.37**
negative consequences	excessive use	0.45	0.45	-	0.03	7.15**
excessive use	loneliness	-0.17	-0.17	-	-3.16	<0.00**
social benefit/social comfort	excessive use	0.52	0.17	0.35	0.05	2.80**
negative consequences	loneliness	0.46	0.53	-0.76	0.05	9.95**
social benefit/social comfort	loneliness	0.33	-	0.33	-	-

a: Total effect= Direct effect + Indirect effect; **: $p < 0.01$

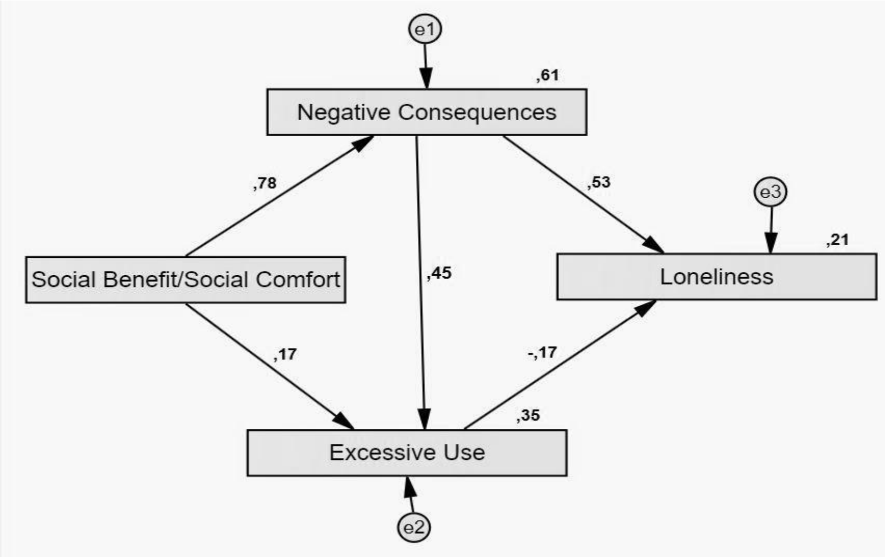


Figure 1. Research Model

In Figure 1, social benefit/social comfort, which is the exogenous variable in the model, has a direct and positive effect on negative consequences (β=0.78) and excessive use (β=0.17). Furthermore, social benefit/social comfort also has an indirect and positive effect on excessive use (β=0.35) and loneliness in the model. Negative consequences of the Internet has a direct and positive effect on both loneliness (β=0.53, p<0.01) and excessive use (β=0.45). In addition, negative consequences of the Internet has an indirect and negative effect on loneliness (β=-0.76). Finally, excessive use had a direct (β=-0.17) and negative effect on loneliness (Table 2).

When each of the separate equations in the model was examined, it can be seen that social benefit/social comfort explains approximately 61% of the variance in negative consequences. Also, negative consequences and social benefit/social comfort together explain approximately 35% of the variance in excessive use. Social benefit/social comfort, excessive use, and negative consequences together explain 21% of the variance in loneliness.

4. DISCUSSION and CONCLUSIONS

This study has demonstrated that excessive Internet use caused individuals to feel themselves less lonely. This finding can be interpreted as meaning that lonely individuals who have difficulties in communicating with their environment feel more comfortable in the Internet environment, satisfy their needs for socializing, and feel themselves less lonely in this environment. Roshoe and Skomski (1989) explain that the individuals who feel themselves lonely consider the Internet as a tool to help relieve loneliness and want to use it gradually more and more. Sheeks and Birchmeier (2007) state that since the online communication environment decreases the anxiety and worry that individuals experience in face-to-face interaction and communication, those with social anxiety tend to use the Internet more compared to others. Other studies conducted on this topic emphasize that the reason for the preference of Internet environments by individuals who feel loneliness could be based on these individuals finding a way to cope with loneliness by interacting with the other individuals in these environments (Ryan & Xenos, 2011; Sheldon, 2008).

When the literature is reviewed, it is seen that contrary to the results of this study, there are several studies that show excessive Internet use leads to increased loneliness (Engelberg & Sjoberg, 2004; Kraut et al., 2002; Moody, 2001; Morahan-Martin, 1999; Pawlak, 2002). It is considered that the difference between these research findings and results reported in the literature originates from cultural factors and differences in the samples. In a study conducted to examine the relationships between increased Internet use, and loneliness and interpersonal styles, Batigun and Hasta (2010) found that individuals with higher Internet use had higher loneliness levels. However, since their research was a correlative study, they could not provide a clear conclusion regarding whether loneliness was an indicator of excessive Internet use, or vice versa. In this context, Morahan-Martin (1999) states that it would not be possible to determine the direction of the mentioned relationship, and asserts that Internet use could cause loneliness. In addition, Morahan-Martin stated that the time individuals with excessive Internet use spend online harms face-to-face communication and social activities, and that Internet use isolates individuals from society and the real world and deprives them of the sense of belonging. Considering the purpose of Internet use among university students in Turkey, it is observed that many college students use to spend time on the social networking (Akar, 2015; Ceyhan, 2010; Çelik, 2012; Karal & Kokoç, 2010). Individuals who use the Internet to socialize on sites such as Facebook, Twitter or Instagram may feel less alone. Facebook usage is very common in Turkey compared to other countries in the world and 37% of Facebook users in Turkey are college students in the 18-24 age range. (Karal & Kokoç, 2010 Aktürk, Emlek, & Çelik, 2017). The level of loneliness of a college student attending a group on Facebook and meeting with his friends may be reduced.

Previous studies showed that individuals spent more time on the Internet to fulfill interpersonal communication needs, create alternative social channels, try to obtain the satisfaction provided by interpersonal relations that cannot be achieved in real life from the Internet, and express themselves more freely on the Internet compared to in daily life (Papacharissi & Rubin, 2000; Peris et al., 2002). Ratunda et al. (2003) state that such individuals reveal significant Internet use characteristics such as spending unnecessary time online and spending more time than actually planned on the Internet. In addition, a study conducted by Caplan (2007) showed that online social interaction provided more privacy compared to face-to-face communication and that individuals with social anxiety perceived less social risk on the Internet.

The findings of this study have revealed that the negative outcomes that emerge as the result of problematic Internet use increased loneliness. This finding indicates results similar to those of previous studies conducted by Kraut et al. (2002), Caplan (2002), Pawlak (2002), and Ozcan and Buzlu (2005), which show a positive relationship between loneliness and problematic Internet use. However, the researchers focus on two different points in explaining this relationship. While some researchers state that problematic Internet use does not increase the level of loneliness, but problematic Internet use emerges as the result of loneliness (Ceyhan & Ceyhan, 2008; Hamburger & Ben-Artzi, 2003), others suggest the opposite view (Morahan-Martin & Schumacher, 2000). Esen and Siyez (2011) state that the causality of this relationship between loneliness and problematic Internet use could be explained through a longitudinal study, which is missing in the literature.

Another finding obtained in this study is the relationship between problematic Internet use and social benefit. The findings of the study showed that the increase in social benefit/social comfort obtained from the Internet also increased excessive Internet use and the negative effects of the Internet. A review of the literature shows studies revealing the existence of both positive and negative relationships between problematic Internet use and social benefit. According to a study conducted by Tanrıverdi (2012), there is a significant, strong, and negative relationship between excessive Internet use and social benefit. In other words, social benefit decreases as excessive Internet use increases. In a study conducted on university students, Ozcan and Buzlu (2005) also found a negative significant relationship between social benefit and excessive Internet use. Contrary to those studies, several other studies in the literature show a positive relationship between problematic Internet use and social benefit (Shaw & Gant, 2002; Silverman; 1999; Winzelberg, 1997). Mossbarger (2008) states that this situation mostly occurs in individuals who use the Internet to play online games with their friends or make new friends in chat rooms, rather than to get information. In a study conducted on high school students, Pawlak (2002) determined that loneliness and social benefit were associated with excessive Internet use, and stated that the lack of social benefit could lead students to excessive Internet use.

When the literature is reviewed, it is seen that there are two different prevailing views regarding social benefit and excessive Internet use. While some researchers (Kraut et al., 1998; Ozcan & Buzlu, 2005) point out that face-to-face interaction and relationships decrease as a result of increased time spent on the Internet, which could cause a decrease in affection, sincerity, and closeness in real life, others assert that the Internet develops the social relationship networks of individuals (Valkenburg & Peter, 2007), and increases social interaction and support (Shaw & Gant, 2002; Silverman, 1999).

Yeh et al. (2008) state that the lack of social benefit causes problematic Internet use. Individuals who experience obstacles in establishing social relationships often refer to the Internet to recreate and continue their personal relationships and tend to prefer the Internet to face-to-face communication (Inderbitzen, Walters, & Bukowski, 1997; Kubey, Lavin, & Barrows, 2001). Individuals who cannot get support from their environments turn to use of the Internet more to socialize in different environments and to create unique social channels. This reveals that the lack of social benefit could be closely related to loneliness (Batigun & Kilic, 2010). In fact, the present study revealed that increased social benefit/social comfort from the Internet, although indirectly, increases loneliness.

In this study, the relationships between university students' problematic Internet uses and their loneliness levels were investigated. The findings of the study revealed that while university students' social benefit/social comfort from the Internet has a direct effect on their excessive Internet use and negative consequences, it is related to the loneliness level indirectly. In addition, it is seen in the research model that increased negative consequences of the Internet are associated with higher levels of loneliness. Another result from the present study is that when university students' excessive Internet use increased, their loneliness level decreased. It is very important to take steps to prevent youth from engaging in problematic Internet use in today's world. Factors related to problematic Internet use should be considered in research studies and the results should be shared with the university students to increase their levels of awareness regarding this problem. Moreover, some university courses may involve contents about safe Internet usage and Internet ethics. This study was carried out with university students. Future studies can be conducted with different sample groups, or could use research designs such as mixed or qualitative methods to yield further insights regarding the relationship between loneliness and problematic Internet use.

ORCID

Mustafa Tevfik Hebebci https://orcid.org/0000-0002-2337-5345
Mack Shelley https://orcid.org/0000-0002-0414-5843

5. REFERENCES

Akar, F. (2015). Purposes, causes and consequences of excessive internet use among Turkish adolescents. *Eurasian Journal of Educational Research, 60,* 35-56. Doi: 10.14689/ejer.2015.60.3

Aktürk, A. O., Emlek, B., & Çelik, İ. (2017). Üniversite Öğrencilerinin Facebook Bağlanma Stratejilerinin ve Yaşam Doyumlarının İncelenmesi. *Mersin University Journal of the Faculty of Education*, *13*(2). Doi:10.17860/mersinefd.336739

Ang, R. P., Chong, W. H., Chye, S., & Huan, V. S. (2012). Loneliness and generalized problematic Internet use: Parents' perceived knowledge of adolescents' online activities as a moderator. *Computers in Human Behavior*, *28*(4), 1342-1347. Doi: 10.1016/j.chb.2012.02.019

Batigun, A. D., & Hasta, D. (2010). İnternet bağımlılığı: Yalnızlık ve kişilerarası ilişki tarzları açısından bir değerlendirme. *Anadolu Psikiyatri Dergisi*, *11*(3), 213-219.

Bonetti, L., Campbell, M. A., & Gilmore, L. (2010). The relationship of loneliness and social anxiety with children's and adolescents' online communication. *Cyberpsychology, Behavior, and Social Networking*, *13*(3), 279-285. Doi: 10.1089/cyber.2009.0215

Boz, B. & Adnan, M. (2017). How do freshman engineering students reflect an online calculus course*? International Journal of Education in Mathematics, Science and Technology (IJEMST)*, *5*(4), 262-278. DOI:10.18404/ijemst.83046.

Bozoglan, B., Demirer, V., & Sahin, I. (2014). Problematic Internet use: Functions of use, cognitive absorption, and depression. *Computers in Human Behavior*, *37*, 117-123.

Can, G. (2004). Kişilik Gelişimi. B. Yeşilyaprak (Ed.) *Gelişim ve öğrenme psikolojisi* (113-142), Ankara: Pegem A Yayıncılık.

Caplan, S.E. (2005). A social skill account of problematic internet use. *Journal of Communication, 55(4), 721-736. Doi: 10.1111/j.1460-2466.2005.tb03019.x*

Caplan, S.E. (2007). Relations among loneliness, social anxiety, and problematic internet use. *CyberPsychology and Behavior*, *10*, 2, 234-242. Doi: 10.1089/cpb.2006.9963

Casale, S., Caplan, S. E., & Fioravanti, G. (2016). Positive metacognitions about Internet use: The mediating role in the relationship between emotional dysregulation and problematic use. *Addictive Behaviors*, *59*, 84-88. Doi: 10.1016/j.addbeh.2016.03.014

Çelik, İ. (2012). Öğretmen adaylarının sosyal ağ (Facebook) kullanımlarının incelenmesi. *Yayınlanmamış yüksek lisans tezi, Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü. Konya*.

Celik, I., Sahin, I., & Aydin, M. (2014). Reliability and Validity Study of the Mobile Learning Adoption Scale Developed Based on the Diffusion of Innovations Theory. *International Journal of Education in Mathematics, Science and Technology*, *2*(4), 300-316. Doi: 10.18404/ijemst.65217

Ceyhan, A. & Ceyhan, E. (2008). Loneliness, depression, and computer self-efficacy as predictors of problematic internet use. *CyberPsychology and Behavior*, 11(6), 699–701. Doi: 10.1089/cpb.2007.0255

Ceyhan, A. (2011). İnternet kullanma temel nedenlerine göre üniversite öğrencilerinin problemli internet kullanımı ve algıladıkları iletişim beceri düzeyleri. *Kuram ve Uygulamada Eğitim Bilimleri*, 59-77.

Ceyhan, E. (2010). Problemli internet kullanım düzeyi üzerinde klinik statüsünün, internet kullanım amacının ve cinsiyetin yordayıcılığı. *Kuram ve Uygulamada Eğitim Bilimleri,* 1323-1355.

Ceyhan, E., Ceyhan, A. A., & Gürcan, A. (2007). The validity and reliability of the Problematic Internet Usage Scale. *Educational Sciences: Theory & Practice*, *7*(1), 411-416.

Demir, A. (1989). UCLA yalnızlık ölçeğinin gecerlik ve güvenirliği [Validity and reliability of UCLA loneliness scale]. *Psikoloji dergisi, 23,* 14–18.

Demirer, V., Bozoglan, B., & Sahin, I. (2013). Preservice teachers' internet addiction in terms of gender, internet access, loneliness and life satisfaction. *International Journal of Education in Mathematics, Science and Technology, 1*(1), 56-63.

Derbyshire, K. L., Lust, K. A., Schreiber, L. R., Odlaug, B. L., Christenson, G. A., Golden, D. J., & Grant, J. E. (2013). Problematic Internet use and associated risks in a college sample. *Comprehensive psychiatry*, *54*(5), 415-422. Doi: 10.1016/j.comppsych.2012.11.003

Engelberg, E., & Sjöberg, L. (2004). Internet use, social skills, and adjustment. *CyberPsychology & Behavior*, *7*(1), 41-47. Doi: 10.1089/109493104322820101

Erdogan, N. (2016). Communities of practice in online learning environments: A sociocultural perspective of science education. *International Journal of Education in Mathematics, Science and Technology*, *4*(3), 246-257. DOI:10.18404/ijemst.20679.

Eren, F., Çelik, İ., & Aktürk, A. O. (2014). Ortaokul öğrencilerinin Facabook algısı: bir metafor analizi. *Kastamonu Eğitim Dergisi*, *22*(2), 635-648.

Esen, E. & Siyez, D. M. (2016). Ergenlerde internet bağımlılığını yordayan psiko-sosyal değişkenlerin incelenmesi. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *4*(36).

Hamburger, Y. A. ve Ben-Artzi E. (2003). Loneliness and Internet use. *Computers in Human Behavior*, *19,* 71–80. Doi: 10.1016/S0747-5632(02)00014-6

Harrington, D. (2008). *Confirmatory factor analysis*. New York, NY: Oxford University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. Doi: 10.1080/10705519909540118

Inderbitzen, H. M., Walters, K. S., & Bukowski, A. L. (1997). The role of social anxiety in adolescent peer relations: Differences among sociometric status groups and rejected subgroups. *Journal of Clinical Child Psychology*, *26*(4), 338-348. Doi: 10.1207/s15374424jccp2604_2

Joreskog, K. G., & Sörbom, D. (1984). *LISREL–VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Karahan, E. & Roehrig, G. (2016). Use of web 2.0 technologies to enhance learning experiences in alternative school settings. *International Journal of Education in Mathematics, Science and Technology*, *4*(4), 272-283. DOI:10.18404/ijemst.32930.

Karal, H., & Kokoç, M. (2010). Üniversite öğrencilerinin sosyal ağ siteleri kullanım amaçlarını belirlemeye yönelik bir ölçek geliştirme çalışması. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *1*(3). Doi: 10.16949/turcomat.85676

Kesici, S., Sahin, I. & Thompson, A. (2010). Prediction of Preservice Teachers' Problematic Internet Use by Their Psychological Needs and Internet Use Functions. In D. Gibson & B. Dodge (Eds.), *Proceedings of Society for Information Technology & Teacher*

Education International Conference 2010 (pp. 1471-1478). Chesapeake, VA: Association for the Advancement of Computing in Education (AACE).

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Koc, M., & Ferneding, K. A. (2013). An Ethnographic Inquiry on Internet Cafes within the Context of Turkish Youth Culture. *International Journal of Education in Mathematics, Science and Technology, 1*(3).

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues*, 58, 49–74. Doi: 10.1111/1540-4560.00248

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being?. *American Psychologist*, *53*(9), 1017. Doi: 10.1037/0003-066X.53.9.1017

Kubey, R. W., Lavin, M. J, & Barrows, J. R. (2001). Internet use and collegiate academic performance decrements: Early findings. *J Communication*, 51:366-382. Doi: 10.1111/j.1460-2466.2001.tb02885.x

Kurtaran, G. T. (2003). *İnternet bağımlılığını yordayan değişkenlerin incelenmesi*. Yayınlanmamış Yüksek lisans tezi. Mersin Üniversitesi, Sosyal Bilimler Enstitüsü, Mersin.

Lam, L. T., & Wong, E. M. (2015). Stress moderates the relationship between problematic Internet use by parents and problematic Internet use by adolescents. *Journal of Adolescent Health*, *56*(3), 300-306. Doi: 10.1016/j.jadohealth.2014.10.263

Li, X., Li, D., & Newman, J. (2013). Parental behavioral and psychological control and problematic Internet use among Chinese adolescents: The mediating role of self-control. *Cyberpsychology, Behavior, and Social Networking*, *16*(6), 442-447. Doi: 10.1089/cyber.2012.0293

Li, X., Newman, J., Li, D., & Zhang, H. (2016). Temperament and adolescent problematic Internet use: The mediating role of deviant peer affiliation. *Computers in Human Behavior*, *60*, 342-350. Doi: 0.1016/j.chb.2016.02.075

Lopez-Fernandez, O., Honrubia-Serrano, M. L., Gibson, W., & Griffiths, M. D. (2014). Problematic Internet use in British adolescents: An exploration of the addictive symptomatology. *Computers in Human Behavior*, *35*, 224-233. Doi: 10.1016/j.chb.2014.02.042

Mazzoni, E., Baiocco, L., Cannata, D., & Dimas, I. (2016). Is internet the cherry on top or a crutch? Offline social support as moderator of the outcomes of online social support on Problematic Internet Use. *Computers in Human Behavior*, *56*, 369-374. Doi: 10.1016/j.chb.2015.11.032

Moody, E. J. (2001). Internet use and its relationship to loneliness. *CyberPsychology & Behavior*, *4*(3), 393-401. Doi: 10.1089/109493101300210303

Morahan-Martin, J. (1999). The relationship between loneliness and Internet use and abuse. *CyberPsychology & Behavior*, *2*(5), 431-439. Doi: 10.1089/cpb.1999.2.431

Morahan-Martin, J., & Schumacher, P. (2000). Incidence and correlates of pathological Internet use among college students. *Computers in human behavior*, *16*(1), 13-29. Doi: 10.1016/S0747-5632(99)00049-7

Moreno, M. A., Jelenchick, L. A., & Christakis, D. A. (2013). Problematic internet use among older adolescents: A conceptual framework. *Computers in Human Behavior*, *29*(4), 1879-1887. Doi: 10.1016/j.chb.2013.01.053

Mossbarger, B. (2008). Is internet addiction addressed in the classroom? A survey of Psychology text books. *Computers in Human Behavior*, 24, 468–474. Doi: 10.1016/j.chb.2007.02.002

Odabasioglu, G. Ozturk, O. Genc, Y., & Pektas, O. (2007). On olguluk bir seri ile internet bağımlılığı-klinik görünümleri. *Bağımlılık Dergisi*, 8, 46-51. Doi: 10.24289/ijsser.279705

Odaci, H., & Celik, Ç. B. (2013). Who are problematic internet users? An investigation of the correlations between problematic internet use and shyness, loneliness, narcissism, aggression and self-perception. *Computers in Human Behavior*, *29*(6), 2382-2387. Doi: 10.1016/j.chb.2013.05.026

Odaci, H., & Cikrikci, O. (2014). Problematic internet use in terms of gender, attachment styles and subjective well-being in university students. *Computers in Human Behavior*, *32*, 61-66. Doi: 10.1016/j.chb.2013.11.019

Oktan, V. (2015). Üniversite öğrencilerinde problemli internet kullanimi, yalnizlik ve algilanan sosyal destek. *Kastamonu Eğitim Dergisi*, 23(1), 281-292.

Ozcan, N. & Buzlu, S. (2005). Problemli İnternet Kullanımını Belirlemede Yardımcı Bir Araç: "İnternet Bilissel Durum Ölçeği'nin Üniversite Öğrencilerinde Geçerlilik ve Güvenirliği. *Bağımlılık Dergisi*, 6(1); 19-26.

Ozgur, H., Demiralay, T., & Demiralay, I. (2014). Exploration of problematic Internet use and loneliness among distance education students. *Turkish Online Journal of Distance Education*, *15*(2), 75-90. Doi: 10.17718/tojde.43009

Ozturk, E. Ozmen, S. (2011). Öğretmen adaylarının problemli internet kullanım davranışlarının, kişilik tipi, utangaçlık ve demografik değişkenlere göre incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, 1785-1808.

Papacharissi, Z., & Rubin, A. M. (2000). Predictors of Internet use. *Journal of Broadcasting & Electronic Media*, *44*(2), 175-196. Doi: 10.1207/s15506878jobem4402_2

Pawlak, C. (2002). Correlates of Internet use and addiction in adolescents. Dissertation Abstracts International Section A: Humanities & Social Sciences, 63(5-A), pp. 1727.

Peris, R., Gimeno, M. A., Pinazo, D., Ortet, G., Carrero, V., Sanchiz, M., & Ibanez, I. (2002). Online chat rooms: Virtual spaces of interaction for socially oriented people. *CyberPsychology & Behavior*, *5*(1), 43-51. Doi: 10.1089/109493102753685872

Pontes, H. M., Caplan, S. E., & Griffiths, M. D. (2016). Psychometric validation of the Generalized Problematic Internet Use Scale 2 in a Portuguese sample. *Computers in Human Behavior*, *63*, 823-833. Doi: 0.1016/j.chb.2016.06.015

Roscoe, B., & Skomski, G. G. (1989). Loneliness among late adolescents. *Adolescence*, *24*(96), 947.

Rotunda, R. J., Kass, S. J., Sutton, M. A., & Leon, D. T. (2003). Internet use and misuse preliminary findings from a new assessment instrument. *Behavior Modification*, *27*(4), 484-504. Doi: 10.1177/0145445503255600

Russell, D., Peplau, L. A. & Cutrona, C. E. (1980). The revised UCLA loneliness scale: Concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology*, *39*, 472–480. Doi: 10.1037/0022-3514.39.3.472

Ryan, T. & Xenos, S. (2011). Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, *27*(5), 1658–1664. doi:10.1016/j.chb.2011.02.004. Doi: 10.1016/j.chb.2011.02.004

Shaw, L. H., & Gant, L. M. (2002). In defense of the Internet: The relationship between Internet communication and depression, loneliness, self-esteem, and perceived social support. *CyberPsychology & Behavior, 5*(2), 157-171. Doi: 10.1089/109493102753770552

Sheeks, M., & Birchmeier, Z. (2007). Shyness, sociability, and the use of computer-mediated communication in relationship development. *Cyber Psychology & Behavior, 10*(1), 64-70. Doi: 10.1089/cpb.2006.9991

Sheldon, P. (2008). The relationship between unwillingness-to-communicate and students' Facebook use. *Journal of Media Psychology, 20*, 67-75. Doi: 10.1027/1864-1105.20.2.67

Siciliano, V., Bastiani, L., Mezzasalma, L., Thanki, D., Curzio, O., & Molinaro, S. (2015). Validation of a new Short Problematic Internet Use Test in a nationally representative sample of adolescents. *Computers in Human Behavior, 45*, 177-184. Doi: 10.1016/j.chb.2014.11.097

Silverman, T. (1999). The Internet and relational theory. *American Psychologist*, 54, 780–781. Doi: 10.1037/0003-066X.54.9.780

Škařupová, K., Ólafsson, K., & Blinka, L. (2015). Excessive internet use and its association with negative experiences: quasi-validation of a short scale in 25 European countries. *Computers in Human Behavior, 53*, 118-123. Doi: 10.1016/j.chb.2015.06.047

Spada, M. M. (2014). An overview of problematic Internet use. *Addictive behaviors, 39*(1), 3-6. Doi: 10.1016/j.addbeh.2013.09.007

Tanriverdi, S. (2012). Ortaöğretim öğrencelerinde internet bağımlılığı ile algılanan sosyal destek arasındaki ilişkinin incelenmesi. *Yayımlanmamış Yüksek Lisans Tezi. Van Yüzüncü Yıl Üniversitesi, Eğitim Bilimleri Enstitüsü, Van.*

Tokunaga, R. S., & Rains, S. A. (2016). A review and meta-analysis examining conceptual and operational definitions of problematic Internet use. *Human Communication Research, 42*(2), 165-199. Doi: 10.1111/hcre.12075

Tutgun, A. (2009). *Öğretmen Adaylarının Problemli İnternet Kullanım Düzeylerinin İncelenmesi. Yayımlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi: İstanbul.*

Tutgun, A. (2012). BÖTE bölümü öğrencilerinin internet kullanım özellikleri ve tercihlerinin incelenmesi. *Online Academic Journal of Information Technology, 3*(6), 27-45. Doi: 10.5824/1309-1581.2012.1.002.x

Unsal, H., Sahin, I., Celik, I., Akturk, A. O., & Shelley, M. (2012). Relationship between secondary school students' perceived social support and their self-efficacy in educational Internet use. In *EdMedia: World Conference on Educational Media and Technology* (pp. 2512-2517). Association for the Advancement of Computing in Education (AACE).

Valkenburg, P. M., & Peter, J. (2007). Preadolescents' and adolescents' online communication and their closeness to friends. *Developmental psychology, 43*(2), 267. Doi: 10.1037/0012-1649.43.2.267

Winzelberg, A. (1997). The analysis of an electronic support group for individuals with eating disorders. *Computers in Human Behavior, 13*(3), 393-407. Doi: 10.1016/S0747-5632(97)00016-2

Yeh, Y. C., Ko, H. C., Wu, J. Y. W. ve Cheng, C. P. (2008). Gender differences in relationships of actual and virtual social support to internet addiction mediated through depressive symptoms among college students in Taiwan. *CyberPsychology & Behavior, 11*, 485-487. Doi: 10.1089/cpb.2007.0134

Yıldız, C., & Bölükbaş, K. (2005). *İnternet kafeler, gençlik ve sosyal sapma*. Ahmet Tarcan (Edt). İnternet ve toplum Ankara: Anı Yayıncılık.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 235–247

DOI: 10.21449/ijate.401069

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

Reliability, Validity and Turkish Adaptation of Self-Directed Learning Scale (SDLS)

Zeynep Işıl Demircioğlu [1], **Burak Öge** [1], **Emine Ezgi Fuçular** [1], **Tuğçe Çevik** [1]
Merve Denizci Nazlıgül [1], **Erol Özçelik** [1]

[1]Cankaya University, Yukarıyurtçu Mahallesi, Eskişehir Yolu 29. Km, Mimar Sinan Caddesi No:4, 06790 Etimesgut, Ankara, Turkey

Abstract: Self-Directed Learning Scale (SDLS) developed by Lounsbury, Levy, Park, Gibson, and Smith (2009) was used for determining individuals' self-directed learning. The purpose of this study was to translate the SDLS into Turkish and to investigate its reliability and validity with a sample of 272 university students. The SDLS, the Modified Schutte Emotional Intelligence Scale (MSEIS), Self-Directed Learning Inventory (SDLI), and the Causal Uncertainty Scale (CUS) for determining convergent validity was applied to the participants. Factor analyses results verified the uni-dimensionality of the scale. The test–retest correlation of SDLS was 0.82, whereas Cronbach alpha coefficient of the scale was founded as 0.85 in the reliability analyses. Correlation coefficients representing for convergent validities varied from -0.30 to 0.72 (p < .01) and criterion validity of the scale was determined as 0.236 when cumulative GPA was used as criterion in the assessment of concurrent validity. The findings suggest that the Turkish adaptation of SDLS is a valid and reliable tool to measure self-directed learning in Turkish samples.

ARTICLE HISTORY

Received: 29 October 2017

Revised: 10 February 2018

Accepted: 15 February 2018

KEYWORDS

self-directed learning,
validity,
reliability,
scale

1. INTRODUCTION

With the advance of technology, it is now easier to access information but difficult to decide on which ones are relevant. Moreover, there is even no obligation to learn this information at schools. Therefore, rather than old-fashioned learning styles, new learning styles are needed. As a result, the concept of self-directed learning gains more importance in this new era. Considering these needs, schools are gradually changing their classical teaching methods and creating more learner-centered environments. Being a self-directed learner is a requirement for all individuals in this information society (Garrison, 1997).

Self-directed learners are "individuals who take primarily initiative action in describing what to learn, why to learn, identifying a personal and material resource for learning; choosing,

CONTACT: Erol Özçelik ✉ ozcelik@cankaya.edu.tr 🖃 Cankaya University, Yukarıyurtçu Mahallesi, Eskişehir Yolu 29. Km, Mimar Sinan Caddesi No:4, 06790 Etimesgut, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

practicing and evaluating the learning outcomes" (Knowles, 1975, p.18). Self-directed learning (SDL) encourages people not only to stay in an observer position but also to have an active role in learning. In SDL, individuals have the control of managing of their own learning. Learners are independent in determining and deciding their own learning goals (Morrow,1993). Self-directed learners act as autonomously and take responsibility for planning, initiating, and evaluating their own learning efforts (Wilcox, 1996). As a result, SDL develops field-specific knowledge as well as the ability to transfer conceptual knowledge to new situations. Individuals can fill the gap between school knowledge and real-world problems more easily (Temple & Rodero, 1995).

According to Kreber (1998), SDL is not only related with a goal, but also with all learning activities to reach this goal. Independent learning is a similar concept as SDL, but it has some differences, as well. Basically, independent learning occurs only if it is based on experimentation and exploration. For instance, Thomas Edison's discovery of the ampoule can be accepted as an example of *independent learning*. However, self-directed learning includes taking responsibility of deciding about what, when and how to learn.

Past research suggests that self-directed learning is affected not only by individual factors but also by environmental factors (Song & Hill, 2007). According to Brockett and Hiemstra (1991), the tendency to be self-directed is higher for women and bachelors than for men and marrieds. Roberson and Merriam suggest that life changes in late ages are directly related to the process of self-directed learning (2005). It was observed that the students who determined their performance standards were more successful than those who did not self-determine their standards (Brownell, Colletti, Ersner-Hershfield, Hershfield, & Wilson, 1977). From this perspective, evolvement of the learner's SDL ability is closely related to the environment and the teacher. For instance, during experiment, teachers bring some tools to the classroom to work on real-life problems. If the duties are meaningful, students will come up with an entertaining approach to tasks, that is to say, students will voluntarily work on them. Thus, students should also be allowed to cooperate with the teacher in determining the deadlines and other arrangements (Temple & Rodero, 1995). On the other hand, if the instructor changes the decision-maker position with learner, SDL can be enhanced. Learners can understand their own needs more deeply and choose more appropriate learning activities (Taylor, 1995). Another example of the effects of environment on SDL is experiment which is demonstrated by Agran and Wehmeyer (2000). They observed that when a lecturer teaches students to set goals, take actions for these goals and revise goals according to the observed improvements, the level of mental retardation of children increased significantly.

There were lots of studies which stressed the positive effects of features of SDL in the literature. For instance, considering that self-evaluation and self-judgment are SDL's characteristics, Schunk (1981) found that the mathematical achievements of students, who evaluated their cognitive strategies verbally and in writing, were increased. With the contribution of proper planning and implementation, leadership patterns of learners evolve through to SDL (Morrow, 1993). It has been found that students become more effective learners and social beings with the help of SDL. They pointed out that self-directed learners have the ability to search for multiple texts, use different strategies to reach the targets, and present their ideas in different forms such as drawing and writing (Guthrie et. al., 1996).

In the literature there is one scale about self-directed learning, namely Self-Directed Learning Readiness Scale (SDLRS) which was developed by Guglielmino (1978). The scale is used to measure attitudes, skills and characteristics that compromise individuals' current level of readiness to manage their learning. In addition, another frequently used scale is Self-Directed Learning Inventory (SDLI) developed by Suh, Wang, and Arterberry (2015). This scale has the goal to measure self-directed learnings in collective cultures in which environmental factors are different from individualistic cultures. According to Suh and colleagues (2015), self-directed learning in Korean culture is different from self-directed learning in other individual cultures. SDLI has 8 subscales which are learning needs, utilizing skills, enduring challenges, self-efficacy in learning, planning skills, completing tasks, evaluation skills, and internal attributions. This scale was translated into Turkish by Çelik and Arslan (2016). Another scale measuring self-directed learning was developed by Lounsbury, Levy, Park, Gibson, and Smith (2009) including 10 items based on a personality approach. This scale's major advantage is its briefness (Lounsbury, et al., 2009).

Noticeably, SDLS's psychometric properties including confirmatory factor analysis, internal consistency and construct validity, was reported by Lounsbury and colleagues (2009). Primarily, internal consistency indicated by correlation coefficient varied from 0.84 to 0.87 in a study with on college students. Moreover, the one-factor structure of the scale was verified by an applied confirmatory factor analysis. To determine convergent validity of SDLS, SDLRS was used and the correlation was found as .82. In addition, a significant relationship between SDLS and a number of personality traits was found. Specifically, the results suggested that although SDLS was positively associated with emotional stability and optimism, it was negatively associated with neuroticism and tension (Lounsbury et al., 2009).

Another important concept in regard with self-directed learning is the average of cumulative grade (GPA) used as an academic performance indicator in education. It is assumed that self-directed capabilities of students have a significant impact on their GPA scores. However, few research studies have examined the relationship between SDL and cumulative GPA. For instance, Hsu and Shiue (2005) found that self-directed learning was related to performance of distance learning. Moreover, Okabayashi and Torrance (1984) found that gifted students had higher self-directed learning. However, none of these studies investigates the relationship between GPA and SDL. To address this need, the present research aims to examine the relationship between self-directed learning and cumulative GPA for university students.

Although a reliability and validity of the SDLS was conducted by Lounsbury and colleagues (2009), there has been no cross-cultural validation of this scale. Thus, the major aim of this study was to examine the psychometric properties of the SDLS in Turkish context with a sample of university students in Turkey. The psychometric examination includes (i) test-retest reliability, (ii) internal consistency, (iii) convergent validity, (iv) factor analyses, (v) and criterion validity of the scale. With respect to criterion validity, this study examined the correlation between cumulative GPA and SDL, unlike previous studies. Moreover, the current study also investigated the relationship between emotional intelligence and self-directed learning to provide the convergent validity of the scale. Emotional intelligence has three subscales including being aware of the own and others' feelings and emotions, noticing different emotions, and using this knowledge to direct thinking and action (Schutte et al., 1998). This research has a potential to reveal the relationship between self-directed learning and emotional intelligence with its subscales. Besides, since the SDLI was administered in a collectivist culture like Korea, the current study can verify the applicability of SDLS in a collectivist culture like Turkey.

To sum up, it is expected that the current research can provide important evidences for reliability and validity of SDLS in a Turkish sample. Moreover, this study may help us to understand the effectiveness of learning processes in educational settings. Also, the results of this study may give more information about self-directed learning of Turkish university students. Lastly, the study may explain differences between individualistic and collectivistic culture's perception of self-directed learning.

2. METHOD

2.1. Participants

Totally, 272 undergraduate students [97 males (35.7%), 175 females (64.3%)] from various universities including Çankaya University, Başkent University, Middle East Technical University, Gazi University, Hacettepe University, Ankara University, Yıldırım Beyazıt University and Karabük University recruited in the study by convenience sampling method. Their ages ranged from 18 to 35, with a mean age of 21.45 (*SD* = 1.99). All participants were Turkish students. The grades and universities of students were shown in Table 1. Of these participants, 166 [53 males (31.9 %), 113 females (68.1%)] of them received the SDLS twice for examining retest reliability. Their ages ranged from 18 to 30 with a mean age of 21.25 (*SD* = 2.35).

Table 1. Descriptive Statistics of participants in this study.

	Frequency	Percentage	Mean±Standard Deviation
Gender			
Female	175	64.3	
Male	97	35.7	
Grade			
1st grade	15	5.5	
2nd grade	114	41.9	
3rd grade	62	22.8	
4th grade	78	28.7	
Unstated	3	1.1	
University			
Çankaya University	111	40.8	
Başkent University	103	37.9	
Middle East Technical University	14	5.1	
Gazi University	12	4.4	
Hacettepe University	13	4.8	
Ankara University	7	2.6	
Yıldırım Beyazıt University	5	1.8	
Karabük University	7	2.6	
Department			
Psychology	85	31.1	
Banking and Finance	37	13.6	
Management Information Systems	28	10.3	
Accounting and Financial Management	25	9.2	
Education	17	6.2	
Political Science and International Relation	13	4.8	
International Trade	10	3.7	
Economics	10	3.7	
Engineering	8	2.9	
Management	6	2.2	
English Language and Literature	5	1.8	
Insurance and Risk Management	5	1.8	
Chemistry	5	1.8	
Others	19	6.8	
Age			21.45±1.98

2.2. Measures

Self-Directed Learning Scale (SDLS). The original Self-Directed Learning Scale was created by Lounsbury et. al. (2009) as a self-report scale. It measures to what extent individuals learn in an autonomous manner through a unidimensional structure. It consists of 10 items rated on a five-point Likert Scale from 1 (strongly disagree) to 5 (strongly agree). Individuals who get higher scores are associated with stronger self-directed learning. Lounsbury at al. (2009) obtained Cronbach alpha of .87 when their sample included middle and high school students. The Cronbach alpha was .84 when the sample included college students. In another study, Zhoc and Chen (2016) applied SDLS in Chinese university students. They obtained internal consistency reliability coefficient of 0.79.

Modified Schutte Emotional Intelligence Scale (MSEIS). The Modified Schutte Emotional Intelligence Scale was developed by Schutte and colleagues (1998) to measure dimensions of emotional intelligence (e.g., optimism/mood regulation, utilization of emotions and appraisal of emotions). It has 41 items and 21 of them are reverse-scored. Its responses are rated between 1 (totally disagree) and 5 (totally agree). Higher scores indicate higher emotional intelligence. Its internal reliability was 0.87. It was translated into Turkish by Tatar, Tok and Saltukoğlu (2011). The Cronbach alpha for the Turkish version of the scale was found as 0.82.

Self-Directed Learning Inventory (SDLI). The Self-Directed Learning Inventory was developed by Suh, Wang, and Arterberry (2015) to measure for elementary to middle school students' self-directness in collectivist cultures. This scale has 8 subscales which are learning needs, utilizing skills, enduring challenges, self-efficacy in learning, planning the process, evaluating the process, completing tasks, and internal attribution. Its internal reliability was 0.82. The Turkish adaptation and validation of the scale was established by Çelik and Arslan (2016). Internal consistency of this inventory was found 0.93. It consists of 28 items and responses are rated between 1 (totally disagree) and 5 (totally agree).

Causal Uncertainty Scale (CUS). The Causal Uncertainty Scale was developed by Weary and Edwards (1994) to measure uncertainty about understanding the cause and effect relationship in social world. The internal consistency of the scale was founded as 0.83 (Weary & Edwards, 1994). It consists of 14 items and responses are rated between 1 (totally disagree) and 5 (totally agree). Higher scores indicate higher uncertainty. This scale was adapted into Turkish by Uz (2015). The Turkish version of the scale's internal consistency was found as 0.82.

2.3. Procedure

First of all, ethical approval was obtained from Çankaya University Ethics Committee. SDLS was translated to Turkish by three expert psychologists. In addition, back-translations were separately done by a psychologist with a specialist degree in i) cognitive psychology, ii) social psychology and by a iii) professional translator. The final version of the translation was approved again by the same three psychologists.

All subjects voluntarily participated in the current study. Before attending, information about the study was explained and informed consent was obtained from all participants. A demographic information form was administered to measure variables including gender, age, university, department, grade, and cumulative GPA. MSEIS, SDLI, and CUS were also applied to all participants in order to examine convergent validity of SDLS. To measure test-retest reliability, SDLS was re-administered after two to four weeks after the first application of the scale.

3. RESULTS

3.1. Reliability Analyses

The test–retest correlation of SDLS was r = 0.820, p < .01. Guttman Split-Half Coefficient was computed for determining internal consistency (split-half correlation). Guttman Split-Half Coefficient of SDLD was found as 0.816. Item total correlations and Cronbach's alpha values (if an individual item deleted) were calculated to assess internal consistency. The Cronbach alpha coefficient of SDLS was found to be 0.853. Item-total item correlations were between 0.43 and 0.63. If item deleted Cronbach's α values were calculated for each item and it was found that α values varied from 0.823 to 0.841 (see Table 2).

Table 2. Mean, standard deviation, item-total correlations, and alpha values of items

Items	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Item1	34.25	29.49	,478	,837
Item2	34.69	27.60	,582	,828
Item3	34.35	28.47	,582	,828
Item4	34.41	28.77	,515	,834
Item5	33.91	30.98	,428	,841
Item6	34.01	29.83	,528	,833
Item7	34.08	30.00	,442	,840
Item8	34.08	28.93	,632	,824
Item9	34.44	28.05	,629	,823
Item10	34.25	27.50	,623	,823

N = 272, α = 0.85

3.2. Validity Analyses

Convergent validity. There was a strong significant positive correlation between participants' SDLS and SDLI scores (r = 0.73, p < .01). The correlation between SDLS and the MSEIS was also significant (r = 0.38, p < .01). There was a significant negative moderate correlation between SDLS and the CUS (r = -0.30, p < .01) (Table 3). The correlations between SDLS and subscales of SDLI varied between 0.33 and 0.60 (p < .05). As seen in Table 4, the correlations between SDLS and subscales of MSEIS varied between 0.082 and 0.402 (p < .05).

Table 3. Descriptive Statistics, alpha coefficients, and correlations of the scales

Scale	Mean	S.D.	SDLS	MSEIS	CUS	SDLI
SDLS	38.01	5.92	(0.853)			
MSEIS	156.53	15.09	0.376**	(0.837)		
CUS	31.05	9.25	-0.304**	-0.473**	(0.879)	
SDLI	105.57	14.03	0.728**	0.459**	-0.338**	(0.905)

N = 272
Alpha coefficients are on the diagonal, in parentheses.
** p < .01.

Table 4. *Correlations between SDLS Scores and (i)MSEIS Scores, (ii) SDLI Scores, (iii) Cum GPA*

Scales	Mean±SD	r
Modified Schutte Emotional Intelligence Scale (MSEIS)		
Optimism/Mood Regulation	47.03±5.39	0.402**
Utilization of Emotions	22.06±3.24	0.082
Appraisal of Emotions	38.62±5.58	0.266**
Self-directed Learning Inventory (SDLI)		
Learning Needs	20.46±3.54	0.503**
Utilizing Skills	14.50±2.60	0.604**
Enduring Challenges	14.31±2.98	0.596**
Self-Efficacy in Learning	11.29±2.34	0.584**
Planning the Process	10.39±2.84	0.371**
Evaluating the Process;	11.06±2.80	0.325**
Completing Tasks;	11.18±2.34	0.408**
Internal Attribution	12.37±1.94	0.411**
Cumulative GPA Scores	2.69±0.58	0.236**

** *p* < .01.

Factor analyses. The one-factor structure of the scale, which was formed by Lounsbury and Gibson (2006), was tested with a confirmatory factor analyses by LISREL 9.2. For one-factor structure, Goodness of Fit Index was found as 0.97, Comparative Fit Index was found as 0.99, Root Mean Square Error of Approximation was found as 0.04, and other scores can be seen in Table 5. The path diagram of the one factor model of the SDLS can be seen in Figure 1.

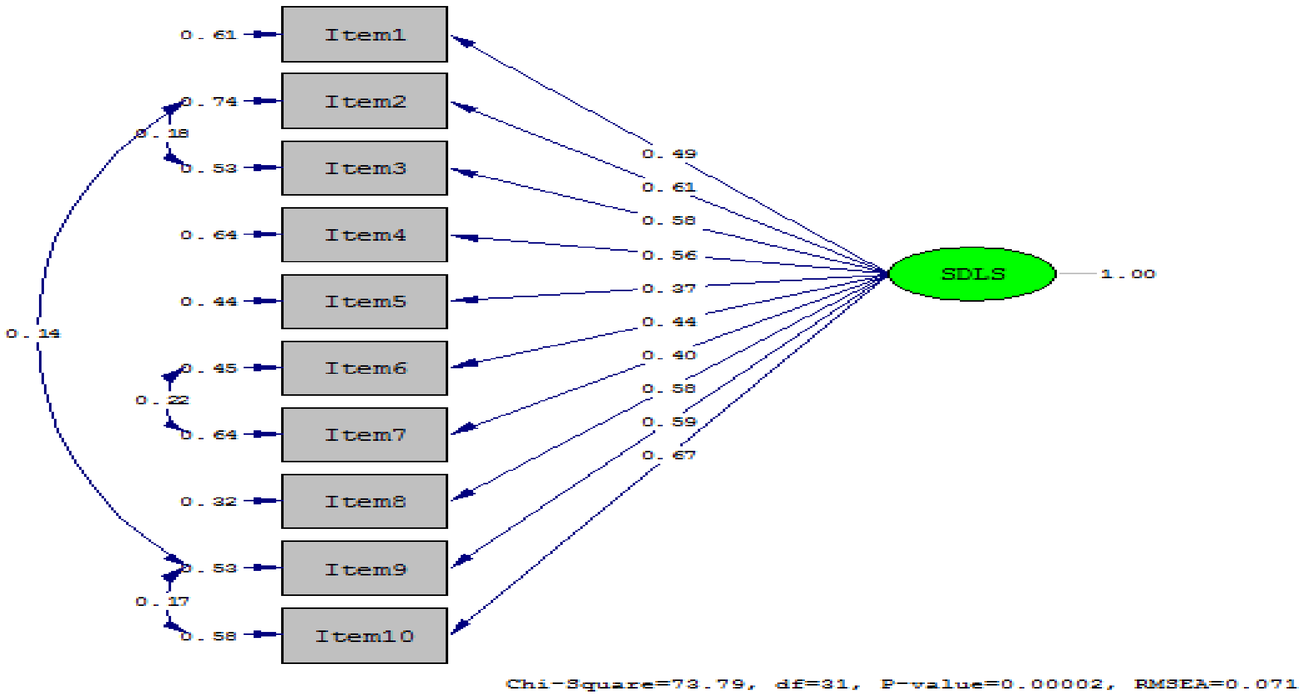


Figure 1. The one-factor Structure of SDLS

Table 5. CFA results for the one-factor model

Fit Indices	Fit Range	Research Model Uni-dimensional Model
Total Fit Index		
χ^2 */df*	*0 ≤* χ^2 */df ≤ 3*	73.79/31= 2.38
Comparative Fit Index		
NNFI	.90 ≥ - ≥ .94	.96
CFI	≥ .95	.97
RMSEA	0.05 ≤ - ≤ 0.08	0.07
Absolute Fit Index		
GFI	≥ .90	.95
AGFI	≥ .85	.91
Residual Based Indexes of Compliance		
SRMR	.06 ≤ - ≤ .08	.05
RMR		.04

Criterion validity. Cumulative GPA scores were used to determine concurrent validity. There existed a positive significant correlation between individuals' SDLS scores and Cumulative GPA scores (r = 0.236, p < .01).

4. DISCUSSION

The purpose of this study was to translate the SDLS into Turkish and to investigate the psychometric properties of the Turkish adaptation of the SDLS. The majority of the sample was composed of university students from Ankara. The psychometric evaluation of the Turkish version of SDLS included examining (i) test-retest reliability, (ii) internal consistency, (iii) convergent validity, (iv) factor analyses, (v) and criterion validity of the scale.

To test-retest reliability, the correlation coefficient was found as 0.82. This result suggests that SDLS was consistent over time, meaning that student who got high self- directed learning scores tend to have high scores in the same scale after some time. Past research studies did not determine test-retest reliability of this scale. For this reason, this study provides information about the reliability of SDLS. Moreover, internal consistency was examined, and the Cronbach's alpha coefficient was found as fairly high, demonstrating that the one-factor structure was internally consistent. This score is similar to the one obtained in Zhoc and Chen's (2015) study, as well as Lounbury and Gibson's (2009) research. Besides, Guttman Split-Half Coefficient was greater than 0.8, indicating that SDLS was reliable.

Additional three scales were used in this study for determining convergent validity of the scale as a part of construct validity examination. Firstly, a significant strong positive correlation was found between SDLS and SDLI. This result indicates not only the convergent validity of SDLS but also applicability of SDLS in collectivist cultures like Turkey. SDLS was used to examine self-directed learning in individualistic cultures. On the other hand, Suh, Wang, and Arterberry (2015) developed SDLI to determine people's self-directed learning in collectivistic cultures. In fact, culture is one of the determinant of measuring self-directed learning (Mok, Leung, & Shan, 2005). According to Brockett (1983), self-directed learners are willing to learn new concepts and they like to learn information independently. On the other hand, independence-interdependence dimension is the most important determinant when distinguishing between individualistic and collectivistic cultures (Triandis, 2001). Considering all these information, the 'self-directed learning' concept can vary according to individualistic or collectivistic cultures. High correlation between SDLS and SDLI demonstrates that SDLS measures self-directed learning not only for individualistic cultures, but also for collectivistic cultures.

In addition, the current study revealed the relations of self-directed learning and emotional intelligence. As founded, the significant positive correlation between SDLS and the MSEIS indicates that students who learn more self-directed tend to be more emotionally intelligent or vice versa. The observed correlation coefficient is lower than previous studies. There can be two reasons for this result. First of all, number of males and females were not balanced in the current study. MSEIS scores of males were significantly lower than MSEIS scores of females. On the other hand, there was no significant difference between SDLS scores of men and of females. Thus, gender can be a confounding variable for determining correlation between SDLS and MSEIS for this study. Second reason may be the small sample size employed in the current work.

There is a significant negative moderate correlation between SDLS and the CUS, indicating that students who are more self-directed tend to be less causal uncertain. According to Markant, Settles, and Gureckis (2016), people generally start learning with a little piece of information. For this reason, self-directed learning people should have little causal uncertainty not only for determining correct sources but also for finding proper methods for themselves. The negative correlation between SDLS and CUS supports this expectation.

In the original study, Lounsbury and Gibson (2006) found a uni-dimentional factor structure of the scale. Supporting past findings, confirmatory factor analysis (CFA) results shows that the SDLS is uni-dimensional. As Browne and Cudeck (1993) suggested, Root Mean Square Error of Approximation (RMSEA) score obtained in our study was lower than 0.08, conforming adequate fit model. Similarly, Goodness of Fit Index (GFI) score of the present study reached the suggested cut off score of 0.95 (Munro, 2005). Adjusted Goodness of Fit Index (AGFI) should be higher than 0.95, but allowance value was suggested to be 0.90 (Munro, 2005). AGFI score in this study was within this range. In addition, Bentler (1990) suggested 0.90 as an allowance score of Comparative Fit Index (CFI). CFI score of the current work was quite higher than this value. Furthermore, Root Mean Square Residual (RMR) and Standardized Root Mean Square Residual (SRMR) should be lower than 0.05 (Hu & Bentler, 1999). In this study, RMR and SRMR scores were found as lower than this threshold. All these results demonstrate that the Turkish version of SDLS fit the one-factor model.

The correlation between SDLS and cumulative GPA of participants was examined for determining the criterion validity of the scale. A significant positive correlation between SDLS scores and cumulative GPA of participants was observed. These results support those of Lounsbury et al. (2009) who found a positive correlation between self-directed learning and academic achievements. All these findings suggest that self-directed learners who are motivated and open to new experiences tend to have higher academic achievement. However, there existed some missing values for cumulative GPA in the data. These missing values might decrease the magnitude of the relationship between SDLS scores and cumulative GPA.

In sum, the results of the current study show that SDLS was a reliable and valid tool to measure self-directed learning for university students in Turkey. SDLS is uni-dimensional and can measure self-directed learning in different cultures. The scale's factor structure was internally consistent. The scale also showed test-retest reliability. Criterion validity of the scale was provided by its correlation with university achievement (i.e., Cumulative GPA). Moreover, the study has broadened the nomothetic span of self-directed learning by relating to emotional intelligence and causal uncertainty.

Although this study will contribute the area of education with clarifying the learning orientation of individuals, the current study has the following limitations. Firstly, the majority of the participants were from one city, Ankara. Secondly, sample size was small. Additionally, the number of students was not equally distributed across universities and gender. Future studies are suggested to select participants from different cities in different cultures to enhance the generalizability of the findings and applicability of SDLS in collectivistic cultures. Additionally, future studies are recommended to collect data from larger samples to strengthen the external validity of the scale. Moreover, future research studies should balance the male and female ratio to minimize a possible confounding effect of gender.

ORCID

Zeynep Işıl Demircioğlu https://orcid.org/0000-0002-5249-5514
Burak Öge https://orcid.org/0000-0002-0029-9626
Emine Ezgi Fuçular https://orcid.org/0000-0002-9740-9202
Tuğçe Çevik https://orcid.org/0000-0002-6744-0583
Merve Denizci Nazlıgül https://orcid.org/0000-0002-6516-7341
Erol Özçelik https://orcid.org/0000-0003-0370-8517

5. REFERENCES

Agran, M., Blanchard, C., & Wehmeyer, M. L. (2000). Promoting transition goals and self-determination through student self-directed learning: The self-determined learning model of instruction. *Education and Training in Mental Retardation and Developmental Disabilities*, 351–364.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Brockett, R. (1983). Self-directed learning and the hard-to-reach adult. *Lifelong Learning: The Adult Years*, *6*(8), 16–18.

Brockett, R. G. & Hiemstra, R. (1991*) Self-direction in adult learning: perspectives on theory, research, and practice.* London: Routledge

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *154*, 136.

Brownell, K. D., Colletti, G., Ersner-Hershfield, R., Hershfield, S. M., & Wilson, G. T. (1977). Self-control in school children: Stringency and leniency in self-determined and externally imposed performance standards. *Behavior Therapy*, *8*(3), 442–455.

Corno, L. (1992). Encouraging students to take responsibility for learning and performance. *The Elementary School Journal*, 93(1), 69-83.

Çelik, K., & Arslan, S. (2016). Turkish adaptation and validation of Self-Directed Learning Inventory. *International Journal of New Trends in Arts, Sports & Science Education (IJTASE)*, 5(1), 19-25.

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. *Adult Education Quarterly*, *48*(1), 18 –33.

Guglielmino, L. M. (1978). *Development of the self-directed learning readiness scale* (Doctoral dissertation, ProQuest Information & Learning).

Guthrie, J. T., Meter, P., McCann, A. D., Wigfield, A., Bennett, L., Poundstone, C. C., & Mitchell, A. M. (1996). Growth of literacy engagement: Changes in motivations and strategies during concept-oriented reading instruction. *Reading Research Quarterly*, *31*(3), 306–332.

Hodgkinson, G. P., Langan-Fox, J., & Sadler-Smith, E. (2008). Intuition: A fundamental bridging construct in the behavioural sciences. *British Journal of Psychology*, *99*(1), 1-27.

Hu, L. T., & Bentler, P. M. (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.

Hsu, Y. C., & Shiue, Y. M. (2005). The effect of self-directed learning readiness on achievement comparing face-to-face and two-way distance learning instruction. *International Journal of Instructional Media*, *32*(2), 143.

Knowles, M. S. (1975) *Self Directed Learning: A Guide for Learners and Teachers*. Chicago: Association Press.

Kilmann, R. H., & Thomas, K. W. (1975). Interpersonal conflict-handling behavior as reflections of Jungian personality dimensions. *Psychological Reports*, *37*(3), 971–980.

Kreber, C. (1998). The relationships between self-directed learning, critical thinking, and psychological type, and some implications for teaching in higher education. *Studies in Higher Education*, *23*(1), 71-86.

Lounsbury, J. W., & Gibson, L. W. (2006). Personal Style Inventory: A personality measurement system for work and school settings. *Knoxville, TN: Resource Associates Inc*.

Lounsbury, J. W., Levy, J. J., Park, S. H., Gibson, L. W., & Smith, R. (2009). An investigation of the construct validity of the personality trait of self-directed learning. *Learning and Individual Differences*, *19*(4), 411-418.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-Directed Learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120.

McCombs, B. L., & Whisler, J. S. (1989). The role of affective variables in autonomous learning. *Educational Psychologist*, *24*(3), 277–306.

Mok, M. C.M., Leung, S. O., & Shan, W. J. P. (2005). A comparative study on the self-directed learning of primary students in Hong Kong and Macau. *International Journal of Self-directed Learning*, *2*(2), 39–54.

Morrow, L. M. (1993). Promoting Independent Reading and Writing through Self-Directed Literacy Activities in a Collaborative Setting. Reading Research Report No. 2.

Munro B.H. (2005) *Statistical Methods for Health Care Research*, 5th edn. Lippincott Williams and Wilkins, Philadelphia, PA.

Okabayashi, H., & Torrance, E. P. (1984). Role of style of learning and thinking and self directed learning readiness in the achievement of gifted students. *Journal of Learning Disabilities*, *17*(2), 104-106.

Roberson Jr, D. N., & Merriam, S. B. (2005). The self-directed learning process of older, rural adults. *Adult Education Quarterly*, *55*(4), 269–287.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology*, *73*(1), 93.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, *25*(2), 167–177.

Song, L., & Hill, J. R. (2007). A conceptual model for understanding self-directed learning in online environments. *Journal of Interactive Online Learning*, *6*(1), 27–42.

Suh, H. N., Wang, K. T., & Arterberry, B. J. (2015). Development and Initial Validation of the Self-Directed Learning Inventory with Korean College Students. *Journal of Psychoeducational Assessment*, *33*(7), 687–697.

Tatar, A., Tok, S., & Saltukoğlu, G. (2011). Gözden geçirilmiş Schutte Duygusal Zekâ Ölçeği'nin Türkçe'ye uyarlanması ve psikometrik özelliklerinin incelenmesi. *Klinik Psikofarmakoloji Bülteni*, *21*(4), 325–338.

Taylor, B. (1995). Self-Directed Learning: Revisiting an Idea Most Appropriate for Middle School Students. Paper presented at *the Combined Meeting of the Great Lakes and Southeast International Reading Association*, Nashville, TN, Nov 11-15. [ED395287]

Temple, C., & Rodero, M. L. (1995). Reading around the World: Active Learning in a Democratic Classroom: The" Pedagogical Invariants" of Célestin Freinet. *The Reading Teacher*, *49*(2), 164–167.

Triandis, H. C. (2001). Individualism, collectivism and personality. *Journal of Personality*, *69*(6), 907–924.

Uz, İ. (2015). Nedensel Belirsizlik Ölçeğinin Türkçeye uyarlanması. *Anatolian Journal of Psychiatry/Anadolu Psikiyatri Dergisi*, *16*.

Weary, G., & Edwards, J. A. (1994). Individual differences in causal uncertainty. *Journal of Personality and Social Psychology*, *67*(2), 308.

Wilcox, S. (1996). Fostering self-directed learning in the university setting. *Studies in Higher Education*, *21*(2), 165–176.

Zhoc, K. C., & Chen, G. (2016). Reliability and validity evidence for the Self-Directed Learning Scale (SDLS). *Learning and Individual Differences*, *49*, 245–250.

Appendix A

Tablo A. Öz Yönetimli Öğrenme Ölçeği (Turkish Version)

Öz Yönetimli Öğrenme Ölçeği						
Aşağıda çeşitli durumlara ilişkin ifadeler bulunmaktadır. Lütfen ifadeyi okuduktan sonra size uyma derecesini sağ taraftaki kutucuklardan birini işaretleyerek belirtiniz.		Kesinlikle Katılmıyorum	Katılmıyorum	Fikrim Yok	Katılıyorum	Kesinlikle Katılıyorum
1.	Sınıf dışında, düzenli olarak kendi kendime bir şeyler öğrenirim.					
2.	Öğretmenin sınıfta açıklamadığı şeylerin cevabını kendi kendime bulmak konusunda oldukça iyiyimdir.					
3.	Sınıfta anlamadığım bir şey olursa, onu kendi kendime öğrenmenin her zaman bir yolunu bulurum.					
4.	Okulda başarılı olmamda yardımcı olacak doğru kaynakları bulmada iyiyimdir.					
5.	Kendi insiyatifim temelinde, öz yönetimli öğrenmeyi (belirlediğim amaca yönelik, kendi öğrenme yöntemimle öğrenmeyi) okulda ve gelecekteki kariyerimde başarı için çok önemli buluyorum.					
6.	Öğreneceğim şeyler için hedeflerimi kendim koyarım.					
7.	Neyi ne zaman öğreneceğimden kendim sorumlu olmak isterim.					
8.	Eğer öğrenmem gereken bir şey varsa, onu öğrenmenin bir yolunu hemen bulurum.					
9.	Çoğu öğrenciye kıyasla, kendi kendine öğrenme konusunda çok daha iyiyimdir.					
10.	Diğer insanlara bel bağlamadan kendi kendime öğrenme konusunda oldukça motiveyimdir.					

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 248–262

DOI: 10.21449/ijate.407193

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

An Iterative Method for Empirically-Based Q-Matrix Validation

Ragip Terzi [1*], **Jimmy de la Torre** [2]

[1] Educational Measurement and Evaluation, Harran University, Sanliurfa, Turkey
[2] Division of Learning, Development and Diversity, The University of Hong Kong, Pokfulam, Hong Kong

Abstract: In cognitive diagnosis modeling, the attributes required for each item are specified in the Q-matrix. The traditional way of constructing a Q-matrix based on expert opinion is inherently subjective, consequently resulting in serious validity concerns. The current study proposes a new validation method under the deterministic inputs, noisy "and" gate (DINA) model to empirically validate attribute specifications in the Q-matrix. In particular, an iterative procedure with a modified version of the sequential search algorithm is introduced. Simulation studies are conducted to compare the proposed method with existing parametric and nonparametric methods. Results show that the new method outperforms the other methods across the board. Finally, the method is applied to real data using fraction-subtraction data.

ARTICLE HISTORY

Received: 02 February 2018
Revised: 14 March 2018
Accepted: 15 March 2018

KEYWORDS

Cognitive diagnosis model,
Q-matrix validation,
DINA,
sequential search algorithm

1. INTRODUCTION

Cognitive diagnosis models (CDMs) require a Q-matrix (Tatsuoka, 1983) to identify the specific subset of attributes measured by each item. The entry q_{jk} in row j and column k of the Q-matrix is 1 if the k^{th} attribute is required to correctly answer item j, and 0 otherwise. Due to its nature, constructing a Q-matrix is usually subjective, which has raised serious validity concerns among researchers. For instance, the estimation of model parameters, and ultimately the accuracy of attribute classifications may be negatively affected by including or omitting multiple q-entries in the Q-matrix (de la Torre, 2011; Rupp & Templin, 2008). However, the Q-matrix is usually assumed to be correct once specified by domain experts. This assumption is generally made because until recently, few well-established methods have become available to detect misspecifications in the Q-matrix (Chiu, 2013; de la Torre, 2008; Rupp & Templin, 2008), particularly when general CDMs are involved (de la Torre & Chiu, 2016; Liu, Xu, & Ying, 2012; Terzi, 2017). Any analysis, such as model-fit evaluation, that does not check the correctness of the Q-matrix, becomes questionable.

CONTACT: Ragıp Terzi ✉ terziragip@harran.edu.tr Harran University, School of Education, E Blok, Osmanbey Yerleskesi, 63190, Sanliurfa, Turkey

This study is derived from a part of Ph.D. dissertation of the first author.

ISSN-e: 2148-7456 /© IJATE 2018

These concerns have led to developments of some statistical methods for validating the appropriateness of Q-matrix specifications. One of the earlier studies on the Q-matrix validation was introduced by de la Torre (2008) for the deterministic inputs, noisy “and” gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model. It is an empirically based δ-method that defines the correct q-vector for each item. In doing so, the discrimination index of item *j*, δ_j, is estimated. The index δ_j is the difference in the probabilities of correct responses between examinees who have mastered the required attributes and those who have not. Using the δ-method, two algorithms were discussed in de la Torre (2008). However, the algorithms have some limitations. As noted by de la Torre (2008), an incorrect Q-matrix because of over- and under-specifications of attributes can cause bias in parameter estimation. This issue cannot be completely addressed by the algorithms because they usually choose q-vectors with all attributes specified. For one of the algorithms, the *sequential search algorithm* (SSA), it is also not clear what cut-off values should be used in practice because it could vary depending on many conditions, such as changes in sample sizes, test lengths, item qualities, and amount of misspecifications, all of which were fixed in de la Torre (2008)'s paper. It should also be noted that the algorithm was not implemented iteratively, meaning that the validation method stops after one full iteration even if changes are made in the provisional Q-matrix.

Another method, the Q-matrix refinement method (QRM), was proposed by (Chiu, 2013) based on a nonparametric classification procedure (Chiu & Douglas, 2013). This method aims to minimize the residual sum of squares (RSS) between the observed and ideal responses among all the possible q-vectors given a Q-matrix. The RSS is used to identify any misspecified q-entries for an item. In the algorithm, the item vector with the highest RSS gets replaced by the one having the lowest RSS. The process is repeated iteratively until the convergence criterion is met. Due to its nature as a nonparametric method, it neither relies on the estimation of model parameters nor makes any assumptions other than those made by the CDM itself (Chiu, 2013). However, if the underlying model is known, parametric methods should provide more powerful results particularly when *N* is large.

DeCarlo (2011) introduced a model-based approach using a Bayesian extension of the DINA model. In this method, possible misspecified entries in the Q-matrix were identified in advance. Then, these entries were treated as random (Bernoulli) variables and estimated with the rest of the model parameters. Limitations of this method are that it is computationally time-consuming and any misspecified q-entries have to be identified in advance. Unlike DeCarlo (2011)'s study, Liu et al. (2012) proposed a data-driven approach in that any expert involvement in Q-matrix design is not required for identifying misspecified entries in the Q-matrix. However, when unknown guessing parameters exist, the identifiability of the Q-matrix can be difficult.

Recently, de la Torre and Chiu (2016) developed a discrimination index, as an extension of the empirically based δ-method (de la Torre, 2008), using the G-DINA model. This index can be applied under a wider class of CDMs. However, the findings of the study were limited to the fixed sample size and test length. Moreover, the index does not determine optimal ε values that prevent q-entries from over- or under-specifications, and the procedure is not iterative, meaning that it stops further identifying attribute specifications after the first round of validation step.

The purpose of this current study is to introduce an iterative procedure in conjunction with a modified version of the SSA, and is called *iterative modified* SSA (IMSSA). The new method aims to make three crucial contributions to the Q-matrix validation literature. First, using simulation, an approximation was made to generally define an empirically based a cut-off value applicable across all conditions. Second, the search algorithm only focuses on single-attribute specifications so that it can eliminate additional complications that could happen due to q-vectors with more than single-attribute specifications. Third, the algorithm is implemented iteratively, such that, if any q-vectors are changed in the previous iteration, a new calibration is carried out using the updated Q-matrix as the provisional Q-matrix. The iterative algorithm aims to alleviate negative effects of any misspecified attribute specifications given in the preceding iteration. In this present study, iterative and non-iterative algorithms were compared to examine if an iterative algorithm can further identify and correct misspecifications in succeeding iterations.

Given the purpose, the rest of the paper consists of the following sections: First is a brief background on the DINA model, Q-matrix refinement method, and exhaustive and sequential search algorithms. Second is a presentation of the new method proposed in this paper. This is followed by simulation study design and results. Then, real data analysis and its results are introduced. Finally, the paper concludes with a discussion and conclusion for future studies.

2. BACKGROUND

2.1. The DINA Model

The DINA model has been commonly used in many studies (e.g., de la Torre & Douglas, 2004, 2008; de la Torre, 2009a; DeCarlo, 2011; Kuo, Pai, & de la Torre, 2016; Liu, Ying, & Zhang, 2015; Park & Lee, 2014; Rupp & Templin, 2008). This study focuses on the DINA model because of its more straightforward interpretations, smaller sample size requirements for accurate parameter estimation (Rojas, de la Torre, & Olea, 2012), and flexibility for extension to more general cognitive diagnostic models. The DINA model is an example of a conjunctive model for dichotomously scored test items, where all required attributes of an item should be mastered by examinees before an examinee can be expected to correctly answer the item. Nonmastery of one or more required attributes for an item is equivalent to nonmastery of all required attributes. Let examinee *i*'s binary attribute vector be denoted by $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$. The item response function of the model is defined as:

$$P\left(X_{ij} = 1 \middle| \alpha_i\right) = \left(1 - s_j\right)^{\eta_{ij}} {g_j}^{(1-\eta_{ij})}, \tag{1}$$

which is the probability of answering an item *j* correctly by examinees with the attribute pattern $\boldsymbol{\alpha}_i$, X_{ij} is the response of examinee *i* (*i* = 1, 2, …, *N*) to item *j* (*j* = 1, 2, …, *J*), and η_{ij} is the ideal response computed as:

$$\eta_{ij} = \prod_{k=1}^{K} {\alpha_{ik}}^{q_{jk}}, \tag{2}$$

an indicator of whether all of the required attributes associated with item *j* have been mastered by examinee *i*.

2.2. Q-Matrix Refinement Method

The RSS of item *j* across all examinees is defined as:

$$RRS_j = \sum_{i=1}^{N} \left(X_{ij} - \eta_{ij}\right)^2 = \sum_{m=1}^{2^K} \sum_{i \in C_m} \left(X_{ij} - \eta_{jm}\right)^2 , \tag{3}$$

where X_{ij} and η_{ij} are the observed and ideal item responses of examinee *i* to item *j*, respectively, C_m is the latent proficiency class *m*, and *N* is the number of examinees. Note that the index j of η_{ij} in Equation 3 was replaced by *m* because ideal item responses are class-specific, meaning that every examinee in the same latent class is assumed to have the same ideal response to an item (Chiu, 2013).

2.3. Exhaustive Search Algorithm

The *exhaustive search algorithm* (ESA) for Q-matrix validation computes δ_{jl} for all $l = 2^K - 1$ possible q-vectors for each *j* item (de la Torre, 2008). The q-vector that gives the largest difference in the probabilities of correct response between examinees who have all the required attributes ($\eta_{jl} = 1$) and those who do not have ($\eta_{jl} = 0$) among all the possible attribute patterns is chosen as the correct q-vector for item *j*. However, the algorithm becomes impractical when *K* is reasonably large. Additionally, the *ESA* has the tendency to choose over-specified q-vectors (de la Torre, 2008).

2.4. Sequential Search Algorithm

The *sequential search algorithm* (SSA), in comparison to the ESA, is considered more efficient because it does not require the comparisons of δ_{jl} for all the possible attribute patterns. More specifically, δ_{jl} is computed for $(K_j + 1)K - (K_j^2 + K_j)/2$ q-vectors for item *j*, where K_j is the number of attributes required for item *j* (de la Torre, 2008).

The SSA starts by comparing δ_{jl}^1 of single-attribute q-vectors with the superscript (1) referring to single-attribute q-vectors. Let δ_j^1 be the largest of δ_{jl}^1 from single-attribute q-vectors, and assume that this is due to α_1. The process continues by examining δ_{jl} of two-attribute q-vectors, δ_j^2, where α_1 is one of the required attributes. If $\delta_{jl}^2 > \delta_{jl}^1$, the single-attribute q-vector is replaced by a two-attribute q-vector. However, if $\delta_{jl}^1 > \delta_{jl}^2$, the process is terminated choosing α_1 as the correct attribute specification for the q-vector. Otherwise, the process continues with such comparisons until a *K*-attribute q-vector is chosen as long as the difference of succeeding δ_{jl} values (i.e., $\hat{\delta}_j^{(K_j+1)} - \hat{\delta}_j^{(K_j)}$) is larger than a predetermined cut-off value.

As stated earlier, estimation that involves some misspecified q-vectors can affect the quality of parameter estimation (Rupp & Templin, 2008) and this in turn affects the accuracy of the validation method. Similarly, the noise due to the stochastic nature of the response process makes it possible to obtain a q-vector with more attribute specifications than necessary. Especially using real data can cause $\hat{\delta}_j^{(K_j+1)} > \hat{\delta}_j^{(K_j)}$ or the reverse, resulting in over- or under-specifications, respectively. A suggested solution is to assign ε, which is a minimum increment in the discrimination index of the item before an additional attribute can be included, as in, $\hat{\delta}_j^{(K_j+1)} - \hat{\delta}_j^{(K_j)} > \varepsilon$ (de la Torre, 2008).

3. THE PROPOSED METHOD

3.1. An Iterative Method for Empirically-Based Q-Matrix Validation

This study introduces an iterative procedure in conjunction with a modified version of the SSA, and is called *iterative modified* SSA (IMSSA). The IMSSA differs from the SSA in two respects. First, the IMSSA determines required attribute specifications based on only the single-attribute q-vectors. Similar to the empirically based δ-method (de la Torre, 2008), the IMSSA starts by estimating the item parameters via an empirical Bayesian implementation of the expected-maximization (EM) algorithm (de la Torre, 2009b) using a provisional Q-matrix. The *K* numbers of $\hat{\delta}$s corresponding to the single-attribute q-vectors (i.e., δ_j^1) are then estimated and ordered from the highest to the lowest. The correct attribute specification is determined based on the proportion of $\hat{\delta}^1_{jl^*}$ relative to the maximum $\hat{\delta}^1_{j(\max)}$ (i.e., $\hat{\delta}^1_{jl^*}/\hat{\delta}^1_{j(\max)}$, for $l^* = 1,2,\dots,K$) for item j. $\hat{\delta}^1_{j(\max)}$ is $\hat{\delta}^1_{jl^*=1}$ because it corresponds to the best suggested attribute specification. The noise due to the use of the estimated posterior distribution should be controlled so as to not cause any over- or under-specifications. That can be done by using a cut-off point denoted by $\varepsilon^{(1)}$, which represent the minimum ratio between single-attribute q-vectors and the best single-attribute q-vector corresponding to $\hat{\delta}^1_{j(\max)}$. Specifically, if $\hat{\delta}^1_{j2}$ is considerably smaller than $\hat{\delta}^1_{j(\max)}$ (i.e., $\hat{\delta}^1_{j2}/\hat{\delta}^1_{j(\max)} < \varepsilon^{(1)}$, the required attribute would be an attribute specified in the single-attribute q-vector corresponding to $\hat{\delta}^1_{j(\max)}$; if not, the attribute specifications in the first two q-vectors are chosen. It continues by checking the ratio $\hat{\delta}^1_{j3}/\hat{\delta}^1_{j(\max)}$. If the ratio is larger than $\varepsilon^{(1)}$, the attribute specification in the third q-vector is also added on the top of the previous two specifications, and it continues; otherwise, the process is terminated. The ratio between $\hat{\delta}^1_{jl^*}$ and $\hat{\delta}^1_{j(\max)}$ was determined based on some preliminary findings, and the values of $\varepsilon^{(1)}$, the cut-off point, were defined using simulated response data.

At this point, an example can be helpful to lay out the rationale as to how the study determines the correctness of attribute specifications based on the ratio of $\hat{\delta}$s to the maximum $\hat{\delta}$. For illustration purposes, we considered two items, each with a misspecified attribute specification. In practice, the provisional Q-matrix may not have entirely correct specifications. However, data based on parameter estimates using the provisional Q-matrix can be generated. The $\hat{\delta}$-computation for the simulated data can be monitored, which can allow us to define extreme changes in the ratio of $\hat{\delta}$s.

Examples of $\hat{\delta}^1_{jl^*}$ computations for the simulated data can help determine whether or not the algorithm could identify correct specifications. Assume that $K = 5$. Table 1 displays examples of items that have over- and under-specifications. In the first misspecification, the q-vector $(1,0,0,0,0)'$ is over-specified as in $(1,0,1,0,0)'$. The EM estimation is carried out with the latter q-vector, and $\hat{\delta}$s of single-attribute q-vectors are estimated and sorted from the highest to the lowest. The result suggests that the correct attribute specification is only α_1 ($\hat{\delta}^1_{j(\max)} = .41$) due to a large drop in $\hat{\delta}^1_{j2}$ (i.e., $\hat{\delta}^1_{j2}\ /\hat{\delta}^1_{j(\max)} = .15 < \varepsilon^{(1)}$), in that a value of $\varepsilon^{(1)}$ will be determined later. A similar result is also observed for an item that has been under-specified. The misspecification appears as $(1,0,0,0,0)'$ from the correct vector of $(1,1,0,0,0)'$ in the right-hand side of Table 1. The ratio of the second $\hat{\delta}^1_{j2}$ to the maximum $\hat{\delta}^1_{j(\max)}$ shows a small drop (i.e., $\hat{\delta}^1_{j2}/\hat{\delta}^1_{j(\max)} = .73 >\ \varepsilon^{(1)}$); however, the next ratio is rather small (i.e., $\hat{\delta}^1_{j3}/\hat{\delta}^1_{j(\max)} = .13 < \varepsilon^{(1)}$). Therefore, the attributes in the first two single-attribute q-vectors are accurately specified (i.e., α_1 and α_2). Note that the criterion is similar to the method proposed by de la Torre and Chiu (2016), which is the proportion of variance accounted for (PVAF) by a particular q-vector relative to the maximum $\hat{\delta}^2$ that is achieved when all the attributes are specified (i.e., $(1,1,1,1,1)'$). However, the criterion in this study is not exactly the same, because it is relative to the best attribute specification, not the attribute vector with all the attributes specified.

Second, the IMSSA becomes more efficient than the original SSA because $\hat{\delta}$ is not computed beyond single-attribute vectors. As such, the maximum number of comparisons for the new algorithm is K, which is considerably smaller than SSA (i.e., $(K_j + 1)K - (K_j^2 + K_j)/2$) and ESA (i.e., $2^K - 1$), where K is the total number of attributes and K_j is the number of attributes being measured by item j. For example, let $K = 10$ and $K_j = 3$. The maximum number of comparisons is 10 for the IMSSA, 34 for the SSA, and 1023 for the ESA. Thus, using the IMSSA can lessen complications associated with multiple search steps. In summary, examining the proportion of $\hat{\delta}$ by a particular single-attribute q-vector to the maximum $\hat{\delta}$ using a provisional q-vector could suggest which attributes should be specified -- $\hat{\delta}$ of required attributes are considerably larger compared to $\hat{\delta}$ of other attributes.

Table 1. Examples for Over- and Under-Specifications

	$(1,0,0,0,0)' \rightarrow (1,0,1,0,0)'$								$(1,1,0,0,0)' \rightarrow (1,0,0,0,0)'$							
l^*	α_1	α_2	α_3	α_4	α_5	$\hat{\delta}_{jl^*}^1$	$\hat{\delta}_{jl^*}^1 / \hat{\delta}_{j(max)}^1$		α_1	α_2	α_3	α_4	α_5	$\hat{\delta}_{jl^*}^1$	$\hat{\delta}_{jl^*}^1 / \hat{\delta}_{j(max)}^1$	
1	**1**	0	0	0	0	.41	**1.00**	√	**1**	0	0	0	0	.40	**1.00**	√
2	0	0	1	0	0	.06	0.15		0	**1**	0	0	0	.29	**0.73**	√
3	0	0	0	0	1	.04	0.10		0	0	0	0	1	.05	0.13	
4	0	1	0	0	0	.04	0.10		0	0	1	0	0	-.01	-0.03	
5	0	0	0	1	0	-.00	0.00		0	0	0	1	0	-.03	-0.08	

Note. The symbol √ displays the chosen attributes based on the associated δ-ratio. $(1,0,0,0,0)' \rightarrow (1,0,1,0,0)'$: $(1,0,0,0,0)'$ is over-specified as in $(1,0,1,0,0)'$. $(1,1,0,0,0)' \rightarrow (1,0,0,0,0)'$: $(1,1,0,0,0)'$ is under specified as in $(1,0,0,0,0)'$. Negative values in the ratio come from the negative $\hat{\delta}$. For example, .52 and .49 for the slip and guessing parameters, respectively, $\hat{\delta}_{jl^*=4}^1 = 1 - s_{jl^*=4} - g_{jl^*=4} = 1 - .52 - .49 = -.03$.

4. SIMULATION STUDY DESIGN

To evaluate the viability of the proposed method, two simulation studies were conducted with the following goals: (1) to determine an optimal $\varepsilon^{(1)}$ value, which could be generalized across the conditions; and (2) to compare the effectiveness of different validation methods with an iterative and noniterative algorithm. For each simulation condition, 100 datasets were replicated using the DINA model with the following factors: sample sizes (N = 1,000 and 2,000), test lengths (J = 15 and 30), item qualities ($s_j = g_j$ = 0.1, 0.2, and 0.3), and amount of misspecifications (5% and 10%). In this study, the three sets of item qualities were considered similar to Hou, de la Torre, and Nandakumar (2014). In each condition, 100 misspecified Q-matrices were generated, which contain 5% or 10% randomly misspecified q-entries. Two constraints were imposed on altering the q-vectors, namely, the misspecified q-vectors cannot have more than two-attribute misspecifications, and at least one attribute should be specified as 1. For example, if a Q-matrix has 10% misspecifications for J = 30 and K = 5, 15 of 150 entries were randomly altered by producing over- or under-specified q-vectors, where almost eight to 15 q-vectors are misspecified. In doing so, the study was able to focus on the impact of the amount of misspecifications rather than the type of misspecifications. It should be noted that the true Q-matrices in Table 2 for J = 15 and 30 are related in two ways. Each attribute is measured six and 12 times when J = 15 and 30, respectively, and there are equal numbers of 1-, 2-, and 3-attribute q-vectors in the each Q-matrix. Finally, the attribute profiles were generated from a uniform distribution in that all the possible attribute patterns were generated with equal probabilities from a multinomial distribution.

To define an optimal $\varepsilon^{(1)}$ value for the IMSSA, the item quality was generated from *Unif*(0.05,0.45). Based on the results of a pilot study, the performance of the proposed method was examined given $\varepsilon^{(1)}$ values in the range 0.10 to 0.90, with an increment of 0.1. After defining an optimal ε value, the second simulation study was conducted to compare the five validation procedures: IMSSA, MSSA, ESA, SSA, and QRM. These methods were compared based on the proportions of correctly identifying attribute specifications at the vector level. The code to implement the IMSSA, MSSA, ESA, and SSA was written in Ox (Doornik, 2009), whereas, the NPCD R package (Zheng & Chiu, 2015) was used (R Core Team, 2014) for the QRM analyses.

Table 2. True Q-Matrix for the Simulated Data

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1*	1	0	0	0	0	11*	1	1	0	0	0	21*	1	1	1	0	0
2*	0	1	0	0	0	12*	1	0	1	0	0	22*	1	1	0	1	0
3*	0	0	1	0	0	13	1	0	0	1	0	23	1	1	0	0	1
4*	0	0	0	1	0	14	1	0	0	0	1	24	1	0	1	1	0
5*	0	0	0	0	1	15	0	1	1	0	0	25*	1	0	1	0	1
6	1	0	0	0	0	16	0	1	0	1	0	26	1	0	0	1	1
7	0	1	0	0	0	17*	0	1	0	0	1	27	0	1	1	1	0
8	0	0	1	0	0	18*	0	0	1	1	0	28	0	1	1	0	1
9	0	0	0	1	0	19	0	0	1	0	1	29*	0	1	0	1	1
10	0	0	0	0	1	20*	0	0	0	1	1	30*	0	0	1	1	1

Note. Items with * are used for J = 15.

5. FINDINGS

5.1. Simulation Study I

In the first simulation study, the performance of the IMSSA was observed to define an optimal $\varepsilon^{(1)}$ value which can be used under all conditions. Focusing in the range 0.10 to 0.90, values were derived based on the highest proportions of correctly identifying attribute specifications on average throughout all conditions, as shown in Table 3. When $\varepsilon^{(1)}$ = 0.50 and 0.60, 92% of the q-vectors were correctly identified on average, which is the highest proportions of recovery across all $\varepsilon^{(1)}$ values. Thus, $\varepsilon^{(1)}$ was set at 0.50 in the second simulation study.

Table 3. Proportions of Recovery for Various Cut-off Values

N	*J*	%	ε								
			0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
1,000	15	5	0.01	0.18	0.66	0.84	0.94	**0.96**	0.94	0.90	0.78
		10	0.06	0.38	0.63	0.79	**0.87**	0.85	0.80	0.67	0.41
	30	5	0.33	0.86	0.96	0.98	**0.99**	0.97	0.93	0.83	0.51
		10	0.12	0.60	0.87	0.93	**0.98**	0.95	0.87	0.73	0.47
2,000	15	5	0.23	0.64	0.80	0.86	0.88	**0.91**	0.88	0.78	0.44
		10	0.11	0.39	0.55	0.68	0.74	**0.77**	0.75	0.62	0.39
	30	5	0.38	0.89	0.96	**0.98**	**0.98**	0.97	0.93	0.86	0.64
		10	0.13	0.68	0.91	0.95	**0.97**	0.96	0.90	0.80	0.54
Average			0.17	0.58	0.79	0.88	**0.92**	**0.92**	0.87	0.77	0.52

Note. Numbers in bold are the highest proportions of recovery for each condition.

5.2. Simulation Study II

Table 4 shows results reported at the vector level, which are divided into two, with and without iterative algorithms. Among the methods with a non-iterative algorithm, the MSSA outperformed the others for each simulation condition considered in this study. In addition, the SSA (76%) provided better recovery than the ESA (74%) on average across all the conditions. As noted in de la Torre (2008), the SSA procedure was originally proposed to be a more efficient algorithm that does not require computing δ_{jl} for the $2^K - 1$ possible q-vectors; however, results based on the ESA and SSA did not show considerable differences, which was only 2% on average. In general across the conditions, the average recovery of the MSSA was 89% of the q-vectors; whereas, the average recoveries of the ESA and SSA were 74% and 76%, respectively. In particular, recovery based on the MSSA was 15% and 13% higher than that of the ESA and SSA, respectively. Therefore, even though an iterative algorithm was not implemented in the MSSA, we could state that the modified version of the SSA (MSSA) improved the recovery in comparison to the SSA.

Specifically in comparing the iterative methods (i.e., IMSSA and QRM) under the high quality item, the QRM worked usually equally well as or better than the IMSSA. In this quality of items, both methods had perfect or above 97% of recovery. In continuing the comparison of the IMSSA and QRM under the medium quality item, both methods had recovery of attribute specifications above 99% when J = 30. The lowest recovery was 69% for the QRM and 86% for the IMSSA. When data were generated from the low quality item, the IMSSA (81%) had 9% more recovery than the QRM (72%) on average. The QRM only outperformed the IMSSA under four conditions, where the proportions of recovery differed only by 1% to 2% when the item qualities were medium (i.e., N = 1,000, J = 30 with 5% and 10% misspecifications) and high (i.e., N = 1,000, J = 15 with 10% misspecifications and N = 2,000, J = 15 with 10% misspecifications), respectively. Other than these differences, the IMSSA provided a better overall recovery than the QRM.

It is interesting to report that the performance of the QRM was equally well or worse when the sample size was doubled. For example, when the item quality was low under a condition where N = 1,000, J = 30 with 5% misspecifications, doubling the sample size to N = 2,000 resulted in the recovery dropping from 85% to 83%. In contrast, considering the same conditions, the recovery improved from 70% to 77% for the ESA, from 73% to 79% for SSA, from 88% to 92% for MSSA, and from 89% to 93% for the IMSSA. However, doubling the test items from 15 to 30, the recovery increased for all the methods. This finding can indicate that doubling the test length can lead to better improvement in recovery more than doubling the sample size.

Similarly, with regards to the difference in recovery rates due to the amount of misspecifications within the same conditions (i.e., N and J), a larger test length provided a smaller gap than a larger sample size. That is, recovery differences between 5% and 10% misspecifications were higher with a larger sample size than a longer length test. For example, among the non-iterative methods when N = 1,000 and J = 15 under the high quality item, recovery differences between 5% and 10% misspecifications were 22%, 21%, and 9% for the SSA, ESA, and MSSA, respectively, which dropped to 9%, 8%, and 0% when J = 30 holding the sample size constant. However, doubling the sample size with a fixed test length did not change the recovery differences that much, which was only 20%, 19%, and 9% for the SSA, ESA, and MSSA, respectively. In taking the amount of misspecifications into account for the non-iterative methods, doubling the test length had a considerably positive impact on the recovery than doubling the sample size.

For the iterative methods, again, doubling the test length decreased the difference in recovery rates between 5% and 10% misspecified Q-matrices. Under the same conditions, when N = 1,000 and J = 15, it was 1% for the QRM (i.e., 100 - 99 = 1) and 3% for the IMSSA (i.e., 100 - 97 = 13). However, that gap was smaller when J = 30 than N = 2,000. The difference substantially dropped for both methods after doubling the test length with a constant sample size. Therefore, based on these findings, it can be stated that doubling the test length substantially improved the recovery for both iterative methods and decreased the recovery differences due to a different amount of misspecifications.

Table 4. Proportions of Recovery for Misspecification in the Q-Matrix

				Non-Iterative			Iterative	
Quality	*N*	*J*	*%*	ESA	SSA	MSSA	QRM	IMSSA
H	1,000	15	5	0.83	0.85	0.99	1.00	1.00
			10	0.61	0.64	0.90	0.99	0.97
		30	5	0.94	0.96	1.00	1.00	1.00
			10	0.85	0.88	1.00	1.00	1.00
	2,000	15	5	0.86	0.87	1.00	1.00	1.00
			10	0.66	0.68	0.91	0.99	0.97
		30	5	0.98	0.99	1.00	1.00	1.00
			10	0.91	0.93	0.99	1.00	1.00
M	1,000	15	5	0.80	0.82	0.94	0.90	0.96
			10	0.60	0.60	0.79	0.70	0.86
		30	5	0.78	0.82	0.99	1.00	0.99
			10	0.64	0.68	0.97	1.00	0.99
	2,000	15	5	0.83	0.84	0.95	0.90	0.97
			10	0.64	0.64	0.80	0.69	0.89
		30	5	0.85	0.89	1.00	1.00	1.00
			10	0.69	0.74	0.98	1.00	1.00
L	1,000	15	5	0.69	0.71	0.82	0.81	0.82
			10	0.51	0.51	0.64	0.61	0.64
		30	5	0.70	0.73	0.88	0.85	0.89
			10	0.56	0.59	0.74	0.65	0.81
	2,000	15	5	0.80	0.81	0.84	0.81	0.85
			10	0.59	0.59	0.64	0.61	0.67
		30	5	0.77	0.79	0.92	0.83	0.93
			10	0.61	0.62	0.77	0.62	0.87
	Average			0.74	0.76	0.89	0.87	0.92

Note. ESA: exhaustive search algorithm, *SSA*: sequential search algorithm with $\varepsilon = .01$, *MSSA*: non-iterative modified sequential search algorithm, *QRM*: Q-matrix refinement method with an iterative algorithm, *IMSSA*: iterative modified sequential search algorithm, H: high quality, M: medium quality, L: low quality, *N*: sample size, *J*: test length, *%*: amount of misspecification.

In summary, the proposed MSSA and IMSSA worked much better than the other methods. That is, after averaging the proportions of recovery across the conditions (i.e., *N*, *J*, item qualities, and amount of misspecifications), recovery based on the IMSSA (92%) and MSSA (89%) was 5% and 2% higher than that of the QRM (87%), respectively, and rather larger than the ESA and SSA. Note that the number of iterations in the iterative procedures was usually between two and three, and did not go beyond four.

6. REAL DATA ANALYSIS

6.1. Data

In addition to the simulation study, real data were analyzed to investigate the applicability of the method. The fraction-subtraction data (Tatsuoka, 1984) with 536 middle school students' responses to 12 fraction subtraction problems were examined. The four attributes for this dataset are: (a) performing a basic fraction subtraction operation, (b) simplifying/reducing, (c) separating a whole number from fraction, and (d) borrowing one from a whole number to fraction. The 12 items with the corresponding attribute specifications and $\hat{\delta}$ values are shown in Table 5.

Table 5. Q-Matrix for Fraction-Subtraction Items

		Attribute				
Item		α_1	α_2	α_3	α_4	$\hat{\delta}$
1	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0	0.72
2	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0.66
3	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	0.83
4	$3\frac{7}{8}-2$	1	0	1	0	0.42
5	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	0.74
6	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1	0.86
7	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	0.80
8	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	0.86
9	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	0.80
10	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1	0.84
11	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	0.71
12	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	0.82

Note. α_1 - performing a basic fraction subtraction operation; α_2 - simplifying/reducing; α_3 - separating a whole number from fraction; and α_4 - borrowing one from a whole number to fraction.

Note that the data set of Tatsuoka (1984) has been one of the most commonly examined real data designed for cognitively diagnostic assessment (Chiu, 2013; Chiu & Köhn, 2015; de la Torre, 2008; de la Torre & Chiu, 2016; DeCarlo, 2011). In CDM analyses, one of the main concerns is the completeness of the Q-matrix. Unfortunately, the fraction-subtraction data do not appear to have a complete Q-matrix. It was demonstrated by Chiu, Douglas, and Li (2009) that a complete Q-matrix should identify all possible attribute patterns and require each attribute to be represented by at least one single-attribute vector. This issue has been further discussed with the original data (see Table 4 on pp. 615, Chiu, 2013; DeCarlo, 2011) or subsets of it (see de la Torre, 2008; de la Torre & Chiu, 2016). The incompleteness of the Q-matrix in this dataset occurs because of the fact that only 58 of 256 (*K* = 8; Chiu, 2013) and 10 of 32 (*K* = 5; Chiu & Köhn, 2015) possible attribute patterns can be identified by the items, meaning that multiple classes may be merged (Chiu, 2013). Therefore, results of this data analysis should be interpreted with caution.

6.2. Results

For the IMSSA, $\hat{\delta}^{1}_{jl^*}$ statistic and $\hat{\delta}^{1}_{jl^*}/\hat{\delta}^{1}_{j(max)}$ ratios for 12 items are reported in Table 6, and the suggested Q-matrix is further shown in Table 7. Given the results in the first simulation study, the $\varepsilon^{(1)}$ values were set at 0.50 and 0.60.

Table 6. Suggested Single-Attribute Specifications with $\hat{\delta}$-values for the Fraction-Subtraction Test

Item	α_1	α_2	α_3	α_4	$\hat{\delta}^1_{jl^*}$	$\hat{\delta}^1_{jl^*}/\hat{\delta}^1_{j(max)}$	Item	α_1	α_2	α_3	α_4	$\hat{\delta}^1_{jl^*}$	$\hat{\delta}^1_{jl^*}/\hat{\delta}^1_{j(max)}$
1	1*	0	0	0	0.72	1.00	7	1*	0	0	0	0.73	1.00
	0	0	1*	0	0.45	0.63		0	1*	0	0	0.71	0.97
	0	1*	0	0	0.40	0.56		0	0	1*	0	0.56	0.77
	0	0	0	1	0.34	0.47		0	0	0	1	0.15	0.21
2	0	0	0	1*	0.55	1.00	8	1*	0	0	0	0.82	1.00
	1*	0	0	0	0.34	0.62		0	0	1*	0	0.75	0.91
	0	1*	0	0	0.30	0.55		0	1*	0	0	0.51	0.62
	0	0	1*	0	0.30	0.55		0	0	0	1	0.13	0.16
3	1*	0	0	0	0.83	1.00	9	1*	0	0	0	0.75	1.00
	0	0	1*	0	0.45	0.54		0	0	1*	0	0.71	0.95
	0	1	0	0	0.37	0.45		0	1*	0	0	0.49	0.65
	0	0	0	1	0.07	0.08		0	0	0	1	0.15	0.20
4	1*	0	0	0	0.39	1.00	10	0	0	0	1*	0.66	1.00
	0	0	1*	0	0.37	0.95		1*	0	0	0	0.52	0.79
	0	1*	0	0	0.26	0.67		0	0	1*	0	0.49	0.74
	0	0	0	1	0.08	0.21		0	1*	0	0	0.46	0.70
5	0	0	0	1*	0.57	1.00	11	1*	0	0	0	0.56	1.00
	1*	0	0	0	0.47	0.82		0	0	0	1*	0.51	0.91
	0	1*	0	0	0.42	0.74		0	0	1*	0	0.50	0.89
	0	0	1*	0	0.41	0.72		0	1*	0	0	0.48	0.86
6	0	0	1*	0	0.67	1.00	12	0	0	0	1*	0.64	1.00
	1*	0	0	0	0.53	0.79		1*	0	0	0	0.48	0.75
	0	1*	0	0	0.51	0.76		0	1*	0	0	0.47	0.73
	0	0	1*	0	0.49	0.73		0	0	1*	0	0.44	0.69

Note. * indicates a suggested attribute specification, $\varepsilon^{(1)}$ = 0.50.

Table 7. Suggested Q-Matrix by the IMSSA and QRM for the Fraction-Subtraction Test

	IMSSA ($\varepsilon^{(1)}$ = 0.50)				IMSSA ($\varepsilon^{(1)}$ = 0.60)				QRM			
Item	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1	1	1*	1*	0	1	1*	0	0	1	0	0	1*
2	1	1	1	1	1	0*	0*	0*	1	1	1	1
3	1	1*	0	0	1	0	0	0	1	0	0	0
4	1	1*	1	0	1	1*	1	0	1	0	1	0
5	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1*	0	1	1	1*	0	1	1	0	0
8	1	1*	1	0	1	1*	1	0	1	0	1	0
9	1	1*	1	0	1	1*	1	0	1	0	1	0
10	1	1*	1	1	1	1*	1	1	1	1*	1	1
11	1	1	1	1	1	1	1	1	1	0*	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1

Note. α_1 - performing a basic fraction subtraction operation; α_2 - simplifying/reducing; α_3 - separating a whole number from fraction; and α_4 - borrowing one from a whole number to fraction; * indicates a modified attribute specification.

The results of the fraction-subtraction data obtained from the IMSSA were compared to the QRM. The IMSSA suggested attribute changes in seven items (i.e., items 1, 3, 4, 7, 8, 9, and 10) when $\varepsilon^{(1)} = 0.50$; whereas, the QRM suggested attribute changes in three items (i.e., items 1, 10, and 11). Based on the IMSSA, the result indicated that item 1 (i.e., $\frac{3}{4} - \frac{3}{8}$) should require two more attributes (i.e., α_2 and α_3) in addition to α_1. This suggestion may have occurred because this item requires more than just α_1, performing a basic fraction subtraction problem. Another suggestion was for item 3 (i.e., $\frac{6}{7} - \frac{4}{7}$), where α_2 was deemed required. Items 4 (i.e., $3\frac{7}{8} - 2$), 8 (i.e., $3\frac{4}{5} - 3\frac{2}{5}$), 9 (i.e., $4\frac{5}{7} - 1\frac{4}{7}$), and 10 (i.e., $7\frac{3}{5} - \frac{4}{5}$) required α_2 in addition to $\alpha_1$and α_3. Note that another strategy for solving the problem in one of these four items – borrowing one from a whole number to fraction, performing a basic fraction, and simplifying/reducing – happens to give the correct answer. The following example shows another strategy to solve item 9:

$$4\frac{5}{7} - 1\frac{4}{7} = \frac{(4 \times 7) + 5}{7} - \frac{(1 \times 7) + 4}{7}$$

$$= \frac{33 - 11}{7} = \frac{22}{7} = 3\frac{1}{7}.$$

Another attribute suggestion (i.e., α_3) was for item 7 (i.e., $\frac{11}{8} - \frac{1}{8}$) on the top of α_1 and α_2. Similar to the preceding example, a different strategy – separating a whole number from fraction, performing a basic fraction subtraction operation, and simplifying/reducing – could also give the correct answer to item 7, as in,

$$\frac{11}{8} - \frac{1}{8} = 1\frac{3}{8} - \frac{1}{8} = 1\frac{3-1}{8}$$

$$= 1\frac{2}{8} = 1\frac{1}{4}.$$

In applying the QRM, Chiu (2013) found that item 4, which appears as item 2 in this study, did not require the possession of α_3 to be correctly answered. In contrast, the QRM in this study suggested that α_3 was necessary. An explanation could be because of the fact that Chiu (2013) used 20 items with 8 attributes. Whereas, the IMSSA indicated that the mastery of the third attribute was required to answer item 2 correctly. The QRM also suggested to include and exclude α_2 in items 10 and 11, respectively.

As demonstrated by the examples, a deeper analysis is needed. The IMSSA has more 1s than the QRM that can be controlled by adjusting the cut-offs. The cut-off values defined in the simulation study do not perfectly fit to the real data analysis in this case because it did not have a complete Q-matrix. The latter values were just approximations based on the conditions defined in the simulation study. Further discussions about multiple strategies in cognitive diagnosis using the fraction subtraction data can be found in de la Torre and Douglas (2008), Hou and de la Torre (2014), and Mislevy (1996). Other reasons could be because the fraction subtraction data have fewer number of items and attributes than the simulation study. Also note that when $\varepsilon^{(1)}$ was set at 0.60, three items presented different attribute specifications (i.e., items 1, 2, and 3). α_3 in item 1, and α_2, α_3, and α_4 in item 2 were altered to 0s; however attribute specifications in item 3 was consistent with the Q-matrix given for the data.

7. DISCUSSION AND CONCLUSION

CDMs aim to classify the attribute mastery or nonmastery of examinees, and the Q-matrix is needed for specifying required attributes for each item in a test. The importance of revising attribute specifications in the Q-matrix should not be underestimated due to the inherent subjectivity of domain experts, consequently resulting in serious validity concerns.

The IMSSA for Q-matrix validation presented in this study aimed to extend the SSA (de la Torre, 2008) in several ways. First, it offered a more efficient solution as it only examines the first K single-attribute q-vectors. Second, in addition to less number of computational requirements, an iterative algorithm was included in the method to decrease negative effects of any misspecified attribute specification given in the previous iteration. And, third, an approximation was made to generally define optimal cut-off values applicable across the specific set of conditions.

In this work, three methods without an iterative algorithm were compared to two methods with an iterative algorithm. Among the noniterative methods, the MSSA reported better results, which had higher recovery than the QRM on average across all the factors. As expected, the results showed that the IMSSA worked much better than the noniterative methods. According to the simulation studies, the IMSSA showed promising improvements in Q-matrix validation that could enhance the estimation of model parameters, model-data fit analyses, and ultimately, the accuracy of attribute-classifications.

Using a 3.50-GHz I7 computer, it took the code the least amount of time to run the validation procedures for MSSA, followed by IMSSA, ESA, SSA, and QRM. For instance, it took 1.64, 3.11, 9.89, 24.35, and 30.00 minutes using MSSA, IMSSA, ESA, SSA, and QRM procedures, respectively, for 100 iterations under the condition in that N = 2,000, J = 30, and medium quality items with 10% misspecifications in the Q-matrix.

This present study had some limitations. For instance, the number of attributes was assumed to be known and fixed to K = 5. It would be interesting to investigate the method by relaxing this assumption. The findings of this study were based on the attribute structure generated from a uniform distribution. The performance of the methods should be investigated under a condition where attributes were generated from a higher order distribution (de la Torre & Douglas, 2004). Also, in addition to the δ-statistic used in this study, other statistics can be carried out for Q-matrix validation. This study should also be extended to make it applicable to a wider class of CDMs such as the G-DINA model (de la Torre, 2011). This will obviate the need to assume the specific CDMs involved. Finally, this method should be applied to other real data sets (e.g., Akbay, Terzi, Kaplan, & Karaaslan, 2018) with a complete Q-matrix so that further insights can be gained on how the proposed method could work in practice.

ORCID

Ragip Terzi https://orcid.org/0000-0003-3976-5054

Jimmy de la Torre https://orcid.org/0000-0002-0893-3863

8. REFERENCES

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K. G. (2018). Expert-based attribute identification and validation: An application of cognitively diagnostic assessment. *Journal on Mathematics Education*, *9*, 103-120.

Chiu, C. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement 37*, 598-618.

Chiu, C., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification 37*, 225-250.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*. *74*, 633-665.

Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement, 39*, 465-479.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35(1)*, 8-26.

de la Torre, J. (2008). An empirically based method of Q-Matrix validation for the DINA model: development and applications. *Journal of Educational Measurement*, *45*, 343-362.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163-183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253-273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive

diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595-624.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. [Computer software]. London, UK: Timberlake Consultants Ltd.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98-125.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464-485.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*, 315-330.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 609-618.

Liu, J., Ying, Z., & Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, *80*, 468-490.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379-416.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, *38*, 376-390.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada.

Rupp, A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*, 78-98.

Tatsuoka, K. K. (1983). Rule-space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems* (Report No. NIE-G-81-0002). Urbana: Computer-based Education Research Laboratory, University of Illinois.

Terzi, R. (2017). New Q-matrix validation procedures (Doctoral dissertation). Retrieved from https://doi.org/doi:10.7282/T3571G5G

Zheng, Y., & Chiu, C.-Y. (2015). NPCD: The R package for nonparametric methods for cognitive diagnosis.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 263–273

DOI: 10.21449/ijate.377138

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

Effects of Various Simulation Conditions on Latent-Trait Estimates: A Simulation Study

Hakan Kogar [1*]

[1] Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Turkey

Abstract: The aim of this simulation study, determine the relationship between true latent scores and estimated latent scores by including various control variables and different statistical models. The study also aimed to compare the statistical models and determine the effects of different distribution types, response formats and sample sizes on latent score estimations. 108 different data bases, comprised of three different distribution types (positively skewed, normal, negatively skewed), three response formats (three-, five- and seven-level likert) and four different sample sizes (100, 250, 500, 1000) were used in the present study. Results show that, distribution types and response formats, in almost all simulations, have significant effect on determination coefficients. When the general performance of the models are evaluated, it can be said that MR and GRM display a better performance than the other models. Particularly in situations when the distribution is either negatively or positively skewed and when the sample size is small, these models display a rather good performance.

ARTICLE HISTORY

Received: 10 January 2018
Revised: 03 March 2018
Accepted: 16 March 2018

KEYWORDS

Item response theory,
Classical test theory,
Factor analysis,
Latent trait scores,
Data simulation

1. INTRODUCTION

In the Classical Test Theory (CTT), known to be the first theory developed to measure latent traits, the fundamental concept is the true score. The true score is defined as the expected value of the observed scores. The expected value expressed in this definition can be obtained by means of an infinite number of repetitions of the independent observations (Lord & Novick, 1968). In other words, if a psychological test is to be administered, the test taker's true score can be obtained by administering the test to the person an infinite number of times. According to this theory, the mathematical representation of which is rather simple, the observed score is obtained by adding the true score and the random error (Mellenberg, 1996). The latent score in CTT refers to the observed scores obtained by adding the item scores (Lord & Novick, 1968).

Item Response Theory (IRT), known to be a modern test theory, was developed based on the argument that it is not realistic to make infinite observations and that repeated measurements are not statistically independent of each other. IRT and CTT are different in

CONTACT: Hakan Kogar ✉ hkogar@gmail.com Akdeniz University Education Faculty Measurement and Assessment Department Antalya Turkey

ISSN-e: 2148-7456 /© IJATE 2018

terms of their theoretical basics and statistical formulations (Borsboom & Mellenbergh, 2002). When both are compared, it is believed that IRT is superior as psychometric traits can be obtained independent of the sample and to which test or item an ability or trait belongs to can be determined from the participants' responses (Crocker & Algina, 1986). IRT models seek to determine the latent traits based on their item stimulators (such as item difficulty and estimate of parameters) and the interaction of the ability. In these models, instead of the total score, the patterns in the responses are focused on. IRT, which is widely used in the fields of education and psychology, has various latent trait models which can be applied to dichotomous or polytomous datasets (Brzezińska, 2016).

While IRT models make use of all the information in the response patterns in order to obtain all the item parameters, factor analysis (FA) techniques estimate the relationships between items and latent traits by means of correlation matrices (Cyr & Davies, 2005). Principal component analysis (PCA), which is considered as the basic method of factor analysis, is a dimension reduction method. It seeks to derive a small number of independent principal components from a larger number of correlated variables (Saporta & Niang, 2009). While latent variables can directly be measured in PCA, in factory analysis, data reduction can only be used for traits that cannot be directly measured (e.g. intelligence, anxiety). A theoretical definition is needed for these traits that cannot be directly measured (Bartholomew, Knott, & Moustaki, 2011). Researchers who seek to determine how many factors have an effect on a variable and which factors have a combined effect utilize exploratory factor analysis (EFA), which is based on an exploratory technique (DeCoster, 1998). When the relationship between the observed and latent variables is revealed, confirmatory factor analysis (CFA) is used. CFA is a measurement model that seeks to estimate the population covariance matrix of the theoretical model based on the observed covariance matrix (Raykoy & Marcoulides, 2000, 95).

Not many studies are encountered in the related literature which comparisons are made between the different parameter estimation methods on these techniques, namely CTT, IRT, and FA (Dumenci & Achenbach, 2008; Hauck Filho, Machado, & Damásio, 2014). In one study, conducted by Dumenci and Achenbach (2008), six statistical models that could estimate different latent traits were compared: CTT, PCA, CFA using maximum likelihood estimation, CFA using weighted least squares, graded response model (GRM) and partial credit model (PCM). CTT, PCA and CFA using the maximum likelihood estimation method yielded similar findings. Likewise, similar findings were observed among the PCA, GRM and CFA using weighted least squares models. In each group of methods, the estimations of the linear relationships (r^2) were found to be close to 1.00. As real data were used in the study, the lack of control variables made it difficult for the models to be compared. In another study, conducted by Hauck Filho et al. (2014), seven different statistical models that could estimate latent traits were compared: CTT, PCA, EFA using Maximum Likelihood, EFA with Minimum Rank, RSM, GRM and CFA with weighted least squares. This comparative study was performed with a total of 15 different simulative datasets comprised of three different item difficulty distributions and five different sample sizes. In each dataset, based on 10 items, true scores of latent traits were obtained. The comparison between the true scores and the estimated trait scores were tested by means of various statistical techniques. It was found that the estimations that were closest to the true scores were those estimations obtained from RSM, GRM and CFA using weighted least squares. These three models are ones that are least affected by inconsistencies among the items and sample distributions. However, the findings of these three models were not found to be statistically significant.

The present simulation study, which took into consideration previous studies, aimed to determine the relationship between true latent scores and estimated latent scores by including various control variables (distribution types and response formats) and different statistical models (unweighted least squares and diagonally weighted least squares). The study also aimed to compare the statistical models and determine the effects of different distribution types, response formats and sample sizes on latent score estimations.

2. METHOD

2.1. Procedures of Data Simulations

Based on three different item difficulty distributions (which is defined below), 108 different data bases, comprised of three different distribution types (positively skewed, normal, negatively skewed), three response formats (three-, five- and seven-level likert) and four different sample sizes (100, 250, 500, 1000) were used in the present study. In these data bases, the discrimination parameter (parameter a) was kept constant between 0.5 and 2.8 owing to the fact that the distribution of the simulative datasets was similar to that of the true datasets. The item responses were produced via the Generalized Partial Credit Model (GPCM). Ability parameters (theta) were calculated for each database. These values were recorded as true latent scores. Total of 20 items were simulated.

Among the three different item difficulty distributions, the first (Situation-1) aimed to include the individuals who were in the lower 20% of the sample distribution, that is between -3.00 and -0.84 in terms of the item difficulty parameter (parameter b). The second item difficulty distribution (Situation-2) was simulated with a standard normal distribution having a mean of 0 and a standard deviation of 1. The third item difficulty distribution (Situation-3) included the individuals in the top 20% of the sample distribution that is between 0.84 and 3.00 in terms of the item difficulty parameter (parameter b). These values were obtained by means of the z-score table. These values are adapted from Hauck Filho, et al. (2014).

Of the three different distribution types, the first was a negatively skewed distribution. Taking into consideration beta distribution, this distribution was produced with an expected skewness of 0.40 and an expected kurtosis of -0.30. For this purpose, in the beta distribution, value a was 5.7 and value b was 2.9. The normal distribution, which is the second distribution type, was mean of 0 and the standard deviation of 1. Taking into consideration beta distribution, the positively skewed distribution, which was the third distribution type, was produced with an expected skewness of 0.40 and an expected kurtosis of -0.30. For this purpose, in the beta distribution, value a was 2.9 and value b was 5.7. These values are adapted from Hauck Filho, et al. (2014).

The difference in the sample size was determined, considering previous simulation studies (Dawber, Rogers, & Carbonaro, 2009; Hauck Filho, et al., 2014). Even though one of the factors affecting the psychometric traits of measurement instruments is the response formats (Jafari, Bagheri, Ayatollahi, & Soltani, 2012), the same number of response formats was used in almost all simulation studies. However, there are simulation studies that seek to determine the most appropriate response format for psychological measurement instruments. The response formats in the present study were determined by taking into consideration the findings of studies in which the most appropriate number of response categories was stated (Lozano, García-Cueto, & Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol, & Coffman, 2009). Data simulation was implemented using the WINGEN program (Han, 2007).

2.2. Data Analysis

In the present study, latent trait score estimates were made by means of the different models stated below:

Classical Test Theory (CTT): In congruence with this theory, for every database, the raw scores (total score) were calculated based on a 20-item test.

Principal Component Analysis (PCA): Component scores were obtained by using this method, which produced weighted scores from indicators (items). Regression scoring method was used for estimate. Factor scores were obtained using the Factor 10.5 program.

Minimum Rank Factor Analysis (MR): This parameter estimation method was developed by Ten Berge and Kiers (1991) with the purpose of explaining the common variance at the highest level. By using the Factor 10.5 program and this parameter estimation method, the polychoric correlation matrix (Lorenzo-Seva & Ferrando, 2006) and the factor scores were determined.

Unweighted Least Squares (ULS): With this method, which can independently make parameter estimations based on distribution types (Kline, 2015, p. 159), a confirmatory factory analysis was conducted. The factor values were obtained via LISREL 8.7.

Diagonally Weighted Least Squares (DWLS): DWLS is a CFA model specifically designed for ordinal data. DWLS does not have any distribution assumptions (Li, 2016). The factor values were obtained via LISREL 8.7.

Graded Response Model (GRM): This model, which is a IRT method used in multiple score scales, such as Likert type scales (Samejima, 1968), was used in combination with estimated a posteriori (EAP) and the R 3.4.2 program and the psych (Revelle, 2017) and Itm (Revelle, 2017) packages to estimate ability parameters.

The Pearson correlation coefficients and determination coefficients (r^2) between the obtained latent trait estimates (scores and indices) and the true latent scores were obtained. In addition, in all the simulation conditions, the factorial ANOVA test was run to test the mean differences and the common variance.

3. FINDINGS

The relationship between six different methods used to estimated latent trait scores and true latent scores in a total of 108 different simulative datasets consisting of three different item difficulty distributions, three different distribution types, three different response formats and four different sample sizes, and the findings regarding determination coefficients are presented in Tables 1, 2 and 3.

In Situation-1, there were huge differences between the correlation and determination coefficients obtained from the negative skewed distribution. Particularly in sample size-1 and response format-1 conditions, zero correlation was found between the true score and the latent trait scores that the models yielded. Nor was zero correlation found for sample size-1 and response format-3. It was found that there was a high correlation between latent trait estimates obtained via a negatively skewed distribution in MR and true scores only in sample size-1 and response format-4, while the relationships in the other simulation conditions were close to zero. The estimations of the other five models yielded moderate or high correlation coefficients in the other simulation conditions. CTT produced a correlation coefficients with the highest average. In the normal distribution in Situation-1, the correlation coefficients in all the simulation conditions were moderate or high. The estimations that the MR model yielded had correlation coefficients with the highest average. In the positively skewed distribution in Situation-1, the correlation coefficients obtained in all the simulation conditions were very high (r>.88). The estimations that GRM yielded had a correlation coefficients with the highest average.

The correlation coefficients obtained in the simulation condition with a negatively skewed distribution (Situation-2), except for the estimations made for sample size-1 and response format-1 via MR model, were found to be very high (r>.90). It was observed that the estimations obtained via the MR model were affected by a negatively skewed distribution, particularly in situations with a small sample size. It was also found that in a simulative database obtained from a normal distribution, it was the MR model estimations that were mostly affected, but all the models yielded estimations with high correlation coefficients. It was found that in positively skewed distributions, the estimations that the DWLS model yielded were affected by small sample sizes. In Situation-2, the higher the response format and sample size were, the higher the correlations and determination coefficients turned out to be. In Situation-2, the estimations that GRM yielded in all conditions had coefficients of relationship with the highest averages.

Table 1. Correlation and determination coefficients for situation-1

D	RF	S	Situation-1											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.016	.000	.006	.000	-.008	.000	-.022	.000	.055	.003	.015	.000
		S2	.753	.567	.681	.463	.635	.403	.687	.472	.695	.483	.676	.457
		S3	.049	.002	.039	.001	.048	.002	.037	.001	.034	.001	.048	.002
		S4	.696	.484	.660	.435	.701	.492	.654	.428	.658	.433	.642	.413
	RF -2	S1	.720	.519	.642	.413	.009	.000	.642	.413	.642	.413	.458	.209
		S2	.707	.499	.645	.415	.024	.001	.645	.415	.572	.328	.687	.472
		S3	.706	.499	.665	.443	-.037	.001	.654	.428	.621	.386	.637	.406
		S4	.692	.479	.617	.381	.040	.002	.615	.378	.569	.324	.734	.539
	RF -3	S1	.699	.488	.675	.455	-.091	.008	.655	.429	.559	.313	.537	.288
		S2	.634	.401	.590	.348	-.051	.003	.559	.312	.451	.203	.695	.482
		S3	.692	.479	.667	.445	-.116	.013	.652	.425	.651	.424	.762	.580
		S4	.717	.513	.674	.454	.042	.002	.666	.444	.629	.396	.778	.605
Δ			.737	.567	.675	.463	.817	.492	.709	.472	.661	.482	.763	.605
Mean			.590	.411	.547	.354	.100	.077	.537	.345	.511	.309	.556	.371
D-2	RF -1	S1	.849	.721	.827	.684	.818	.669	.826	.682	.823	.678	.881	.849
		S2	.801	.641	.768	.589	.813	.661	.764	.584	.755	.570	.727	.529
		S3	.837	.700	.817	.668	.856	.733	.812	.660	.791	.626	.887	.787
		S4	.834	.695	.819	.670	.885	.783	.815	.665	.811	.658	.902	.813
	RF -2	S1	.766	.586	.753	.568	.837	.701	.747	.558	.652	.425	.815	.766
		S2	.784	.615	.729	.532	.874	.763	.711	.506	.715	.512	.829	.687
		S3	.788	.621	.773	.597	.847	.717	.774	.599	.772	.597	.827	.683
		S4	.776	.603	.745	.555	.866	.749	.750	.562	.748	.559	.864	.746
	RF -3	S1	.816	.666	.814	.662	.898	.807	.803	.644	.813	.661	.844	.816
		S2	.787	.619	.775	.601	.897	.804	.785	.616	.788	.621	.856	.733
		S3	.788	.621	.775	.601	.900	.810	.769	.591	.766	.587	.859	.738
		S4	.778	.606	.773	.597	.883	.779	.765	.585	.766	.587	.887	.788
Δ			.083	.135	.098	.152	.087	.149	.115	.176	.171	.253	.175	.320
Mean			.800	.641	.781	.610	.865	.748	.777	.604	.767	.590	.848	.745
D-3	RF -1	S1	.907	.823	.904	.816	.903	.815	.900	.811	.898	.806	.937	.907
		S2	.914	.836	.913	.834	.920	.846	.915	.837	.914	.835	.946	.894
		S3	.902	.813	.902	.814	.925	.855	.905	.819	.905	.820	.933	.870
		S4	.906	.820	.903	.816	.927	.860	.906	.820	.905	.819	.938	.880
	RF -2	S1	.897	.805	.893	.798	.934	.872	.890	.792	.887	.787	.941	.897
		S2	.936	.876	.934	.872	.955	.911	.934	.871	.933	.871	.962	.926
		S3	.911	.829	.905	.819	.943	.890	.906	.821	.889	.791	.949	.901
		S4	.910	.827	.908	.824	.944	.891	.909	.826	.909	.826	.951	.905
	RF -3	S1	.917	.842	.914	.836	.946	.895	.917	.840	.914	.835	.951	.917
		S2	.910	.829	.909	.826	.947	.897	.911	.830	.913	.834	.958	.917
		S3	.890	.793	.887	.787	.951	.904	.881	.776	.883	.780	.951	.904
		S4	.912	.833	.908	.825	.953	.908	.908	.824	.908	.825	.958	.917
Δ			.046	.083	.047	.085	.052	.096	.053	.095	.050	.091	.029	.056
Mean			.909	.827	.907	.822	.937	.879	.907	.822	.905	.819	.948	.903

D: Distribution type, RF: Response format, S: Sample Size

Table 2. Correlation and determination coefficients for situation-2

D	RF	S	Situation-2											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.937	.877	.926	.857	.779	.606	.934	.873	.935	.874	.950	.902
		S2	.938	.879	.934	.872	.911	.829	.939	.881	.941	.885	.915	.837
		S3	.940	.883	.933	.871	.906	.822	.933	.870	.929	.863	.956	.915
		S4	.938	.880	.937	.878	.930	.865	.940	.883	.939	.882	.945	.894
	RF -2	S1	.961	.924	.962	.925	.964	.929	.963	.926	.959	.919	.979	.959
		S2	.949	.901	.948	.898	.940	.884	.942	.887	.944	.891	.970	.941
		S3	.956	.913	.955	.911	.956	.914	.953	.909	.952	.907	.973	.947
		S4	.960	.922	.959	.920	.964	.929	.962	.926	.963	.927	.972	.944
	RF -3	S1	.965	.931	.963	.928	.975	.951	.960	.921	.954	.910	.985	.970
		S2	.954	.910	.948	.900	.972	.944	.951	.904	.951	.904	.970	.940
		S3	.963	.928	.963	.927	.969	.939	.962	.926	.963	.927	.977	.954
		S4	.962	.926	.962	.926	.972	.944	.962	.926	.961	.924	.980	.961
Δ			.028	.054	.037	.071	.196	.345	.030	.056	.034	.064	.070	.133
Mean			.952	.906	.949	.901	.937	.880	.950	.903	.949	.901	.964	.930
D-2	RF -1	S1	.942	.887	.944	.892	.883	.779	.943	.890	.943	.889	.946	.895
		S2	.959	.920	.962	.925	.947	.897	.962	.926	.961	.924	.973	.947
		S3	.946	.895	.949	.900	.919	.845	.946	.896	.947	.896	.957	.916
		S4	.947	.896	.950	.902	.954	.909	.950	.903	.950	.902	.963	.927
	RF -2	S1	.971	.942	.971	.942	.973	.947	.968	.938	.967	.935	.977	.955
		S2	.963	.927	.965	.932	.980	.961	.963	.927	.963	.927	.983	.965
		S3	.971	.944	.974	.949	.977	.955	.973	.947	.973	.947	.982	.965
		S4	.967	.935	.969	.938	.970	.941	.966	.934	.966	.934	.976	.952
	RF -3	S1	.977	.954	.977	.955	.985	.970	.976	.952	.977	.954	.989	.978
		S2	.970	.941	.973	.947	.983	.966	.969	.938	.968	.937	.984	.968
		S3	.976	.953	.977	.955	.981	.962	.977	.954	.977	.954	.983	.965
		S4	.970	.941	.970	.942	.978	.956	.969	.939	.969	.939	.982	.964
Δ			.035	.067	.033	.063	.102	.191	.034	.064	.034	.065	.043	.083
Mean			.963	.928	.965	.932	.961	.924	.964	.929	.963	.928	.975	.950
D-3	RF -1	S1	.953	.909	.952	.907	.895	.801	.951	.904	.939	.882	.955	.912
		S2	.928	.861	.920	.847	.863	.744	.920	.847	.659	.435	.936	.876
		S3	.936	.877	.934	.872	.947	.897	.935	.874	.930	.864	.958	.918
		S4	.942	.887	.941	.885	.945	.894	.940	.883	.938	.880	.952	.906
	RF -2	S1	.961	.924	.960	.922	.961	.924	.965	.931	.751	.565	.975	.950
		S2	.958	.917	.959	.920	.968	.937	.959	.919	.961	.924	.972	.945
		S3	.948	.898	.944	.890	.955	.912	.946	.896	.952	.906	.968	.937
		S4	.961	.923	.958	.918	.969	.939	.960	.922	.961	.924	.975	.951
	RF -3	S1	.972	.945	.973	.947	.977	.954	.975	.950	.974	.948	.973	.946
		S2	.968	.937	.968	.937	.973	.946	.966	.933	.961	.924	.967	.934
		S3	.950	.902	.949	.901	.972	.944	.950	.903	.948	.898	.972	.945
		S4	.955	.912	.953	.908	.973	.947	.953	.907	.951	.905	.981	.963
Δ			.044	.084	.053	.100	.114	.210	.055	.103	.315	.513	.045	.087
Mean			.953	.908	.951	.905	.950	.903	.952	.906	.910	.838	.965	.932

D: Distribution type, RF: Response format, S: Sample Size

Table 3. Correlation and determination coefficients for situation-3

D	RF	S	Situation-3											
			CTT		PCA		MR		ULS		DWLS		GRM	
			R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
D-1	RF -1	S1	.837	.701	.813	.660	.808	.653	.799	.639	.805	.648	.877	.769
		S2	.889	.790	.885	.783	.915	.837	.888	.789	.882	.778	.926	.857
		S3	.908	.824	.904	.817	.932	.869	.899	.809	.902	.814	.936	.877
		S4	.886	.785	.879	.773	.913	.834	.881	.777	.884	.781	.915	.838
	RF -2	S1	.907	.823	.902	.814	.951	.905	.901	.811	.899	.808	.957	.916
		S2	.914	.835	.911	.829	.941	.886	.909	.826	.909	.826	.954	.910
		S3	.907	.823	.902	.814	.948	.898	.900	.811	.899	.808	.948	.899
		S4	.918	.842	.910	.829	.953	.908	.909	.826	.903	.816	.953	.909
	RF -3	S1	.915	.837	.911	.829	.963	.927	.910	.829	.752	.566	.956	.914
		S2	.933	.870	.932	.868	.961	.924	.932	.869	.932	.868	.962	.925
		S3	.901	.812	.899	.809	.959	.919	.898	.806	.898	.807	.954	.910
		S4	.884	.781	.882	.778	.961	.924	.880	.774	.879	.773	.953	.907
Δ			.096	.169	.119	.208	.155	.274	.133	.230	.180	.302	.085	.156
Mean			.900	.810	.894	.800	.934	.874	.892	.797	.879	.774	.941	.886
D-2	RF -1	S1	.822	.676	.814	.662	.777	.604	.811	.658	.758	.574	.847	.717
		S2	.819	.671	.800	.640	.803	.644	.797	.636	.798	.637	.868	.753
		S3	.789	.622	.777	.604	.833	.693	.773	.598	.774	.599	.863	.745
		S4	.816	.666	.801	.642	.866	.751	.798	.637	.798	.637	.875	.766
	RF -2	S1	.765	.585	.737	.543	.836	.699	.735	.541	.657	.431	.777	.603
		S2	.811	.658	.794	.631	.875	.765	.790	.625	.790	.624	.750	.562
		S3	.804	.646	.794	.631	.882	.779	.790	.624	.790	.625	.860	.740
		S4	.828	.685	.802	.643	.894	.799	.800	.641	.799	.639	.900	.810
	RF -3	S1	.781	.610	.757	.573	.905	.818	.737	.544	.624	.389	.706	.498
		S2	.818	.669	.808	.652	.911	.829	.791	.626	.795	.632	.896	.803
		S3	.783	.614	.773	.598	.906	.820	.771	.595	.768	.590	.893	.797
		S4	.793	.629	.780	.609	.909	.826	.774	.599	.773	.597	.880	.775
Δ			.063	.100	.077	.119	.134	.225	.076	.117	.175	.250	.194	.312
Mean			.802	.644	.786	.619	.866	.752	.781	.610	.760	.581	.843	.714
D-3	RF -1	S1	.664	.441	.595	.354	.559	.313	.536	.288	.040	.002	.517	.267
		S2	.746	.557	.707	.500	.670	.448	.682	.466	.612	.374	.754	.568
		S3	.727	.529	.670	.449	.646	.418	.667	.444	.669	.448	.734	.538
		S4	.734	.538	.676	.457	.653	.427	.678	.460	.677	.459	.791	.625
	RF -2	S1	.684	.467	.572	.328	.550	.303	.538	.290	.587	.344	.525	.275
		S2	.698	.488	.656	.431	.631	.399	.641	.411	.545	.297	.685	.469
		S3	.705	.497	.623	.388	.737	.544	.613	.376	.637	.406	.713	.508
		S4	.668	.446	.637	.406	.741	.549	.628	.395	.622	.386	.741	.550
	RF -3	S1	.755	.571	.734	.539	.715	.512	.709	.503	.596	.355	.634	.402
		S2	.677	.458	.636	.405	.708	.501	.612	.374	.625	.391	.672	.452
		S3	.721	.520	.695	.483	.684	.468	.691	.478	.695	.483	.770	.593
		S4	.668	.446	.633	.400	.764	.584	.614	.377	.613	.376	.733	.538
Δ			.091	.130	.162	.211	.214	.281	.173	.215	.655	.481	.274	.358
Mean			.704	.497	.653	.428	.672	.456	.634	.405	.577	.360	.689	.482

D: Distribution type, RF: Response format, S: Sample Size

The coefficients of relationship obtained from the negatively skewed distribution in Situation-3 were high (r>.80). The correlation coefficients for the parameter estimates that the MR and DWLS models yielded increased particularly as the sample sizes increased. The average scores of the correlation coefficients that GRM yielded were the highest. The correlation coefficients obtained from the normal distribution in Situation-3 were moderate or high. The correlation coefficients that the DWLS and GRM models yielded were moderate in small sample sizes, but increased as the sample size increased. The correlation coefficients averages obtained from MR were the highest. It was found that there was zero correlation between the true score and sample size-1 and response format-1 conditions of the DWLS model in the positively skewed distribution in Situation-3. A relationship of moderate degree was observed in the other simulation conditions. It was found that the correlation coefficients that the DWLS and GRM models yielded were affected more by the simulation conditions; the correlation coefficients that CTT yielded had the highest average scores.

Whether or not the determination coefficients were affected by different simulation conditions were analyzed by Factorial ANOVA. Separate analyses were run for each Situation. It was found that the distribution types for Situation-1 (F(2, 215)=41.28, p<.001) and the interaction of the distribution types and statistical model effect were significant (F(10, 215)=4.60, p<.01). The effects of the response formats (F(2, 215)=1.24, p=.633), the sample size (F(3, 215)=1.30, p=.534) and the model (F(5, 215)=.68, p=.655) on the determination coefficient was not found to be statistically significant. According to the Bonferroni test, to determine the significance of the distribution type effects, the determination coefficients obtained from a negatively skewed distribution were found to be significantly lower than those obtained from the normal and positively skewed distributions; the determination coefficients obtained from a normal distribution were significantly lower than those obtained from a positively skewed distribution.

It was found that the effect of the response formats (F(2, 215)=27.59, p<.01) and the interaction of the response formats and model (F(10, 215)=2.01, p<.05) in Situation-2 were statistically significant. No statistical significance was found regarding the effects of the distribution types (F(2, 215)=11.75, p=.080), the sample size (F(3, 215)=1.65, p=.416) and the model (F(5, 215) = 1.77, p=.220) on the determination coefficient. According to the findings of the Bonferroni test, the determination coefficients obtained from the datasets that included items scored across seven categories were higher when compared to those items scored across three or five categories.

In Situation-3, the effects of the distribution types (F(2, 215)=156.31, p<.001) and the model (F(5, 215)=4.00, p<.01), the interaction of the distribution types and the model (F(10, 215)=4.94, p<.01), the interaction of the response formats and the model (F(10, 215)=4.55, p<.05) and the interaction of the sample size and the model (F(15, 215)=4.84, p<.01) were found to be statistically significant. It was found that the effects of the response format (F(2, 215)=.85, p=.502) and the sample size (F(3, 215)=11.36, p=.152) on the determination coefficient were not statistically significant. When the Bonferroni test was administered based on the distribution types, the determination coefficient findings obtained from the negatively skewed distribution were found to be significantly higher than those obtained from the normal and the positively skewed distributions. Similarly, the determination coefficients obtained from the normal distribution were significantly higher than those obtained from the positively skewed distribution. Based on the model, it was found that CTT yielded higher determination coefficients than did the ULS and DWLS models; PCA yielded higher determination coefficients than did the DWLS model, and the MR and GRM models yielded higher determination coefficients than did the CTT, PCA, ULS and DWLS models.

4. DISCUSSION AND CONCLUSION

In the present research study, where the basic simulative conditions were an item difficulty level of 20% below average, 20% above average, and normal, various distribution types, the effects of such simulative conditions as response formats and sample sizes on estimating the latent ability distribution were also investigated. To this end, ability parameters of true latent traits were identified and latent trait estimates were made with six different models within related simulative conditions.

In Situation-1, when the item difficulty was low, the distribution was negatively skewed, the response format was three and the sample size was small, all the models yielded values that were not related to the true ability parameters. It is recommended that none of the models should be utilized under these simulative conditions. As the sample size and response categories increased, moderate relationships started to be observed. The MR model, low item difficulty level, and a negatively skewed distribution do not yield accurate parameter estimations; however, in normal distributions, the MR model displays a better performance than do all the other models. All the models, primarily the MR model, are affected more by the negatively skewed distribution and, thus, do not make accurate estimations. However, when compared to normal distributions, positively skewed distributions can be said to yield better findings. Under these simulative conditions, CTT, MR and GRM display the best performances.

In Situation-2, the estimations yielded by the MR model was found to be affected by negatively skewed distributions, especially when the sample size is small. In Situation-2, determination coefficients increase as the response format and sample size increase. Under these simulative conditions, the GRM model displays the best performance.

The coefficients of relationship obtained in Situation-3 were moderate or high. The relationship coefficients that the DWLS and GRM models yielded were found to be moderate when the sample size was small, but higher when the sample size increased. Under these simulative conditions, CTT, MR and GRM displayed the best performances.

The findings of ANOVA, which was administered to determine whether or not simulative conditions affected determination coefficients, showed that particularly distribution types had a significant effect on determination coefficients in negatively skewed and positively skewed distributions. In the present research, where the distribution of item difficulty levels and distribution types were both studied, a significant effect of distribution types was an expected findings. It was found that the response format in Situation-2 and the model in Situation-3 were simulative conditions that had a significant effect. This significant effect in Situation-3 was in favor of particularly GRM and MR. While in Situation-1 and Situation-2 the model did not have a significant effect, the average determination coefficient values of the MR and GRM models were higher than those yielded by the other models. This situation shows that the general performance levels of MR and GRM, which produced latent ability estimations, are high.

In Situation-2, it was found that the significant effect of the response format on the determination coefficient was in favor of a seven-category response format. This finding is consistent with those reported in studies by Allahyari, Jafari and Bagheri (2016) and by Lozano et al. (2008). Allahyari et al. (2016) reported in their study that particularly in situations where the potential distribution was not normal, increasing a three or five-category response format to a higher category level would increase the power of the statistical model of Differential Item Functioning (DIF) by 8%.

The finding that the ability parameters that GRM yielded were higher than almost all other models under different conditions showed consistency with the findings reported in studies by Dumenci and Achenbach (2008) and by Hauck Filho et al., (2014).

When the general performance of the models are evaluated, it can be said that MR and GRM display a better performance than the other models. Particularly in situations when the distribution is either negatively or positively skewed and when the sample size is small, these models display a rather good performance.

The present study can be further developed by means of further studies on different simulation conditions. Iterative and bayesian parameter estimations, such as particularly

Markov Chain Monte Carlo, can be used. In addition, this study, the structure of which was based on a single dimension, can be developed by using multidimensional structures. Moreover, different polytomous parameter estimation models of IRT (such as the rating scale model –RSM) or nonparametric item response theory models can be used.

ORCID

Hakan Kogar https://orcid.org/0000-0001-5749-9824

5. REFERENCES

Allahyari, E., Jafari, P., & Bagheri, Z. (2016). A simulation study to assess the effect of the number of response categories on the power of ordinal logistic regression for differential ıtem functioning analysis in rating scales. Computational and mathematical methods in medicine, vol. 2016, Article ID 5080826. doi.org/10.1155/2016/5080826

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach (Vol. 904). John Wiley & Sons. doi.org/10.1002/9781119970583

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30(6), 505-514. doi.org/10.1016/S0160-2896(02)00082-X

Brzezińska, J. (2016). Latent variable modelling and item response theory analyses in marketing research. Folia Oeconomica Stetinensia, 16(2), 163-174. doi.org/10.1515/foli-2016-0032

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cyr, A., & Davies, A. (2005). Item response theory and latent variable modeling for surveys with complex sampling design: The case of the national longitudinal survey of children and youth in Canada. In conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2009). Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models. Alberta Journal of Educational Research, 55(4), 512.

DeCoster, J. (1998). Overview of factor analysis. Retrieved June 12, 2017 from http://www.stat-help.com/factor.pdf

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. Psychological assessment, 20(1), 55-62. doi.org/10.1037/1040-3590.20.1.55

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. Applied Psychological Measurement, 31(5), 457-459. doi.org/10.1177/0146621607299271

Hauck Filho, N., Machado, W. D. L., & Damásio, B. F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. Psicologia: Reflexão e Crítica, 27(4), 670-678. doi.org/10.1590/1678-7153.201427407

Jafari, P., Bagheri, Z., Ayatollahi, S. M. T., & Soltani, Z. (2012). Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQL TM 4.0 Generic Core Scales in school children. Health and Quality of Life Outcomes, 10(1), 27. doi.org/10.1186/1477-7525-10-27

Kline, R. B. (2005). Principles and practice of structural equation modeling (Second Edition). New York: The Guilford Publications.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods, 48(3), 936-949. doi.org/10.3758/s13428-015-0619-7

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. Behavior research methods, 38(1), 88-91. doi.org/10.3758/BF03192753

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. Methodology, 4(2), 73-79. doi.org/10.1027/1614-2241.4.2.73

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. Behavior Research Methods, 41(2), 295-308. doi.org/10.3758/BRM.41.2.295

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. Psychological Methods, 1, 293 – 299. doi.org/10.1037/1082-989X.1.3.293

Raykov, T. ve Marcoulides, G. A. (2000). *A first course in structural equation modeling.* London: Lawrence Erlbaum Associates, Inc.

Revelle, W. (2017). Package ‘psych’. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rizopoulos, D. (2017). Package ‘ltm’. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monographs, 34(Suppl. 17).

Saporta, G., & Niang, N. (2009). Principal component analysis: Application to statistical process control. Data analysis, 1-23. doi.org/10.1002/9780470611777.ch1

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 274–288

DOI: 10.21449/ijate.407218

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

Power Base Games That School Principles Use Scale: Its Development, Validity and Reliability

Muharrem Gencer [1*], **Türkay Nuri Tok** [2], **Aydan Ordu** [3]

[1] Ministry of National Education, Burdur, Turkey
[2] Izmir Democracy University, Faculty of Education, Izmir, Turkey
[3] Pamukkale University, Faculty of Education, Denizli, Turkey

Abstract: While organizations have a power struggle with their environment and with other organizations in the globalised world, employees who are the most important resource of the organization also have power struggle among themselves. To be successful in this power struggle, employees, especially managers, use a number of political games in the organization. Developing the scale of power base games that school principles use is the aim of this study to detemine how and how much school principles use power base games in schools which are educational organizations. The sample of this study consists of 213 teachers working in Yeşilova and Karamanlı districts of Burdur city in 2015-2016 educational year. In the evaluation of the scale by authorities, Lawshe technique was used and then, Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were used. After the analyses, it was found that this scale is a reliable and valid measurement tool which consists of 41 items and six dimensions as sponsorship, making alliance building, empire building, budgeting, expertise and lording.

ARTICLE HISTORY

Received: 28 October 2017

Revised: 19 February 2018

Accepted: 28 February 2018

KEYWORDS

School Principle,

Power Base Games, Validity,

Reliability.

1. INTRODUCTION

It is difficult to continue their existence for organizations in the globalised world. They have to adapt to changes and compete with their rivals to survive. Distribution of the scarce sources in the organization, not being able to see the future and changes in the organization or in the environment make employees in the organization display power struggle that is to say political games to secure their self-interests.

Gibson, Ivancevich and Donnelly (1988) who define political behaviour as one’s acting apart from normal power system to provide benefit for himself or for another unit indicate that individuals and units always engage in political behaviours. Sonaike (2013) classified the political behaviours which organization members use professionally and which increase collaboration in the organization as positive; and the political behaviours which serve one’s interests, destroy the collaboration among the units of the organization and demage team spirit as negative. Ferris et al. (1996) conceptualize organizational politics as a stressor because it causes stress and tension reactions.

CONTACT: Muharrem Gencer ✉ muharremgencer1978@hotmail.com Ministry of National Education, Burdur, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

Individual and organizational factors cause political behaviours to come to exist. As well as individual factors as high internal control, high machiavelist personality, investment to the organization, being aware of the employment opportunities and having a high success expectancy (Robbins & Judge, 2013), organizational factors as uneven distribution of the sources and power (Al-Tuhaih & Van Fleet, 2011), having a high centralization degree (Kesgen, 1999), institutional size, not having precise policies and heterogeneity of the members (Alp, 2010) are some of the causes for the occurrence of political behaviour. No matter what the reason is, politics is a fact that affects organizational climate in every kind of organization (Bodla & Danish, 2013).

The best way to define organizational politics is to perceive it as the games that organization members play (Mintzberg, 1989). The games displayed as political behaviors Mintzberg (1983) classifies into four categories: Authority Games, Power Base Games, Rivalry Games and Change Games (cited in Cacciattolo, 2014). In this study, Authority Games, Rivalry Games and Change Games were not included in the scale in consideration of the educational system in Turkey and the roles of school principles. The scale was developed regarding Power Base Games.

“Power” concept in the center of the human interest for management and organization can be defined as a source (a kind of power reserve) used by the effective person to change the behaviour of others (Porter, Angle & Allen, 2003). Power base games, on the other hand, can be defined as the games that an individual plays in paralel with the power to improve his organization power (Cacciattolo, 2014). Mintzberg (1983) categorizes power base games in which employees utilize all the opportunities to reach their goals and look out for themselves into six groups as *sponsorship game, alliance building game, empire building game, budgeting game, expertise game and lording game* (cited in Cacciattolo, 2014).

Sponsorship game includes the person who attaches himself to a rising or an established star. It is played by the ones who want to establish their own power center and these people play this game taking advantage of their superiors by declaring their loyalty in return for power (Mintzberg, 1985). Sponsor is generally the boss or someone else having more power and a better status (Gibson et al., 1988). *Alliance building game* is played among the persons who seek support (Yazıcı, Sezgin Nartgün & Özhan, 2015) and who are equal. It is played usually by managers who form an implicit contract in respect of supporting each other to build a power center in the organization and it is sometimes played by experts (Mintzberg, 1989). *Empire building game* is not played with equal persons, it is played by managers to build a power center by collaborating with subordinates on an individual basis. (Mintzberg, 1989). In *budgeting game*, individuals gather power with manipulation or by controlling financial resources. Players are the responsible persons taking part in the budgeting fields. The aim of the game is to guarantee uneven distribution of the undistributed total resources (Medwick, 1996). In *expertise game*, experts try to guarentee their positions using their special knowledge (Yazıcı et al., 2015). This game is played by the persons having technical competence or expertise that an organization needs. In this game players play aggressively laying emphasis on their specialities and competencies. Proficient players try to protect their unique competencies and abilities by keeping their knowledge to themselves (Mintzberg, 1989). In *lording game*, persons try to attain power by using their legitimate powers on subordinates (Yazıcı et al., 2015). The game is played to build a power center by “dominating” the legitimate power on the others, but legitimate power is not used or used a little (Mintzberg, 1989).

In organizations, especially managers use power base games (one of the political games) on the purpose of attaining power and protecting it for their personal gain. Because employees are affected negatively by the strict political games that managers use, some problems may occur such as; a decrease in the job satisfaction and in the organizational commitment perception, a drop in the motivations and performances of the employees, sense of exhaustion and not having an organizational socialization. However, Mintzberg indicates that if they are managed in an effective way, political games ease decision making, realize employees' individual aims for the future expectations and increase organizational productivity.

As a result, schools which are educational organizations can not be thought independently of power base games like other organizations. For this reason, determining how and how much these power base games are used by school principals by teachers who are the most critical members of the process and taking the proper steps are essential. Besides, having very few studies on this topic in the literature is a reason to develop the scale of power base games. So, it was aimed to develop the scale of power base games that the school principles use for the purpose of determining how and how much school principles use these power base games.

2. METHOD

2.1. Research Design

This study aims to develop a scale according to the perceptions of teachers for measuring the power base games that school principals use. The study planned for the purpose of developing a scale was formed on the validity and reliability analyses.

2.2. Sample

In the scope of the research**,** all of the teachers working in Yeşilova and Karamanlı districts of Burdur city were tried to be reached in the spring term of the 2015-2016 educational year. But, survey data was collected from 213 teachers. Tabachnick and Fidell (1989) indicate that data from 200 persons is enough for a factor analysis (cited in Büyüköztürk, 1997). Demographic information of the teachers in the research was given in Table 1.

Table 1. Demographic information of the teachers included in the pilot scheme

		f	*%*
Gender			
	Female	113	53.1
	Male	100	46.9
	Total	213	100.0
Branch			
	Math and Science Courses Teacher	52	24.4
	Non-math Courses Teacher	68	31.9
	Classroom Teacher	42	19.8
	Other Branch Teacher	51	23.9
	Total	213	100.0
Seniority			
	1-10 years	100	46.9
	11-20 years	87	40.8
	21 years and over	26	12.3
	Total	213	100.0
Length of Service at School			
	1-5 years	142	66.7
	6-10 years	41	19.2
	11 years and over	30	14.1
	Total	213	100.0

2.3. Preparing the Measurement Tool

In the first phase of the developing the scale, related literature was examined in detail and theoretical background was formed about the planned scale. The scale developed on the basis of Mintzberg's "Political Games Theory" was prepared about power base games by taking into consideration the roles of the school principals and the system of education in Turkey.

In the study, a text explaining the power base games was prepared by translating the definitions and proposals obtained from international literature. The text was given to 11 teachers and they were asked to write the games that school principals may display as an answer to open ended questions. Candidate scale's statements ware prepared by benefiting from the games that teachers wrote. Acquired statements were broached to four language teachers in order to provide the validity of language and in accordance with the suggestions, necessary corrections were made and the items of the scale were put into their final form. Totally, a pool with 70 items was created about Power Base Games and it was based on 6 factors in the form of 5-Likert scale.

2.4. Content Validity of the Power Base Games Scale

The study of content validity provides information on the representability and explicity of each item and it has the characteristics of a preliminary analysis for construct validity. Expert group offers concrete proposals for the development of the scale. Then, the reviewed draft scale is used in the pilot study to assess the other psychometric features (Rubio, Berg-Weger, Tebb, Lee & Rauch, 2003). Lawshe technique was used for the content validity in the study. The items of the draft scale were examined by 16 academic members (9 associate professors, 7 assistant professors) who were experts in educational administration and their opinions were taken about whether the items were related to the subject of the research or not. In Lawshe technique at least 5, at most 40 expert opinions are needed. (Yurdugül, 2005). Experts were asked to remark their answers about whether the items are proper for the scale on a three-point scale (1: must be cleared, 2: must be corrected, 3: must remain). There was some space under each item for experts to write their explanations and it was stated to the experts that they could make corrections on the items if necessary. After collecting the forms from experts, all the answers were reunited in a single form and content validity ratio was determined for each item. According to the criterion that Lawshe (1975) states, if the number of the experts are 16, minimum content validity ratio (CVR) should be taken as 0.49. The Formula of the Content validity ratio (CVR) for each item is indicated as (Lawshe, 1975):

$$CVR = \frac{N_G - N/2}{N/2}$$

In the Formula N_G stands for the number of the participants who say "necessary" and N stands for the total number of participants.

7 items whose content validity ratio was determined below 0.49 were removed from the scale. The content validity index (CVI) for 63 items remaining in the scale was found as 0.708. Content Validity Index is the mean of the CVR values of the remained items (Lawshe, 1975).

2.5. Data Analysis

The answers by 213 teachers to the items in the draft scale were computerized and data was analyzed. The skewness and kurtosis values of the data were examined in the normality test of the data set. According to Huck (2008) skewness and kurtosis values must be between -1 and +1 in a data set which shows a normal distribution. As a result of the analysis, skewness and kurtosis values were found between -1 and +1, so the data showed a normal distribution.

Firstly, item-total correlation was carried out in order to explain the relationship between the total score of the scale and the scores obtained from the items of the scale. Item analysis was conducted for discriminant validity of the item. Point averages of the groups consisting lower 27% and upper 27% were compared with Independent Two Sample t-Test in order to determine the distinctive strength of the items in the scale. Kaiser-Meyer-Olkin (KMO) and Barlett values were analyzed in order to determine whether the data was appropriate for the factor analysis. Exploratory Factor Analysis (EFA) was conducted for the construct validity of the scale and Confirmatory Factor Analysis (CFA) was conducted in order to look at the fit indices of the factors that came to exist afterward. In the study, the reliability of Power Base Games Scale was assessed with internal consistency (Cronbach Alpha values).

3. FINDINGS

3.1. Item Analysis of the Power Base Games Scale

Firstly, Item-total score correlation was carried out to explain the relationship between the total score of the scale and the scores obtained from the items of the scale. According to Büyüköztürk (2017), correlation coefficient calculated for the validity of the test is interpreted in point of significance .30 and higher correlations calculated for the validity coefficient can be assessed as an indicator of the validity of the test. In this study, the lower value of the item-total correlation was taken as .30. In Table 2 the item-total correlation of all items were given.

Table 2. Pearson product-moment correlation analysis results of the power base games scale

Item No	r	Item No	r	Item No	r
1	.64	22	.58	43	.78
2	.03	23	.49	44	.68
3	.63	24	.53	45	.78
4	.77	25	.67	46	.66
5	.63	26	-.03	47	.42
6	.66	27	.71	48	.63
7	.10	28	.69	49	.59
8	.80	29	.49	50	.82
9	.63	30	.46	51	.75
10	.74	31	.62	52	.76
11	.62	32	.40	53	.65
12	.61	33	.48	54	.67
13	.74	34	.53	55	.64
14	.77	35	.05	56	.63
15	.74	36	.42	57	.61
16	.71	37	.53	58	.64
17	.46	38	.50	59	.57
18	.61	39	.67	60	.64
19	.59	40	.69	61	.64
20	.61	41	.72	62	.68
21	.74	42	.70	63	.26

P<.05

Considering Table 2, 5 items whose item-total correlation was below .30 (2, 7, 26, 35, 63) were removed from the scale. Item-total correlation values of the remained 58 items differed between .82 and .40.

Item analysis was carried out for the discriminant validity of the 58 items in the candidate scale. Raw scores obtained from the scale were put in order from the highest to the lowest with the intent of determining the distinctive strength of the items in the scale. As a result of this ordering, point averages of the groups consisting lower 27% and upper 27% were compared with Independent Two Sample t-Test. In the results of the t-test carried out between the lower and upper groups, all items were found significant at a level of p< .05. All the results show that scale item scores and total scale score are distinctive.

3.2. Construct Validity of the Power Base Games Scale

Kaiser-Meyer-Olkin (KMO) and Barlett values were analyzed in order to determine whether the data was appropriate for the factor analysis. KMO coefficient enlightens whether data matrix is appropriate for factor analysis. KMO is supposed to be higher than .60 for factorability. Besides, if Barlett's test is significant, it means that data matrix is appropriate (Büyüköztürk, 2017). After the analyses in this study, KMO value for the factor analysis was found .942 and Barlett value was significant (X^2= 12984.894; p< .01). KMO and Barlett's Test values are in Table 3.

Table 3. KMO and Barlett's Test values

Kaiser-Meyer-Olkin Sample Adequacy		.942
Bartlett's Test of Sphericity	Chi-square values	12984.894
	S degree	1653
	p	.000

In the factor analysis of the 58 items remained in the draft scale, the scale was tested with Principal Component Analysis. Varimax rotation was used while analyzing the factors in the scale. According to Stevens (1996), twice the critical values in Table 4 should be taken to test the significance of the factor loading values that explains the relationship of the items with the factor. Considering Table 4, because sample is 213 in this study, factor loading value should be at least .36. According to Büyüköztürk (1997) while deciding whether an item should take part in a scale or not, loading value in the first factor must be .45 and higher than it. Also, the difference between the mentioned item's loading value in the first factor and in the other factors must be .10 and higher. In this study, in forming the factors, lower limit for the item factor loading was determined as .45 and loading value difference was determined as .10.

Table 4. Critical values for correlation coefficient in two-tailed tests

Sample Number	Critical Value	Sample Number	Critical Value	Sample Number	Critical Value
50	.361	180	.192	400	.129
80	.286	200	.182	600	.105
100	.256	250	.163	800	.091
140	.217	300	.149	1000	.081

α = .01

Resource: Stevens, J. (1996). *Applied multivariate statistics for the social sciences*, (3rd Edition), New Jersey: Mahwah, Lawrence Erlbaum.

As a result of Principal Component Analysis, 8 factors whose eigenvalue was higher than 1 were found. All of these 8 factors explain 72.89% of the variance. As it was thought to totalize the scale in 6 factors theoretically, we went for a six-factor solution and the number of the factors was determined as six. Findings based on Eigen values and the variances they explain were given in Table 5.

Table 5. Eigen values and the explained variances

Factors	Initial Eigenvalues			Total Factor Loadings		
	Total	Explained Variance (%)	Cumulative variance (%)	Total	Explained Variance (%)	Cumulative variance (%)
1	24.169	41.671	41.671	24.169	41.671	41.671
2	7.067	12.184	53.855	7.067	12.184	53.855
3	3.618	6.238	60.093	3.618	6.238	60.093
4	1.913	3.298	63.391	1.913	3.298	63.391
5	1.773	3.056	66.448	1.773	3.056	66.448
6	1.529	2.636	69.084	1.529	2.636	69.084

As seen in Table 5, when the number of the factors was determined as six, then, the total variance explained by these six factors was found as 69.08%. In the matter of total variance value that a scale must explain, Henson & Roberts (2006) indicate that a value at 52% or more must be provided in the scale studies (cited in Seçer, 2013). When factor loading values and factor structures of the items in the scale were analyzed, because some items (4, 6, 8, 10, 17, 25, 27, 39, 43 and 44) were not thought suitable to take part in the related factor theoretically, because the difference between the highest factor loading and the second highest factor loading of some items (21, 45, 46 and 50) was lower than 0.10 and because factor weight of some items (23, 28 and 49) was lower than .45, they were excluded from the analysis. 41 items remained after the analyses were subjected to factor analysis again specifying six dimensions. Information about factors in Factor Analysis was given in Table 6.

Table 6. Factor structure of the power base games scale and factor loading values of the items

Scale Items	Lording	Budgeting	Sponsorship	Alliance Building	Expertise	Empire Building
61	,896					
60	,881					
57	,849					
52	,846					
56	,844					
54	,832					
59	,780					
53	,779					
62	,778					
51	,767					
55	,761					
58	,757					
37		,791				
32		,782				
34		,780				
36		,750				
31		,745				
38		,730				
29		,678				
30		,648				
33		,571				
11			,915			
9			,914			
5			,908			
1			,902			
3			,890			

12				,754		
13				,711		
14				,685		
16				,682		
15				,673		
40					,755	
42					,734	
41					,728	
48					,589	
47					,524	
22						,677
19						,674
20						,621
24						,573
18						,493

The dimensions of the “Power Base Games Scale” consisting of 41 items were entitled in paralel with literature as “sponsorship”, “alliance building”, “empire building”, “budgeting”, “expertise” and “lording”. In the first dimension, “sponsorship”, there are totally 5 items (1, 3, 5, 9 and 11). Factor loadings of the items differ between 0.91 and 0.89. In the second dimension,” alliance building”, there are 5 items (12, 13, 14, 15 and 16) and the factor loadings of these items are between 0.75 and 0.67. Third dimension, “empire building”, consists of 5 items (18, 19, 20, 22 and 24). The factor loadings of these items differ between 0.67 and 0.49. In the fourth dimension, “budgeting”, there are totally 9 items (29, 30, 31, 32, 33, 34, 36, 37 and 38). The factor loadings of the items differ between 0.79 and 0.57. In the fifth dimensin, “expertise”, there are 5 items (40, 41, 42, 47 and 48) and the factor loadings of these items are between 0.75 and 0.52. Sixth dimension, “lording”, consists of 12 items (51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61 and 62). The factor loadings of these items differ between 0.89 and 0.75. The correlation coefficients between six factors with each other and total scale was given in Table 7. According to the correlation analysis, the relation between factors with each other and the total scale is significant.

Table 7. Correlation of the factors with each other

Factors	1	2	3	4	5	6	Total
Sponsorship(1)							
Alliance Building(2)	.454**						
Empire Building(3)	.477**	.626**					
Budgeting(4)	.298**	.654**	.401**				
Expertise(5)	.312**	.577**	.560**	.452**			
Lording(6)	.545**	.430**	.638**	.155*	.645**		
Total	.678**	.799**	.790**	.646**	.779**	.803**	

** Correlation is significant at the 0.01 level. *. Correlation is significant at the 0.05 level.

To determine the construct validity of the scale, Structural Equation Modeling was used. Structural Equation Modeling is seen as a combination of factor analysis and regression or path analysis (Hox & Bechger, 1998). The fitness of the obtained model tested with $\chi 2$ /df, IFI, CFI, RMSEA, NFI, TLI(NNFI), SRMR and RMR fit indices were given in Table 8.

Table 8. Goodness of fit indices for measurement model

Fit indicates	Fit Range	Research Model
X^2/sd	$0 \leq X^2/sd \leq 3$	1.683
IFI	≤0.90	.94
CFI	≤0.90	.94
RMSEA	0.05 ≤RMSEA≤0.08	.058
NFI	≤0.90	.87
TLI (NNFI)	≤0.90	.94
SRMR	0.05 ≤ - ≤0.10	.061
RMR		.097

As shown in Table 8 in the study, fit indices with respect to factor analysis, chi-square ($\chi2$) value for the scale and statistical significance were determined [$\chi2$= 1260.451, df= 749, p< .05]. When proportioning these values ($\chi2$ /df), the result was 1.683. As the obtained value is below 3, model fit can be interpreted as perfect (Hooper, Coughlan & Mullen, 2008; Schermelleh-Engel, Moosbrugger & Müller, 2003). Consequently, the values mentioned indicate acceptable fit. Path Analysis results showing the appropriateness of the scale items with one another and with the dimension of the scale items were given in Figure 1.

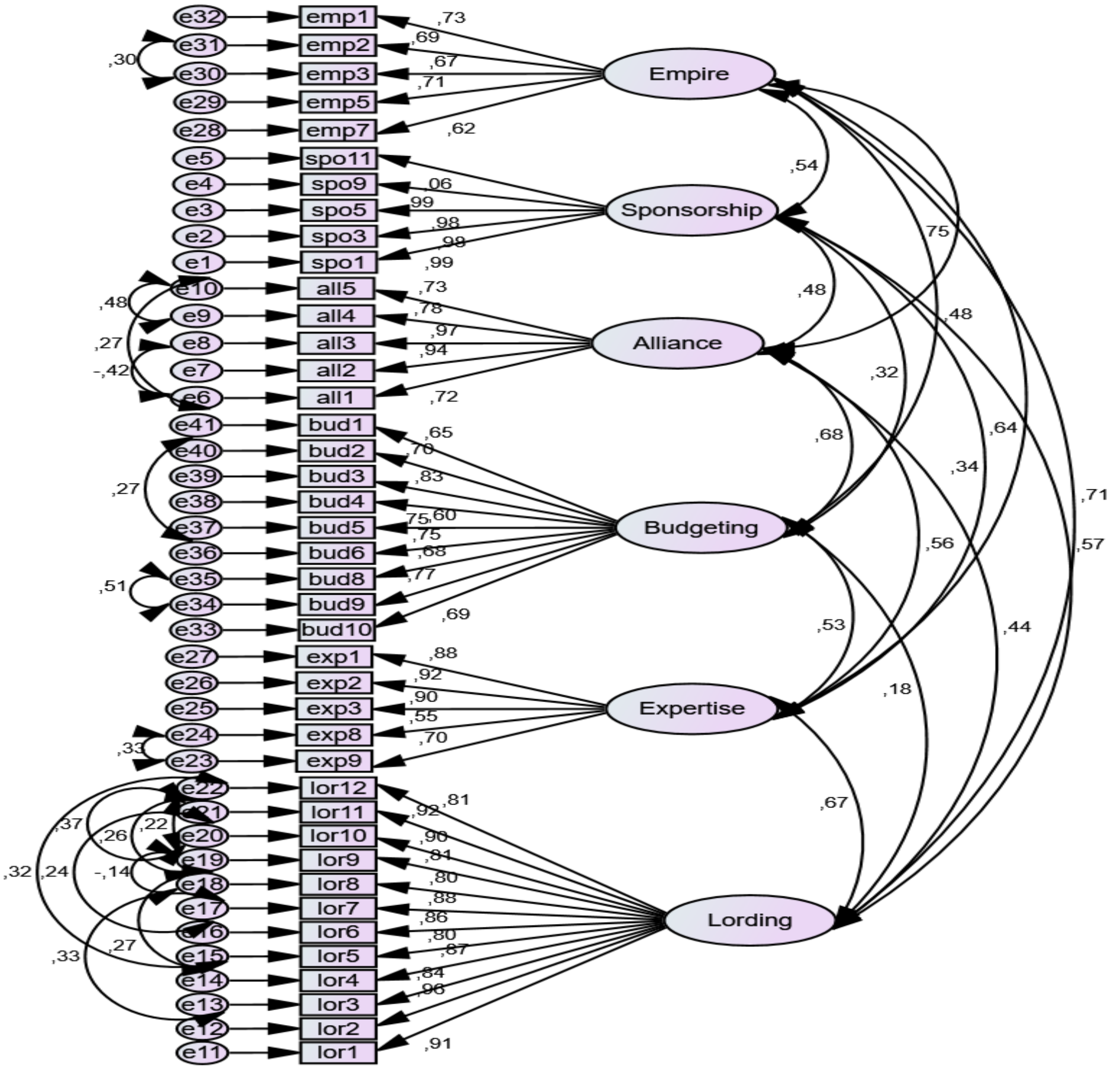


Figure 1. Path analysis results of the power base games scale

3.3. Reliability of the Power Base Games Scale

Reliability of the Power Base Games Scale was analyzed with internal consistency technique. According to this analysis, Cronbach Alpha internal consistency coefficient was found .98 in sponsorship dimension, .92 in alliance building dimension, .81 in empire building dimension, .90 in budgeting dimension, .88 in expertise dimension, .97 in lording dimension and it was found .95 for overall scale. Kılıç (2016) indicates that the scales whose Cronbach Alpha value is above 0.70 have internal consistency, that is to say, the scale is reliable. In Table 9, Cronbach Alpha coefficients of the dimensions and the overall scale were given.

Table 9. Cronbach Alpha Coefficients of the Power Base Games Scale Sub-dimensions

Sub-dimensions	N	Number of Items	Cronbach Alfa
Sponsorship	213	5	.98
Alliance Building	213	5	.92
Empire Building	213	5	.81
Budgeting	213	9	.90
Expertise	213	5	.88
Lording	213	12	.97
TOTAL	213	41	.95

Finally, factor-based Cronbach Alpha coefficients of the items were assessed. Cronbach Alpha coefficients of the dimensions differed between. 98 and .81 without item deleted. When an item was deleted, it differed between .98 and .75. When an item was deleted, Cronbach Alpha coefficients were not higher than the Cronbach Alpha coefficients of the dimensions, so it indicated that the reliability coefficient of the dimension wouldn't increase in case of deleting the item. This finding shows that item scorers are able to distinguish very well on the basis of dimension. In the light of this information, it can be said that the scale is reliable.

4. RESULTS and DISCUSSION

When "being a human" predominates essentially in working environments and when interests of people and groups get ahead of the interests of organizations, a different game is staged. This game is entitled as "organizational politics". The players of this game can be any employee at any level of the organization. The subject of the game is how these actors gain the power, protect the power and use the power to affect the individuals and organizational decisions (Kesgen, 1999). Especially power base games, political games that managers use against other people, may have positive and negative effects. Even so, it is clear that there is not enough knowledge about how to use these games in the organizations and about necessary competence for successful samples. Due to this, it was aimed to develop a scale for power base games which are indispensable in organizational life. On the basis of this aim, by making what the teachers think about the power base games that school principals use measurable, the opportunity to obtain quantitative data on this subject was tried to be created. In this context, "Power Base Games that School Principals Use Scale" was developed and the validity and reliability analyses of the scale were carried out.

On the basis of Mintzberg's "Political Games Theory", what the teachers think about the power base games that school principals use was determined from the results of the answers by teachers to the open ended questions at the beginning of the scale developing process. A pool with 70 items was created about Power Base Games and it was based on six factors in the form of 5-Likert scale. Seven items were removed from the scale by assessing the answers of experts about the appropriateness of the items for the scale according to Lawshe technique. Item-total correlation was conducted to the draft scale with 63 items. Five items having correlation values below .30 were removed from the scale. According to Büyüköztürk (2017), items with the item total correlation of .30 and above are more distinguishing. To determine the distinctive strength of the scale items, point averages of the groups that consist lower 27% and upper 27% were found significant for all items. Through exploratory factor analysis, the factors determined in accordance with Political Games theory were entitled as “sponsorship”, “alliance building”, “empire building”, “budgeting”, “expertise” and “lording”. It was determined that fit values relating to confirmatory factor analysis provide the specified criteria (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003; Hooper et al., 2008; Duyan & Gelbal, 2008).

The scales having Cronbach Alpha values of over .70 have internal consistency, which means the handled scale is reliable (Kılıç, 2016). It was revealed that calculated Cronbach Alpha Coefficients for the total scale and factors were appropriate. Also, when Cronbach alpha coefficients were analyzed, it was determined that the reliability coefficients of the dimensions wouldn’t increase in case of deleting the item.

Finally, in this study, it was found out that this scale developed for the educational institutions is a valid and reliable measurement tool consisting of six dimensions and 41 items. The developed scale revealed the perceptions of teachers' power base games used by school principals In the related literature, there are limited quantitative and qualitative measurement tools about political games and power base games (Chang, 2013; Medwick, 1996; Yazıcı et al., 2015). So, it is believed that this scale will contribute both to researchers and practitioners.

Acknowledgements

The abstract of this study was presented as an oral presentationin The Fourth International Eurasian Educational Research Congress held in Denizli on the 11-14th of May, 2017.

ORCID

Muharrem Gencer https://orcid.org/0000-0002-7212-8551
Türkay Nuri Tok https://orcid.org/0000-0002-2569-0576
Aydan Ordu https://orcid.org/0000-0002-2068-7992

5. REFERENCES

Alp, F. (2010). *Politik davranışın değişime dirence etkisi üzerine bir araştırma [A research on the effect of political behaviour to resistance to change]*. Yayımlanmamış Yükseklisans Tezi (Unpublished Master Thesis). Marmara Üniversitesi (Marmara University), İstanbul.

Al-Tuhaih, S. M. & Van Fleet, D. D. (2011). An exploratory study of organizational politics in Kuwait. *Thunderbird International Business Review.* 53(1), 93-104.

Bodla, M. A. & Danish, R. Q. (2013). The use of influence tactics in politicized organizations: A look from gender perspective. *Information Management and Business Review*,5(9), 456-462.

Büyüköztürk, Ş. (1997). Araştırmaya yönelik kaygı ölçeğinin geliştirilmesi [Development of anxiety scale towards research]. *Eğitim Yönetimi Dergisi (Educational Administration: Theory and Practice)*, 3(4), 453-464.

Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum [Data analysis handbook for social sciences: Statistics, research design, SPSS applications and comment]*. 23. Baskı (23rd Edition). Ankara: Pegem Akademi.

Cacciattolo, K. (2014). Defining organisational politics. *European Scientific Journal*, August Special Edition, 238-246.

Chang, C. L. H. (2013). The relationship among power types, political games, game players, and ınformation system project outcomes - A multiple-case study. *International Journal of Project Management*, 31(1), 57-67.

Duyan, V. & Gelbal, S. (2008). Barnett Çocuk Sevme Ölçeği’ni Türkçeye uyarlama çalışması [The adaptation study of Barnett Liking of Children Scale to Turkish]. *Eğitim ve Bilim Dergisi (Education and Science Journal),* 33(148), 40-48.

Ferris, G. R., Dwight, D. Frink, M. C. G., Jing Z., Kacmar, K. M. & Howard, J. L. (1996). Perceptions of organizational politics: Prediction, stress-related ımplications, and outcomes, *Human Relations*, 49(2), 233–266.

Gibson, J. G., Ivancevich, J. M. & Donnelly, J. H. (1988). *Organizations.* (Six Edition). Illinois: Business Publications.

Hooper, D., Coughlan J. & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.

Hox, J. J. & Bechger, T. M. (1998). An Introduction to structural equation modeling. *Family Science Review*, 11, 354-373.

Huck, S. W. (2008). *Reading Statistics and Research*, Pearson: Boston.

Hu, L. & Bentler, P. M. (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, DOI: 10.1080/10705519909540118.

Kesgen, J. (1999). *Örgütsel politika ve yansımaları. [Organizational politics and reflections of it]*. Yayımlanmamış Doktora Tezi (Unpublished Doctoral Dissertation). Dokuz Eylül Üniversitesi (Dokuz Eylül University), İzmir.

Kılıç, S. (2016). Cronbach’ın Alfa Güvenirlik Katsayısı. [Cronbach’s Alpha Reliability Coefficient]. *Journal of Mood Disorders (JMOOD)*, 6(1), 47-48.

Lawshe, C. H. (1975). A Quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.

Medwick, J. (1996). *An Analysis of the political games played between and among faculty at the K-5 or K-6 Elementary – School Level*. Doctor of Education. Northern Illinois University, Illinois.

Mintzberg, H. (1985). The organization as political arena. *Journal of Management Studies*, 22 (2), 133-154.

Mintzberg, H. (1989). *Mintzberg on management.* New York: The Free Press.

Porter, L. W., Angle, H. L. & Allen, R. W. (2003*). Organizational influence processes.* Armonk, N.Y.: Sharpe.

Robbins S. P. & Judge T. A. (2013). Örgütsel davranış. [Organizational behavior]. Cev. Edit. İ. Erdem (Translation Edit. İ. Erdem). Ankara: Nobel Yayıncılık. Eserin orijinali 2011’de yayımlandı (The original work was published in 2011).

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, S. & Rauch, S. (2003). objectifying contente validity: conducting a contente validity study in social work research. *Soc Work Res*, 27(2), 94-100.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23-74.

Seçer, İ. (2013). *SPSS ve LISREL ile pratik veri analizi [Practical data analysis with SPSS and LISREL]*. 1. Baskı (1st Edition). Ankara: Anı Yayıncılık.

Sonaike, K. (2013). Revisiting the good and bad sides of organizational politics. *Journal of Business & Economics Research,* 11(4), 197, 202.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences,* (3rd Edition), New Jersey: Mahwah, Lawrence Erlbaum.

Yazıcı, E., Sezgin Nartgün, Ş. & Özhan, T. (2015). Political games in universities: A case study. *Procedia - Social and Behavioral Sciences*, 174, 2700–2712.

Yurdugül, H. (2005). Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması [Using content validity indices for content validation in scale development studies]. *XIV. Ulusal Eğitim Bilimleri Kongresi* (*14th National Educational Science Congress).* 28-30 September 2005, Denizli.

Appendix 1. Power Base Games That School Principles Use Scale

ITEMS	Strongly Agree	Agree	Moderately Agree	Disagree	Strongly Disagree
1. Our school principal establishes friendships with his superiors to gain prestige.					
2. Our school principal receives support from his/her union to be able to protect his current position.					
3. Our school principal is in close contact with people who have political identities in order to protect his/her current position.					
4. Our school principal establishes positive relationships with members of the parent-teacher association to protect his/her current position.					
5. When our school principal rewards teachers, he/she looks at their social status in society rather than their achievements.					
6. Our school principal makes common cause with vice-principals - against teachers - by receiving their support.					
7. Our school principal makes common cause with other school principals to gain power.					
8. Our school principal tends to make common cause with some institution directors to gain power.					
9. Our school principal comes to ignore the mistakes of the vice-principals in order to form an alliance.					
10. Our school principal praises vice-principals to get their support.					
11. Our school principal brings parents and teachers, who support his/her views, in parent-teacher association management board.					
12. Our school principal uses the projects of talented teachers to increase his / her own reputation.					
13. Our school principal directs teachers to organize social activities (proms, poetry recitations, etc.) to make his/her own advertising.					
14. Our school principal benefits from the support of members of the parent-teacher association board for his/her reputation.					
15. Our school principal communicates with some teachers outside the school (lunch, home visits, etc.) in order to increase his/her power in the school.					
16. Our school principal states to the higher authorities that allowances received for the school are insufficient.					
17. Our school principal motivates teachers to work more to get support from specific projects (such as overseas projects)**.**					
18. Our school principal makes an effort to obtain support from the parents who have economic power.					
19. Our school principal tries to increase the school budget with social activities.					
20. Our school principal is in the effort to use the school garden for the purpose of income (wedding hall, car park, tea garden, etc.) during the holidays.					
21. Our school principal would like to get more share from the National Education budget for the school.					
22. Our school principle wants teachers and vice-principals to work in a self-sacrificing way to increase school’s budget					

23. Our school principal asks the teachers to increase their efforts so that the school receives more allowance.					
24. Our school principal collects donations from people and organizations for the needs of the school.					
25. Our school principal emphasizes the importance of his/her own knowledge and skills at every turn.					
26. Our school principal states that the school needs his/her own knowledge and skills.					
27. Our school principal talks about the originality of his/her own ideas.					
28. Our school principal makes teachers feel that he/she has mastered the legislation on education.					
29. Our school principal reminds teachers of the position where the school came through his/her knowledge, talent and experience.					
30. Our school principal keeps teachers under pressure to fulfill his/her requests.					
31. Our school principal uses his/her statue to impose his/her ideas on teachers.					
32. Our school principal does not want to take the advices of teachers.					
33. Our school principal makes us feel that he gives the final decision on all issues.					
34. Our school principal creates an image that participation in projects which are not mandatory is mandatory.					
35. Our school principal warns teachers in a rude way about their mistakes.					
36. Our school principal is closed to criticism.					
37. Our school principal loads tasks to teachers outside of their job descriptions.					
38. Our school principal wants to check every incident in the school himself.					
39. Our school principal's wording against teachers is offending.					
40. Our school principal gives orders at every turn.					
41. Our school principal expects responsibility from teachers beyond their duties.					

International Journal of
Assessment Tools in Education

Volume: 5, Number: 2
July, 2018

ISSN-e: 2148-7456 Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Okul Müdürlerinin Kullandıkları Güç Merkezi Oluşturma Oyunları Ölçeğinin Geliştirilmesi: Geçerlik ve Güvenirlik Çalışması

Muharrem Gencer, Türkay Nuri Tok, Aydan Ordu

To cite this article: Gencer, M., Tok, T.N., & Ordu, A. (2018). Okul Müdürlerinin Kullandıkları Güç Merkezi Oluşturma Oyunları Ölçeğinin Geliştirilmesi: Geçerlik ve Güvenirlik Çalışması. *International Journal of Assessment Tools in Education*, 5(2), 274-288.

DOI: 10.21449/ijate.407218

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 289–300

DOI: 10.21449/ijate.408957

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

The Effects of Dynamic Criteria Mapping Assessment on Students' Conceptions and Writing Skills Development

Amare Tesfie Birhan [1]

[1] English Language and Literature, Bahir Dar Institute of Technology, Bahir Dar University, Ethiopia

Abstract: Learner center assessment procedure and application is very crucial for students writing skills improvement. Hence, this study aimed to explore the effects of dynamic criteria mapping assessment on students' conceptions and writing skills development with reference to Vygotsky, zone of proximity development. To examine the issues, time series, quasi experimental research design was employed. The major data gathering tools were pre and post-tests, questionnaire and focus group discussion. Multistage sampling technique was employed to choose the sample of the study, and 63 first year software engineering students were the subjects of the study. Among these participants, 32 students were assigned to experimental group and the other 31 students' were assigned to control group. The findings indicated that dynamic criteria mapping assessment was effective in improving students writing skills development; students were able to construct sentence with better text structure and arguments. Furthermore, they used various cohesive devices, appropriate punctuation marks and dictions in their writing. Moreover, the assessment techniques had changed their conceptions on learning writing skills and engagment in writing assessment. Generally, the researcher learned that dynamic criteria mapping assessment strategy was vital to enhance students writing skills and conceptions on learning writing skills. Lastly, it is recommended that teachers should prepare various and dynamic criteria with their respective students while they assess their students writing skills, and teachers should not use judgmental assessment techniques.

ARTICLE HISTORY

Received: 05 January 2018
Revised: 01 March 2018
Accepted: 04 March 2018

KEYWORDS

Dynamic criteria mapping assessment,
Writing skill,
Conception,
Zone of proximity development

1. INTRODUCTION

It is known that assessment is vital for educational process; it is also significant to follow learners' progress of learning, to make educational decisions, to determine the effectiveness of teaching and learning and to assess the strength and the weakness of a specific instruction (Angelo & Cross, 1993; CERI, 2008; Hyland, 2003). Accordingly, students' language skills and communicative competence have been assessed through various approaches. Researchers such as Isavi, (2012) and Hamp-Lyon (2015) mentioned that the history of foreign language assessment has been characterized by long, traditional and standardized tests, and it was judgmental, learner excluded and lacking support during assessment. Particularly, according to Breland (1983, p. 1), "writing has been assessed through direct way: samples of an examinee's

CONTACT: Amare Tesfie ✉ amaretesfie@gmail.com English Language and Literature, Bahir Dar Institute of Technology, Bahir Dar University, Ethiopia

ISSN-e: 2148-7456 /© IJATE 2018

writing are obtained under controlled condition and evaluated, or indirect way, students writing was assessed through grammar and sentences structure by multiple choices."

Recently, this judgmental and traditional assessments have been changed into learning oriented assessment; learners participate in every assessment procedures. The traditional assessment that dominated the late 1970s and 1980s are no longer meaningful (Fulcher & Davidson, 2007). Fulcher and Davidson (2007) recommended tasks that mirrored language use in the real world should be used in communicative language that reflect the actual purposes of communication in clear defined contexts. Accordingly, Hailay (2017) also asserted using traditional assessment in the writing assessment is no longer sufficient.

This brings a shift of paradigm into participatory, learning oriented assessment and continuous interaction between learners and teachers. With the growing awareness that assessment is more internal to the classroom and can serve as a bridge that connects teaching to further learning, learning-oriented assessment has recently started to receive attention (Colby- kelly & Turner, 2007; Turner, 2012; Purpura & Turner's 2014 as cited in Kim & Kim, 2017).

Furthermore, Whit's 2009 study (as cited in Crusan, Plakans & Gebril, 2016) stated that, assessment remains as major element of any writing classroom instruction, and with the argument that assessment is not simply assigning grades for learners (Hyland, 2003), students have to participate in every procedure to develop their require skills.

According to Ethiopian Ministry of Education (2013), in Ethiopian higher institutions, writing courses are given for all undergraduate students to enable them to use the target skill in their academic, general and professional purposes. Particularly, courses like basic writing skills, intermediate writing skills, advanced writing skills, technical and research report writing and senior essay courses are given for content area and English major students, but students writing skills are being assessed traditionally with holistic approach or static techniques.

These frustrate students to engage in writing skill activities and to be a good writer with the target language. Aghaebrahimian, Rahimirad, Ahmadi & Alamdari (2014) argued that one-shot test administration has always been a challenge for learners by increasing their stress.

In addition, researchers mentioned the assessment technique which does not consider context, learners' language ability and course objectives fail to include essential elements of writing. Moreover, Broad (2003, p.9) stated "writing which is assessed through rubrics made writing less capricious." Xiaoxiao and Yan (2010) also added that writing is a complicated activity, containing abilities, such as choosing suitable topics according to target audience, generating logical and clear ideas, structuring rich and proper content, demonstrating accurate language expressions.

Thus, it seemed that it is very difficult to assess and judge students writing through conventional or static approach. Evaluating students' work is more complex than static rubrics (Broad, 2003). Since static rubrics are used only to secure inter-rater reliability (Beason, 2005; Rezaei & Lovorn, 2010; Janssen, Meier & Trace, 2015), it may not be appropriate for students who have little exposure to use and practice writing skills.

Moreover, it is also believed that feedbacks which are employed in writing classes are an integral part of assessment and helps learners to improve learners writing skills and to minimize errors (Grami, 2005; Tekle, Endalfer & Ebabu, 2012), but researchers such as Yiheyis & Getachew (2014), investigated that the assessment techniques which supposed to implement in higher institutions to maximize students engagement in writing skills were not effective. Yiheyis & Getachew added that teachers provide more of the quantitative feedback, and self and peer assessment were poorly utilized. Amare (2017) also proved that self, peer and teacher feedbacks are practiced rarely in EFL context. These and other factors contributed to students' difficulty in learning writing skills as a foreign and second language context (Richards, 1990; Kim & Kim, 2005).

These may affect students' conceptions on learning writing skills. As Temesgen (2013) investigated EFL students have wrong perception on writing skills. Moreover, the researcher experience revealed that students had low level of self-efficacy on writing skills. They perceived writing is one of the most difficult skills which they could not improve.

These conceptions might be based on lack of writing skills exposure, teaching and learning methodology and assessment techniques. According to (Freeman & Richards, 1993; Mclean, 2001) students develop a conception about their education, language background, schooling, exposure about learning and assessment. Furthermore, teachers understanding about teaching and assessing writing also affects students' perceptions (Endalfer, Ebabu & Tekle 2012; Escorcia, 2015) on learning writing skills.

The conceptions which students hold could be changed through continuing support, follow-up and constructive feedback and learning oriented assessment. These make students effective in their writing skills by engaging more in different activities. According to Tuan, Chin & Shieh (2005, p.641), "when students perceived that they are capable, and they think the conceptual change tasks are worthwhile to participate in and their learning goal is to gain competence, then students will be willing to make a sustained effort and be engaged in making conceptual change".

Currently, writing assessment and learners conceptions have been the focus of researchers such as (Anderson & Mohrweis, 2008; Lovorn & Rezaei, 2011; Li & Lindsey, 2015; Trace, Meier, Janssen, 2016). However, the mentioned researchers did not address interactive, dynamic assessment and the effects of the assessment technique on students' conceptions.

Accordingly, this research concentrated on assessment which helps students learning, which gives a chance to teachers' continuous support and feedback and which makes students in a part of assessment and learning in dynamic assessment criteria. In addition, the research also discovered how dynamic criteria mapping assessment could improve learners' conceptions towards learning writing skills and participating in writing assessment.

1.1. Literature Review

1.1.1. Dynamic Criteria Mapping Assessment

Dynamic assessment idea was practiced mainly by Feuerstein and Vygotsky with the main notion of zon of proximity development (Xiaoxiao & Yan, 2010). It is a strategy which is implemented through teachers help and mediation to develop students learning and to understand the potential for the development in learning (Alavi & Taghizadeh, 2014).

Likewise, dynamic criteria mapping assessment (DCMA) is an assessment and learning approach which was introduced by Broad (2003) and developed with the notion of Vygotsky, zonal proximity development to maximize students learning in writing skills. It is with the assumption that learners progress their learning and acquire the require skill through assisting and mentoring by their adult peers and teachers (Poehner, 2005). Researchers (e.g. Shrestha & Coffin, 2012; Christmas, Kudzai & Josiah, 2012; Chanyalew & Abiy, 2015) explained that learners develop their learning and understand the concept through adult guidance and mediate by capable peers and teachers. Similarly in dynamic assessment approach, students are mediated by their peers and the teacher, and students able to engage in every assessment procedures. In addition, it is implemented continuously with frequent feedback, interactive evaluation, and dynamic criteria. Broad (2003) recommended that instructors prepare dynamic criteria which help them tell the truth about what teachers believe, teach and value in evaluating students text. Sills (2016, p. 3) also claimed that the "writing assessment articulates values about what constitute good writing."

Furthermore, the teacher who uses the approach prepares the dynamic criteria by assuming learners language proficiency, course and program objectives, language policy and what they value in their context. According to Zepernick (nd, p. 137), "DCMA privileges local control in every aspect of the assessment process, celebrates the complexity and diversity of features that might represent good writing in any given context and honors the rhetorical process of negotiating local values."

In addition, West-Puckett (2016) argued it is locally responsive assessment which is designed through engagement of all teachers and all students in active, participatory and critical negotiation of assessment paradigms. Johnson and Schuck (2014) asserted that the assessment approach has been successfully used in writing programs to clarify the rhetorical values at play in the classroom and to engage teachers and learners in dialogue concerning how written works are assessed.

Moreover, it also provides the practitioners with a means of continuous evaluation and more reliable means of assessment, (Aghaebrahimian, et al. 2014). It follows a ground-up approach and provides an opportunity to restructure conversation about learning (Broad, et al. 2009; Breideband, 2016).

1.1.2. Students Conceptions on Teaching/Learning Writing Skills

Conceptions on learning and assessment refer to the personal beliefs and assumptions people have about their own learning and assessment (Steketee, 1996). As Steketee (1996) cited in Van Rossum and Schenk (1984) stated that conceptions are subjective statements which incorporate the assumptions, rules and conventions that influence the way individuals perceive knowledge as well as the way they approach learning task.

Researchers (Mclean, 2001; Brown & Hirschfeld, 2008; Alamdarloo, Moradi & Dehshiri, 2013; Escorcia, 2015) argue that conceptions that students have greatly impacts their academic achievement and motivation. Pajares, (1992) and Thomson (1992) also assert that teachers' beliefs of teaching and learning curricula influence strongly how they teach and how students learn and achieve.

Hence, learners' conceptions could be changed through proper mediation between teachers and students, and learners cognition which is originated shaped through cultural and social interaction process (Watson –Gegeo, 2004; Lantolf & Thorne, 2006 as cited in Isavi 2012).

Hence, this research attempted to answer the following research questions.

1. Is there a significant difference in students writing skills development between students who are assessed through dynamic criteria mapping assessment and students who are assessed through conventional approach?
2. Is there a significance difference in students writing development among the different types of assessments?
3. What is the effect of dynamic criteria mapping assessment on students' conceptions towards learning and assessing writing skills?

2. METHOD

2.1. Research Design

The purpose of this research was to explore the effects of dynamic criteria mapping assessment to students writing development. It also endeavored to investigate the assessment approach on students' conceptions towards learning and assessing in writing skills. Thus, the researcher employed pretest and post-tests techniques to examine their writing skills improvement. Hence, the research was designed through time series quasi experimental research design.

2.2. Population and Sample

The participants of the study were software engineering first year students in the 2016/2017 academic year who were taking basic writing skill course. Multistage sampling technique was employed to choose one section and one department from a total of 29 sections first year students and 14 departments of Bahir Dar Institute of Technology, Bahir Dar University.

Thus, the researcher passed four procedures to determine the study population. First, computing faculty was selected among five faculties (mechanical and industrial engineering, electrical and computer engineering, civil and water resources engineering, chemical and food engineering and computing technology) through systematic sampling technique. Second, of the selected faculty software students were chosen among computer science, information system and information technology departments. Finally, first year students in stated academic year were selected purposively since they were taking basic writing course.

Hence, all population (63 students) which were assigned in the department participated in the study. From these populations, 31 students were assigned to control group and the other 32 students were in experimental group, and the researcher believed that one section with 63 number of population were manageable to give constructive feedback and to follow up their learning progress.

Moreover, the selected subjects were supposed to take two English courses in the 2016/2017 academic year; in the first semester they took communicative English course and in the second semester basic writing skills. Though there is no clear evidence, the participants are believed they are intermediate language users which they can understand main ideas on familiar points and frequent expressions.

2.3. Data Gathering Instruments

Pre and Post-test, questionnaire and focus group discussions were used as instruments of the study. Tests were designed to explore students writing improvement within the time series, and focus group discussion and questionnaire were employed to explore students' conceptions on teaching writing and assessment of writing skills.

The questionnaire was adopted from (Abiy, 2005; Neibling, 2014). It includes 14 close ended items which were developed on a five likert scale (strongly agree, agree, undecided, disagree and strongly disagree). The expected mean determined for the one sample t-test was the middle value 3; hence, a mean value above 3 was considered as significant.

Besides, focus group discussions were carried out, so the researcher prepared checklist with 6 themes, and the themes were what students perceived learning writing skills as technology student, how students found the assessment strategies which we used, students' feedback style preferences, students understanding on dynamic criteria assessment and what students want to focus while they write paragraph and essay.

2.4. Data Collection Procedures

The procedure of the study was based on the principles of social constructivist, zonal proximity development point of view. After the samples selected, the researcher delivered basic writing courses for one semester for both control and experiment groups, and in each content, instructor assessed frequently and mediated in each activities.

The experiment and control groups were assessed different writing discourses. Particularly, the control group wrote six (four paragraphs and two essays) discourses and they did not get peer or teacher feedbacks. Whereas, the experimental group wrote six discourses and they get frequent support from both peers and the teachers based on the criteria prepared by the teacher and the students.

The criteria were focused on sentences structure, logical arguments, cohesive devices, developing unified texts, mechanical aspects and syntax. Both groups were assessed within different time intervals, and instruction and assessment for the two groups were carried out for three months simultaneously. However, for this research purpose four paragraphs (one pretest and three post-tests) which were written by students were taken in to consideration.

Moreover, students' questionnaire and focus group discussion checklist were passed with two validation procedures. Firstly, they were reviewed by PhD students and colleagues at Bahir Dar University. Then, the checklist and the questionnaire were revised and administered to the target population. Finally, it was checked through Cronbach's Alpha and its reliability found as .750.

2.5. Data Analysis

Data gathered from the tests and the questionnaires were analyzed using SPSS version 20. Thus, independent sample t-test, descriptive statistics, repeated measure analysis of variances (ANOVA) and one sample t-test were used. Hence, independent sample test was used to examine the statistical difference between the control and the experimental group, and repeated measure ANOVA was used to determine the time series statistical difference among the tests of the groups.

Moreover, descriptive statistics was employed to identify which writing elements contribute more for this difference. In addition, one sample t-test was also employed to determine the level of students' conception towards students learning writing skills and assessing writing skills. The focus group discussion and the quantitative data were analyzed through concurrent mixed method which quantitative and qualitative data were analyzed thematically.

3. FINDINGS

3.1. Students Writing Skills Development

According to Ismail (2011) the purpose of learning writing skills is to be able to communicate through writing in real life and academic situations. In the study participants wrote paragraphs and essays before and after intervention. Hence, to observe the difference and participants writing skills improvement of the two groups, the independent t-test was run. The statistical difference of the two groups are summarized in Table 1 below.

The table illustrates that there was statistically significance difference between the control and the experiment group which t (61) =6.087, $p<0.05$, and the statistical mean of the experiment group was 10.73, but the control group mean was 8.39 which indicated the experimental group improved their writing skills. Hence, students who were assessed through dynamic criteria assessment improved their writing skills. Students who assessed through dynamic criteria have changed their academic writing results as well as their writing skills. Graham (2008) mentioned that since writing is a complex skill, students require considerable effort and time to use.

Table 1. statistical difference between the control and the experimental group

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Post test3	Equal variances assumed	2.331	.132	-6.087	61	.000	-2.347
	Equal variances not assumed			-6.115	57.248	.000	-2.347

Furthermore, descriptive statistics was run to observe the writing features the students improved. Thus, they abled to develop unified and coherent texts (1.278 in pretest and 2.087 in posttests) and usage of cohesive devises (.976 mean in pre -test and 1.833 in post -test), and the least students writing skills improvements were observed in using explaining ideas logically and using persuasive ideas in their paragraphs (1.266 mean in pre- test and 1.857 in post -test). Finally, they have moderate improvement in sentences structure, proper use of capitalization and punctuation marks in both pre and post tests.

In this study the dynamic mapping assessment helped learners to develop unified paragraph and essay, to use appropriate transitional markers and to improve a text with correct sentences structure, punctuation and other mechanical aspects. As a result, based on what the teacher and students valued and included in the assessment as criteria, they improved their writing skills.

3.2. Analysis of the Time Series Progress of Students Writing

Moreover, the time series progresses of students writing improvement were observed through repeated measure ANOVA. As it is indicated in Table 2, dynamic criteria mapping assessment has improved students writing skills;

Table 2. Students' *writing skill improvement in paragraph*, Tests of within subjects effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
factor1	Sphericity Assumed	634.508	3	211.503	189.139	.000
	Greenhouse-Geisser	634.508	2.350	269.947	189.139	.000
	Huynh-Feldt	634.508	2.449	259.089	189.139	.000
	Lower-bound	634.508	1.000	634.508	189.139	.000

In addition, there was statistically significance difference among tests (pretest, post test1, post test2 and post test3) which F (3)= 189.139, $P< 0.05$ and indicates that students improve their writing skills with simultaneous intervention. The estimated margin mean indicate that the control group has a mean value 5.2, 6.2, 7.13, 8.0 in pre-test, posttest1, post- test 2 and post -test 3 respectively, but the experimental group means were observed 5.0, 6.00, 8.00 and 9.77 in pre- test, post1, post test2 and post test3 respectively. This also indicates students in both experimental and the control group improved their writing skills even the mean margin is different.

3.3. Students Conceptions towards Teaching Writing skills

Students' conceptions towards teaching and assessing writing skills were also investigated by questionnaire and focus group discussion. The purpose of the conception assessment was to check if students have changed their understanding about learning writing skills and assessing writing after the intervention.

Table 3. Students' conceptions towards learning writing skills

	t	df	Sig. (2-tailed)	Mean Difference
Class room practices of teaching writing skills	18.042	33	.000	1.13072

As it is explained in the Table 3, students were interested to participate in learning writing skills activities t (33)= 18.04, P<.0.05).This indicated that students have positive conceptions and they are very much motivated to participate in learning writing skills and participated in writing activities.

As a result, continuing support of learners and mediation among students helped them to improve their writing skills and to engage in activities. Besides, during focus group discussion students agreed that they have changed their perceptions toward improving and engaging in writing skills. They also find out that they could improve their writing skills, if they practiced well, and majority of the students agreed that they focused on their ideas without much worried about their mechanical errors and they believed that engaging in writing activities help them to use the skills in professional and academic writing.

3.4. Students Conceptions towards Writing Assessment

As it can be seen in the Table 4, learners have good understanding towards writing assessment t (33) = 18.4, p<0.05. Hence the data showed that students changed their understanding while they were working with their peers and their teacher, and their conceptions towards participating in peer feedback and accepting teacher feedback during the lesson were improved. Students believed that continuous peer feedback and teacher feedback could help them to develop their writing skills.

Table 4. Students' conceptions towards writing assessment

	t	df	Sig. (2-tailed)	Mean Difference
Students conceptions towards writing assessment	18.448	33	.000	1.47941

In contrast, during focus group discussion, some students explained that they did not think peer feedback help them to improve their writing, and they highly attached to teachers feedback. They also believed that the teacher feedback could show them their gaps more than their peers' feedback. Furthermore, the students agreed on the assessment criteria that the teacher and the students set during writing assessment, and they stated that the class room assessments have impact in their writing development. Lastly, they thought that the criteria that we set told them how much they learnt and the various criteria helped them to see various features of writing.

4. DISCUSSION

Dynamic criteria mapping assessment which was used as an approach in this research encompasses extensive activity, collective feedback and interaction. The ultimate goal of teaching writing is effective written communication (Seifoori, Mozaheb & Beigi, 2012), and this research proved that dynamic criteria mapping assessment is an effective strategy to improve students' written communication.

The students' paragraph and essay had poor introduction in the pre -test sessions with no clear topic sentences and thesis statement, and various unrelated and incoherent ideas were observed in the paragraphs and essays, but as Johnson and Schuck (2014) and West-Puckett

(2016) mentioned, dynamic characteristic assessment impacted students'learning and they gain better understanding of how writing is learned, practiced and valued. Hence, they abled to construct paragraphs and essays with clear topic sentences and thesis statements and with related and coherent supportive details. Aghaebrahimian, et al. (2014) reported the similar finding which the approach improves students writing skills.

Likewise, the zone of proximity development which is the main characteristics of the dynamic assessment approach (Aghaebrahimian, et al. 2014; Nazari, 2017) is also confirmed by (Isavi, 2012; Marzec-Stawiarska, 2016) as an effective strategy to mediate students writing skills, and feedback strategies (self, peer and teacher) are also effective (Diab, 2016) to scaffold their learning even if students prefer to receive feedback from their teachers. However, Diab (2016) recommended that since self and peer feedback are helpful to reduce students' error significantly, teachers should train them how to get and give feedback, and Yu and Lee (2016) also confirmed that peer feedback strategies help learners to improve learners.

In addition, the research showed students' conceptions changed through participatory and dynamic criteria assessment strategy, and this also improved students enegagment in learning and particepating in different writing activities. Researchers such as (Temesgen, 2013; Krawczyk, 2001) stated that students are motivated to engage if they have positive cognition on writing skills. Carless (2007) also asserted assessments which promote the kind of learning, involvement of students in the assessment process and feedback promotes students engagement and action.

5. CONCLUSION

The purpose of this research was to explore the effects of dynamic criteria mapping assessment towards students' writing development and students' conceptions towards writing assessment and learning writing skills. The findings indicated that the assessment contributed for the improvement of students' writing development. Specifically, students develop mechanical aspects like spelling, grammar, punctuation and cohesion (using grammatical elements like connectives, substitution, association and conjunctions).

Furthermore, according to the data, students have changed their conceptions towards learning writing skills and writing assessment. They believed teaching writing contributed for their academic and professional purposes. Hence, students, teachers and other practitioners should work together to enhance students' writing skills, and learning oriented assessment like dynamic criteria mapping assessment can be an alternative assessment technique to develop students' writing skills and to chnage thier concptions.

ORCID

Amare Tesfie Birhan http://orcid.org/0000-0002-8764-8592

6. REFERENCES

Abiy Y. (2005) Effects of Teacher Mediation on students' conceptions and Approaches of reading. Unpublished PhD dissertation. Addis Ababa University.

Abiy Y. (2013). High school English teachers' and students' perceptions, attitudes and actual practices of continuous assessment. *Global Journal of Teacher Education*. 1(1), 112-121.

Alamdarloo, G.H., Moradi, S., & Dehshiri, G.R. (2013). The relationship between students' conceptions of learning and their academic achievement, *Psychology*, 4 (1), 44-49.

Alavi, S.M., & Taghizadeh, M. (2014). Dynamic assessment of writing: The impact of implicit/explicit mediations on L2 learners' internalization of writing skills and strategies. *Educational assessment*, 19 (1), 1-16.

Aghaebrahimian, A., Rahimirad, M., Ahmadi, A., & Alamdari, J. (2014). Dynamic Assessment of writing skill in Advanced EFL Iranian Learners*: International Conference on Current trends in ELT.*

Amare T. (2017). Teachers' Cognition on process genre approach and practices of teaching writing skills in EFL context. *English for specific purposes world*, 54 (19), 1-17.

Anderson, J.S., & Mohrweis, L.C. (2008). Using rubrics to assess accounting students' writing, oral presentations and Ethics skills. *American Journal of Business Education*, 1 (2), 85-94.

Angelo, T.A., & Cross, K.P. (1993). *Classroom assessment techniques*: A handbook for college teachers. San Francisco: Jossey-Bass.

Beason, L. (2005). Review of what we really value: beyond rubrics in teaching and assessing writing: *Council of writing program administration*.

Breland, H. M., (1996). *Writing Skill Assessment Problems and Prospects*. Policy information center. Princeton, Educational Testing service.

Breland, H.M., (1983). The direct assessment of writing skill: A measurement review, College Board report, No. 83-6. Retrieved on October 20, 2017 from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1983-6-direct-assessment-writing-measurement.pdf

Breideband, T. (2016). *Alternative Assessment criteria, but how? George State University Student Innovation.* Retrieved on July, 17, 2017 from http://sites.gsu.edu/innovation/2016/02/01/alternative-assessment-criteria-but-how/.

Brindley, G. (2001). Assessment. In R. Carter & D. Nunan (eds.), *the Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: Cambridge University Press, pp.137-143.

Broad, B., Adler-Kassner, L., Alford, B., Detweiler, J. Estrem, H., Harrington, S., McBride, M., Stalions, E., & Weeden, S. (2009). *Organic writing assessment: Dynamic criteria mapping in action.* Utah, Utah state university press.

Broad, B. (2003). *What we really value: Beyond Rubrics in Teaching and Assessing writing*. Utah. Utah state university press.

Brown, G.T. (2004). Teachers' Conceptions of assessment: implications for policy and professional development. *Assessment in education*. 11(3), 301-318. DOI: 10.1080/0969594042000304609.

Brown, G.T.L. & Hirschfeld, G.H.F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: principles, policy and practice*. 15 (1), 3-17.

Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in education and teaching international*, 44 (1), 57-66.

CERI (2008). Assessment for learning- the case for formative assessment, retrieved on June 20, *2017 from www.oecd.org/site/educeri21st/40600533.pdf.*

Chanyalew E. & Abiy Y. (2015). Effects of teacher scaffolding on students reading comprehension. *Science, Technology and Arts Research Journal*, 4 (2), 263-271.

Christmas, D., Kudzai, C., & Josiah, M. (2012). Vygotsky's Zonal Proximity Development Theory: what are its implications for Mathematical teaching*? Greener Journal of social sciences*, 3 (7), 371-377.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs and practices. *Assessing writing*, 28, 43-56.

Diab, N.M. (2016). A comparison of peer, teacher and self-feedback on the reduction of language errors in student essay, *system*, 57, 55-65. http://dx.doi.org/10.1016/j.system.2015.12.014.

Escorcia, D. (2015). Teaching and assessing writing skills at university level: a comparison of practices in French and Colombian universities, *Educational Research,* 57, (3), 254-271.

Freeman, D. & Richards, J. (1993). Conceptions of teaching and the education of second language teachers. *TESOL Quarterly*, 27(2), 193-216.

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment*. London: Routledge, Taylor and Francis group.

Graham, S. (2008). *Effective writing instruction for all students,* Renaissance learning.

Hailay T. (2017). Investigating the practices of assessment methods in Amharic language writing skill context. The case of selected higher education in Ethiopia. *Educational Research and Reviews,* 12(8), 488-493.

Hyland, K. (2003). *Second Language writing*. Cambridge. Cambridge University press.

Isavi, E. (2012). The effects of dynamic assessment on Iranian L2 writing performance. Retrieved on October 25, 2017 from https://files.eric.ed.gov/fulltext/ED530902.pdf.

Ismail, S. A. A. (2011). Exploring students' perceptions of ESL writing. English Language Teaching, 4(2), 73-83.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing writing*. 26, 51-66.

Johnson, K.E. & Schuck, C. (2014). Using dynamic criteria mapping to improve curriculum alignment across institutions. Retrieved on July, 18, 2017 from http://cop.hlcommission.org/Assessment/johnson.html

Kim, J. & Kim, Y. (2005). Teaching Korean university writing class: Balancing the process and the genre approach. *Asian EFL Journal*. 7 (2). 1-15.

Kim, A.H., & Kim, H.J. (2017). The effectiveness of instructor feedback for learning-oriented language assessment: using an integrated reading-to write task for English for academic purposes. *Assessing writing*. 32, 57-71.

Krawczyk, J. (2001). Writing attitudes: Determining the effect of a community of learners project on the attitudes of composing students, MA thesis, Oklahoma state University.

Li, J., & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing writing,* 26, 67-79.

Lovorn, M. G., & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and In-service teachers. *Practical assessment, research and evaluation,* 16 (16).

Marzec-Stawiarska, M. (2016). The influence of summary writing on the development of reading skills in a foreign language, *system*, 59, 90-99.

McCune, V. (2004). Development of first year students' conceptions of essay writing. H*igher Education*, 47, 257-282.

McLean, M. (2001). Can we relate conceptions of learning to student academic achievement? *Teaching in higher education*, 6 (3), 399-413. DOI: 10.1080/13562510120061241

Ministry of education (2013). English department harmonized curriculum, unpublished curriculum, Addis Ababa.

Nazari, A. (2017). Dynamic assessment in higher education English classes: a lecturer perspective. *The Journal of Language Teaching and Learning,* 7(1), 100-118.

Neibling, J.L. (2014). Teachers' Conceptions Towards Type of Assessment: Grade Level and State Tested Content Area. MA thesis, Kansas University.

Pajares, F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62(3), 307-332.

Poehner, E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French*. Pennsylvania State University.

Rezaei, A.R.,& Lovorn, M. (2010). Reliability and Validity of rubrics for assessment through writing. *Assessing writing*, 15(1), 18-39.

Richards, J. (1990*). The Language Teaching Matrix*. Cambridge: Cambridge University press.

Seifoori, Z., Mozaheb, M.A., & Beigi, A.B. (2012). A profile of an effective EFL writing Teachers (A technology-based approach*). English Language Teaching*, 5 (5), 107-117.

Shrestha, P., & Coffin, C. (2012). Dynamic assessment, tutor mediation, and academic writing development. *Assessing writing*, 17 (1), 55-70.

Sills, E. (2016). Multimodal assessment as disciplinary sense making: Beyond rubrics to framework. *The journal of writing assessment*, 9 (2). Retrieved on October 27, 2017 from http://journalofwritingassessment.org/article.php?article=109

Steketee, C.N. (1996). Conceptions of learning held by students in the lower, middle and upper grades of primary school. Retrieved on September 14, 2017 from http://ro.edu.au/theses_hons/677

Tabar, M., & Davoudi, M. (2015). The Effects of computerized Dynamic Assessment of L2 Writing on Iranian EFL Learner's Writing Development. *International Journal of Linguistics and Communication.* 3 (2), 176-186.

Tekle F., Endalfer M., & Ebabu T. (2012), a descriptive survey on Teachers' perception of EFL writing and their practice of teaching writing: Preparatory schools in Jimma zone in focus. *Ethiopian journal of education and science*, 1 (1), 29-52.

Temesgen E. (2013), Factors that affect learners' motivations towards the writing skills: the case of grade twelve students in Wachemo preparatory school, Hosanna, MA thesis.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research: In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). New York: Macmillan.

Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubrics category interpretations through score negotiation. *Assessing writing*, 30, 32-43.

Tuan, H.L., Chin, C.C., & Shieh, S.H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science and Education.* 27 (6), 639-654.

Vygotsky, S. (1978). *Mind in Society: The Development of Higher psychology Processes:* Cambridge: Harvard University Press.

West-Puckett, S. (2016). Making classroom writing assessment more visible, equitable and portable through digital badging. *College English*, 79 (2), 127-151.

Xiaoxiao, L., &Yan, L. (2010). A case study of Dynamic assessment in EFL process writing. *Chinese Journal of Applied Linguistics*, 33 (1), 24-40.

Yu, S., & Lee, I. (2016). Exploring Chinese students' strategy to use in a cooperative peer feedback writing group, *system*, 58, 1-11.

Yiheyis S., & Getachew S. (2014). The implementation of continuous assessment in writing classes of Jimma College of teacher education. *Ethiopia Journal of Education and science*. 10 (1), 109-135.

Zepernick, J.S. (nd). Reviewed Organic writing assessment: Dynamic criteria mapping in action, by Broad, B., Adler-Kassner, L., Alford, B., Detweiler, J., Estrem, H., Harrington, S., McBrdide, M., Stalions, E., & Weeden, S. (2009). Longon: Utah State UP.

Zoghi, M., & Ma lmeer, E. (2013). The effect of Dynamic assessment on EFL learners' intrinsic motivation. *Journal of Language Teaching and Research*, 4(3), 584-591.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 301–313

DOI: 10.21449/ijate.402806

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

Data Fit Comparison of Mixture Item Response Theory Models and Traditional Models

Seher Yalçın [1*]

[1] Ankara University, Faculty of Educational Sciences, Department of Measurement and Assessment, Ankara, Turkey

Abstract: The purpose of this study is to determine the best IRT model [Rasch, 2PL, 3PL, 4PL and mixed IRT (2 and 3PL)] for the science and technology subtest of the Transition from Basic Education to Secondary Education (TEOG) exam, which is carried out at national level, it is also aimed to predict the item parameters under the best model. This study is a basic research as it contributes to the information production which is fundamental for test development theories. The study group of the research is composed of 5000 students who were randomly selected from students who participated in TEOG exam in 2015. The analyses were carried out on 17 multiple choice items in TEOG science and technology subtest. When model fit indices were evaluated, the MixIRT model with two parameters and three latent classes was found to fit the data best. According to this model, when the difficulties and discrimination averages of the items are taken into account, it can be expressed that items are moderately difficult and discriminative for students in latent class-1; the items are considerably easy and able to slightly distinguish the students in latent class-2; the items are difficult to the students in the third latent class and they can slightly distinguish the students in this group.

ARTICLE HISTORY

Received: 10 February 2018
Revised: 24 March 2018
Accepted: 29 March 2018

KEYWORDS

Item response theory,
Latent class models,
Mixed item response theory,
Model-data fit

1. INTRODUCTION

The purpose in educational and psychological measurements is to ensure that the decisions made about the individual are valid and reliable. To this end, models and theories which try to better demonstrate the state of individual's having the measured characteristics are being developed. Within the scope of the models known as latent variable models; structural equation models, latent class models, latent profile models, and latent trait models (item response theory) are discussed (Skrondal & Rabe-Hesketh, 2007). Commonly used theories in the literature are: Classic Test Theory (CTT) and Item Response Theory (IRT). If the assumptions are met, IRT models are often preferred over CTT because CTT fails to provide as much information as IRT due to the limitations of the theory [e.g. individuals' ability levels depend on the item they receive, the item properties depend on the respondent group; it is difficult to compare individuals who take different tests and the need for parallel tests for

CONTACT: Seher Yalçın ✉ yalcins@ankara.edu.tr Ankara University, Faculty of Educational Sciences, Department of Measurement and Assessment, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

reliability prediction (Hambleton, Swaminathan & Rogers, 1991)]. Some of the reasons for preferring IRT models are; obtaining more reliable results thanks to error prediction on individual level, invariant item parameters across groups, making item independent ability predictions (De Ayala & Santiago, 2017; De-Mars, 2010; Embretson & Reise, 2000). The Item Response Theory (IRT) allows individuals' ability (θ) and item parameters to be predicted by associating the individual's response to the item with the individual's level of ability and item traits (Embretson & Reise, 2000). Since trait or ability cannot be measured directly, item response theory identifies the relationship between individuals' observed performances for items and the unobservable traits or abilities that are assumed to underlie this related performance (Hambleton & Swaminathan, 1985).

Predictions in IRT can be conducted by different models. IRT models are grouped as unidimensional and multidimensional models. The unidimensional models are composed of different models based on item scoring (dichotomous and polytomous items). Models used for dichotomous scoring items are; 1, 2-, 3-, 4- parameter logistic (PL) models. These models are named according to the number of item parameters used in the function which models the relationship between the item response and individual's ability (De Mars, 2010). The possibility of a correct response to the item j for 4PLM is given in Equation 1 (Barton & Lord, 1981):

$$P(\theta_{\mathrm{j}})=c_{\mathrm{j}}+(d_{\mathrm{j}}-c_{\mathrm{j}})\,\frac{e^{Daj(\theta-bj)}}{1+\,e^{Daj(\theta-bj)}} \qquad \text{(Equation 1)}$$

P(θj) is the correct response possibility to item j for a randomly selected individual at θ ability level. "c_j" is the correct response possibility by chance, while "$\mathrm{d_j}$" is the possibility of high-ability individuals' responding wrong to an easy item due to the lack of attention. As a constant, value of *e* is 2.718 while *D* is usually taken as 1.7. Item discrimination parameter of item j is a_j; and $\mathrm{b_j}$ is the difficulty parameter of the item j. When "1" is written instead of the d_j parameter in Equation 1, the formula of 3PLM is obtained. In this formula, if the c_j parameter is taken as "0", the formula of 2PLM is obtained. In the formula of 2PLM, when the a_j parameter is taken as "same value for all items (i.e., usually with 1 at Rasch model)" and when D parameter is subtracted from the formula, the formula for 1PLM is reached.

The latent variable is assumed to be categorical in the latent class analysis (LCA), which is one of the latent class models, while there is a constant latent variable assumption in IRT (De Ayala, 2009). That is, when the observed variable is discontinuous and the latent variable is also discontinuous, LCA is used. LCA is utilized to generate homogeneous subclasses from heterogeneous latent traits (Vermunt & Magidson, 2002). In latent class analysis, it is accepted that all observed variables are the cause of a latent variable. If the latent variable is set as a control variable, the relationship between the observed variables is concluded to be conditionally independent. Under this condition, LCA is conducted to determine the latent variable which is also the control variable (Vermunt & Magidson, 2004).

The use of item response theory and latent class analysis combination brings Mixture item response theory (MixIRT) into light (Cohen & Bolt, 2005). MixIRT model is a powerful statistical method combining the LCA and IRT. Even though the concept of MixIRT has emerged with Rost in the 1980s, it is in the 2000s that it has begun to have a widespread use. The article, in which De Ayala and Santiago (2017) introduced the MixIRT and its applications, was published in 2017. It can be said that models based on MixIRT have become more widespread recently in the literature. MixIRT models (Kelderman & Macready, 1990; Maij-de Meij, Kelderman & van der Flier, 2010; Rost, 1990) have no assumptions about the type or cause of the qualitative differences in participants' responses. In the MixIRT models, latent classes (homogeneous subgroups) are defined and different parameter estimates are made between the latent classes. The MixIRT model assumes that the population consists of a limited number of latent individual, and these classes can be differentiated based on item response patterns (von Davier & Rost, 2017). On the contrary, these different response patterns will indicate themselves as differences in the parameters of the item response model related to each group. The formula for two parameter MixIRT model is as follows (Finch & French, 2012):

$$P\left(U = 1 \middle| g,\ \theta_{ig}\right) = \frac{e^{(a_{jg}(\theta_{ig}-b_{jg}))}}{1+e^{(a_{jg}(\theta_{ig}-b_{jg}))}} \qquad \text{(Equation 2)}$$

In the formula, "g: 1, 2, ..., G" indicates the latent class membership, " b_{jg}" indicates the intraclass difficulty for the item j, "a_{jg}" shows the intraclass discrimination for the item j, and " θ_{ig}" shows the level of latent trait which is measured in classroom for the individual i.

When the literature is reviewed, many studies comparing the traditional models of IRT (Rasch, 1PL, 2PL, 3PL and 4PL) have been found (Barton & Lord, 1981; Can, 2003; Erdemir, 2015; Kılıç, 1999; Loken & Rulison, 2010; Waller & Reise, 2010). Some studies (Can, 2003; Erdemir, 2015; Kılıç, 1999) indicated that 3PL or 4PL models generally fit better to data. However, it is seen in the other studies (Barton & Lord, 1981; Loken & Rulison, 2010; Waller & Reise, 2010) they are generally conducted in the field of psychology, and 4PLM has fitted better to the data in the studies conducted in recent years. Upon looking at the studies conducted for the purpose of scaling with MixIRT models; it is observed that they are employed in various studies in different subject fields such as evaluating the cognitive abilities of students (De Ayala & Santiago, 2017), analysing individual differences according to the response categories they choose in multiple choice items (Bolt, Cohen, & Wollack, 2001), interpretation of response behaviours in personality questionnaires (Maij-de Meij, Kelderman & Van der Flier, 2008), analysis of tobacco dependence in a general population survey (Muthen & Asparouhov, 2006), and scaling of a conscience scale in the context of career development (Egberink, Meijer & Veldkamp, 2010).

This study is important as it provides an application example for an exam conducted at national level regarding the use of MixIRT models. In addition, the validity and reliability of the decisions made in the exams conducted at national level are also important. Different statistical models and theories have been developed to make the most accurate predictions about the individuals' scores. In this study, results according to MixIRT are presented by trying out these models and theories. The MixIRT models allow researchers to obtain more reliable, thus more valid information about the traits of the item and group by dividing the ability of students into latent classes.

1.1. Purpose of Research

The purpose of this study is to determine the best IRT model [Rasch, 2PL, 3PL, 4PL and mixed IRT (2 and 3PL)] for the science and technology subtest of the national transition examination which is conducted for transition from basic to secondary education. It is also aimed to predict the item parameters under the best model. In this context, the questions that are sought to be answered in the study are:

1. Which IRT model (Rasch, 2PL, 3PL, 4PL and MixIRT) do TEOG 2015 science and technology subtest items fit better to?
2. What are the item parameters based on the model that fits best to data?

2. METHOD

2.1. Research Model

This study is a basic research as it aims to determine the model which fits best to the data by trying different IRT models, in other words, it contributes to the production of information necessary for test development theories.

2.2. Study Group

The study group of the research is composed of 5000 students who were randomly selected from the students participated in the Transition from Basic Education to Secondary Education (TEOG) exam in 2015. When the students' gender distribution is examined, it is seen that 48.5% (N: 2425) of the students were female and 51.5% (N: 2575) of them were male. It can be expressed that the gender rates are rather close to each other.

2.3. Data Collection Tools

The data used in this study are obtained from the application that is carried out according to the curriculum which is taught in the lessons with centralized joint exams of six core curriculum (Turkish, Mathematics, Science and Technology, Religion and Ethics, History of Revolution and Kemalism, Foreign Language). It was applied at the end of the first semester in 2015 by the Ministry of National Education. The TEOG exams, which started to be implemented in 2013, gave its place to another exam in the 2017-2018 academic year. Science and Technology subtest data consisting of 20 multiple choice items were used in this study. Because an item (item 13) was cancelled, analyses were conducted on 19 items. The data was obtained with a written permission from the Ministry of National Education (MoNE) General Directorate of Measurement and Evaluation Examination Services with the request of the researcher.

2.4. Data Analysis Procedures

Before analyzing the data, the data missing rates of the items were analysed. It was observed that they varied between 0.1% and 0.2%. The state of having extreme value of the data is examined and no extreme value is encountered. In addition, the normality of the distribution was tested, the skewness coefficient was found to be .06, and the kurtosis coefficient was -1.00. The average score of students' science and technology scores were found to be 10.62 and the standard deviations 4.51. The histogram of students' science and technology scores was also examined and seen to be in line with the normal distribution assumptions. Then, the assumptions of IRT (unidimensionality, local independence, monotone increase of the item characteristics curve and whether the test is a speed test or not) were tested (Hambleton & Swaminathan, 1985).

Confirmatory Factor Analysis (CFA) was carried out with Mplus 8 program to examine whether the nature of data-meets unidimensionality assumptions. As a result of the analyses conducted for 19 items, two items, which are items 16 and 18, were subtracted from the analysis because their factor loading values were below .30, and the CFA analysis was repeated for 17 items. As a result of the analyses, factor loadings of all items are above .30 and are statistically significant. Table 1 demonstrates the results of the analysis. When the fit indices obtained from the unidimensional model are examined, it can be expressed that the data has a good level of fit to the model ($\chi^2_{(119)}$=504.198, p<.01, χ^2/sd=4.24; RMSEA: 0.025, CFI: 0.988, TLI: 0.987).

Table 1. Results of the analysis of the unidimensional model for science and technology subtest

Items	Estimate	Standard error	Estimate/standard error
i1	0.693	0.014	50.996*
i2	0.619	0.014	43.977*
i3	0.555	0.017	32.673*
i4	0.556	0.015	36.628*
i5	0.636	0.014	46.973*
i6	0.504	0.018	28.017*
i7	0.667	0.015	45.340*
i8	0.705	0.012	57.694*
i9	0.488	0.017	29.403*
i10	0.625	0.014	44.942*
i11	0.742	0.012	59.400*
i12	0.723	0.012	59.714*
i13	0.658	0.014	47.823*
i14	0.674	0.013	51.330*
i15	0.672	0.013	50.120*
i16	0.571	0.015	37.171*
i17	0.383	0.018	21.716*

*p< .05

As it can be seen in Table 1, the factor loadings of the items vary between .383 [item17 (i17)] and .742 (i11), and all items appear to make significant contribution to the unidimensional model.

Yen's Q3 statistics was used to examine whether the data validate the local independence assumptions. Although the local independence assumption is stated to be met as well in the case of the unidimensionality assumption (Hambleton & Swaminathan, 1985), the Q3 statistics which is frequently used in testing the local independence is also calculated. The calculations are carried out based on examining the correlations between items under the four different models (Rasch, 2PL, 3PL and 4PL). Q3 statistics are calculated for each model in R with the help of "sirt" package (Robitzsch, 2015). In all models, the correlations between the items were found to be -0.127 (the lowest) and 0.042 (the highest). It can be stated that the local independence assumption is met as the values calculated are less than .20 (DeMars, 2010). Item characteristic curves (ICCs), were examined for four models to see the monotonic increase of the item characteristic curve. The ICCs are drawn for each model in R with the help of the "sirt" package (Robitzsch, 2015), and are presented in Figures 1 and 2.

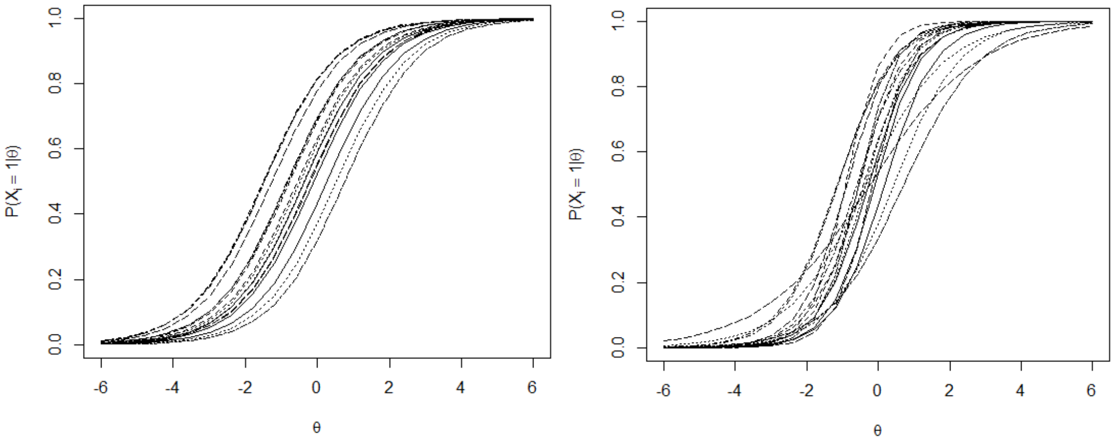


Figure 1. ICCs according to 1PL and 2PL models, respectively

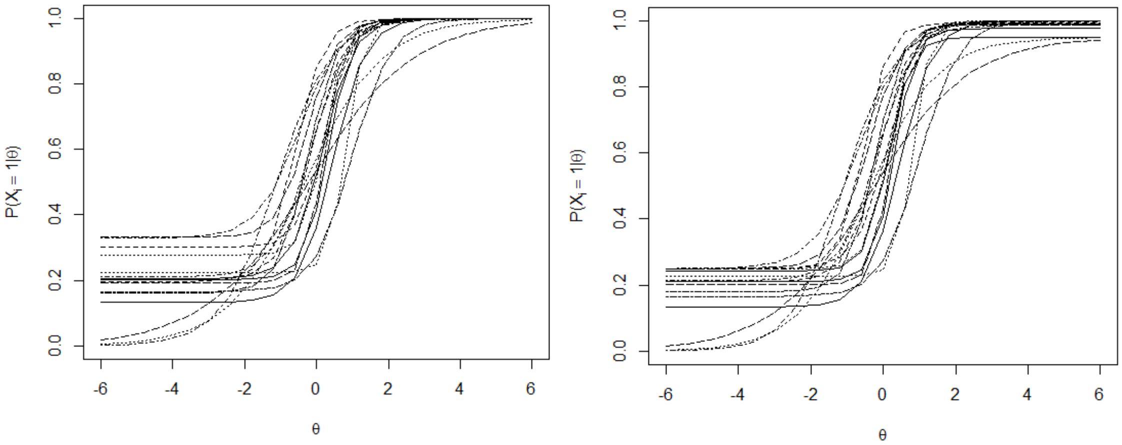
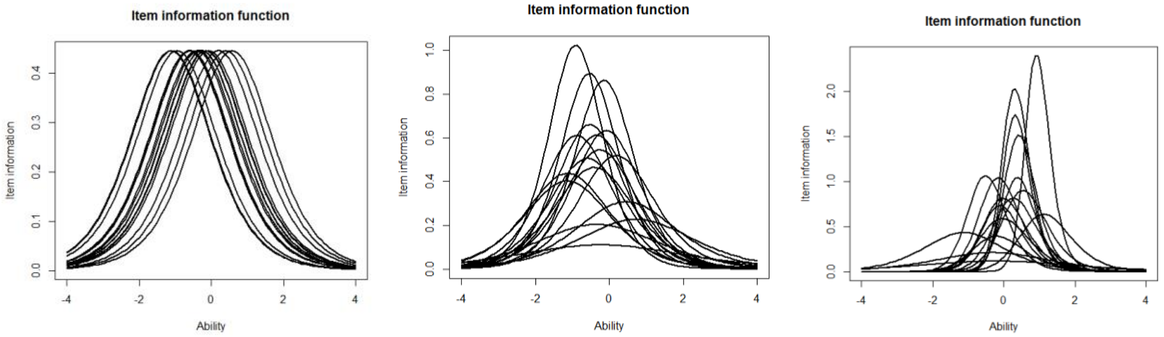


Figure 2. ICCs according to 3PL and 4PL models, respectively

As can be seen in Figures 1 and 2, the probability of correct response to an item increases as the level of the individual's ability increases in the four models, that is, the item characteristic curves increase monotonically.

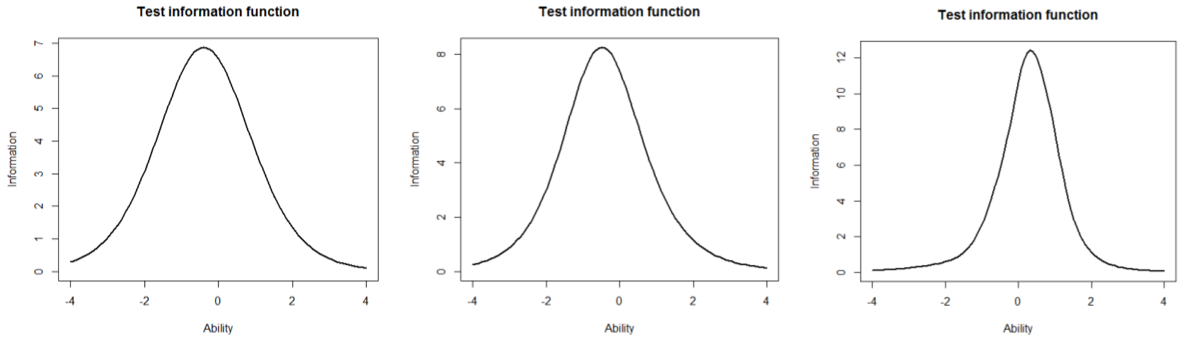
In order to determine whether the test is a speed test, the variance of number omitted items was divided by the variance of the number of incorrectly answered items. The rate found was near zero, and the test is accepted not to be a speed test (Hambleton & Swaminathan, 1985). Moreover, the rate of responding the items correctly is also examined and it is seen that it varied between .36 (item6) and .75 (item17), and that the rates of responding to the final items correctly are similar to those of other items.

Item and test information graphics based on 1, 2 and 3 PL models related to reliability were created. The graphs for items, test information values and functions are calculated and drew in R with the help of the "irtoys" package (Partchev, 2017). Item and test information functions according to three models are given in Figures 3 and 4. Since there is no package which calculates the information function according to the 4PL model, it could not be drew.



1PL model 2PL model 3PL model

Figure 3. Item information functions for three models



1PL model 2PL model 3PL model

Figure 4. Test information functions for three models

As it can be seen in the Figures 3 and 4, predictions under 3 PLM provided the most information for a higher ability group than other models. The model that provides information for the largest ability level is the 1PL model, which is also the one with the least information.

As a results of the examinations, it is concluded that the 17-item science and technology sub-test meets the IRT's assumption. Analyses were conducted according to four models (2-, 3-, 4 PLs and mixture-IRT) to determine which model fits the data better to, in other words to find an answer to the first research question presented above. Estimates were made for 2-, 3-, 4 PL and mixture-IRT in Mplus 8 program (Muthén & Muthén, 2017). The Bayesian information criterion (BIC) value which is recommended in the literature to determine the model data fit (Li, Cohen, Kim & Cho, 2009) and -2 log $\chi 2$ values of the models (Hambleton et al., 1991) is used for comparisons. Then, for the second research question, the parameter values of the fitting model are presented and interpreted.

3. FINDINGS

Analysis which were conducted to determine the most appropriate IRT model for TEOG 2015 science and technology subtest data resulted some model fit indices to be discussed. Some indices such as likelihood- (LL), the degree of freedom (df), BIC and Akaike Information Criterion (AIC) are presented in Table 2.

Table 2. Model data fit results based on models

Models	**LL**	**df**	**BIC**	**AIC**
2PL	-48112.103	34	96513.790	96292.206
3PL	-47744.809	51	95923.994	95591.617
4PL	-47773.491	68	96126.150	95682.981
MixIRT (2PL) 1-Latent Class	-48112.110	34	96513.804	96292.220
MixIRT (2PL) 2- Latent Class	-47757.030	53	95965.471	95620.060
MixIRT (2PL) 3- Latent Class	**-47649.129**	**72**	**95911.496**	**95442.258**
MixIRT (2PL) 4- Latent Class	-47599.948	91	95974.961	95381.896
MixIRT (3PL) 2- Latent Class	-47643.375	86	96019.228	95458.749
MixIRT (3PL) 3- Latent Class	-47588.756	121	96208.093	95419.512

As it can be seen in Table 2, when traditional IRT models (2, 3 and 4PL) are examined solely with the LL, BIC and AIC values, the model that fits best is the three-parameter model. When predictions are made with MixIRT models, the model that best fits the data according to the BIC value, which is the best indicator of model data fit, is the model predicted according to MixIRT with three latent classes (3LC) and two parameters. When deciding on the model data fit, together with taking the BIC value under consideration, -2 log $\chi 2$ values can be compared. In this context, Chi-Square statistics, the degree of freedom and the difference between the values of -2 log $\chi 2$ belonging to the 2- and 3PL models were evaluated at first. Since the calculated value ($\chi 2$ = 48112.103-47744.809 = 367.294) is greater than the table value ($\chi 2_{(17;\, 0.05)}$ = 27.857), the difference between -2 log $\chi 2$ values is significant. In this case, it can be said that the three-parameter model is more suitable for data. Then, the Chi-Square statistics, the degree of freedom and the difference of the -2 log $\chi 2$ values belonging to the 4PL and 3PL models are evaluated. Since the calculated value ($\chi 2$ = 47773.491- 47744.809 = 28.682) is greater than the table value ($\chi 2_{(17;\, 0.05)}$ = 27.857), the difference between -2 log $\chi 2$ values is significant. In this case, it can be stated that the three parameter model for the data is more suitable than the four parameter model. When compared to the model with the lowest BIC value among MixIRT models, since the calculated value ($\chi 2$ = 47744.809-47649.129 = 95.68) is greater than the table value ($\chi 2_{(21;\, 0.05)}$ = 32.671), the difference between the values of -2 log $\chi 2$ is significant. In this case, it is stated that the two parameter MixIRT model with three latent classes is more suitable for the data. The results of the two-parameter MixIRT model with three latent classes are given in Table 3 in order to present the item parameters [(item discrimination (a) and item difficulty (b)] according to the model which fits best to the data.

As shown in Table 3, 37% (N: 1868) of the students are in the first latent class, 37% (N: 1848) of the students are in the second latent class and 26% (N: 1284) of the students are in the third latent class. When the gender distribution of the students in latent classes is examined, it is seen that the ratio of the students in terms of gender in all the latent classes is very close. When item-model fit is evaluated, it is indicated that the difficulty values of one item (i6) in the first latent class, three items (i2, i11 and i13) in the second latent class and two items (i4 and i16) in the third latent class do not fit to the model. It is thought that the reason why different items in different latent classes do not fit the data is resulted from the different traits individuals carry in the latent classes. Within the scope of this research, the emerged latent classes could not be interpreted in more details because information obtained from MoNE is limited to individual responses for items and their gender.

In latent classes, the item discrimination averages are respectively; 1.70, 0.77 and 0.27. It is observed that discrimination decreases from the latent class-1 to the latent class-3. Item difficulty averages in latent classes are respectively; 1.33, -0.79 and 4.20. In this context, it can be expressed that items are moderately difficult and discriminative for students in latent class-1; the items are considerably easy and able to slightly distinguish the students in latent class-2; the items are difficult to the students in the third latent class and they can slightly distinguish the students in this group.

Table 3. Item parameters in each model for 2PLM with three latent classes

	LC1		LC2		LC3	
Gender	Frequency	%	Frequency	%	Frequency	%
Female	872	49	879	48	674	49
Male	921	51	953	52	701	51
Total	1868	37	1848	37	1284	26
Items	a	b	a	b	a	b
i1	1.946	1.010	0.930	0.366	0.306	3.628
i2	1.454	1.732	0.695	-0.029*	0.229	3.149
i3	1.285	0.757	0.614	2.315	0.202	5.232
i4	0.922	3.512	0.441	-2.675	0.145	1.093*
i5	1.574	1.949	0.752	-1.632	0.247	1.612
i6	1.070	0.136*	0.511	1.854	0.168	8.284
i7	1.512	0.734	0.722	0.736	0.238	6.439
i8	2.116	1.753	1.011	-1.939	0.332	1.427
i9	0.731	1.151	0.349	-1.973	0.115	9.767
i10	1.296	1.753	0.619	-1.129	0.204	3.878
i11	2.328	1.062	1.112	0.186*	0.366	3.278
i12	2.675	0.968	1.278	-0.991	0.420	1.899
i13	1.840	0.969	0.879	-0.199*	0.289	2.978
i14	2.059	1.071	0.984	-0.947	0.324	2.277
i15	3.306	0.660	1.579	-0.327	0.519	1.027
i16	1.471	1.490	0.703	-3.027	0.231	0.630*
i17	0.665	0.669	0.318	-1.716	0.105	8.131

*p>.05

Item discrimination values of the items in the first latent class vary between 0.665 (i17) and 3.306 (i15), and the item difficulty values range from 0.660 (i15) to 3.512 (i4). The item discrimination values of the items in the second latent class are between 0.318 (i17) and 1.579 (i15), and the item difficulty values range from -0.327 (i16) to 2.315 (i3). Item discrimination values of the items in the third latent class are between 0.105 (i17) and 0.519 (i15), and the item difficulty values range from 1.027 (i15) to 9.767 (i9). When the difficulty range of items is examined in three latent classes, it is seen that the vast majority of the items in the second latent class have negative difficulty value. In this context, it can be expressed that the items are easier for the students in this group. Yet, in the third latent class, it is seen that the difficulty values of the items increase. This situation makes it possible to state that the items are difficult for the students in this group.

The item with the lowest discrimination value in all three latent classes is item-17, which is the last item in the test. This item is a question asking the relationship between the weight of the objects and the lifting force applied to objects which are in status of swimming and sinking. When the students' response distribution to the choices for this item is examined, 75% of the students have marked the wrong "C" option. Only 6% of students responded correctly to this item. However, when the difficulty levels of the item in the latent classes are examined, it is seen that this is an easy item for the students in the second latent class. This is also constituting as an example of the change of item parameters according to MixIRT in latent groups.

Furthermore, item-15 is the item that has the highest discriminate value in all three latent classes. At the same time, this item has the lowest difficulty value in three latent classes. This item corresponds to the item-17 in the original test (since item-13 is cancelled, and item-16 is excluded from the analysis). When the students' response distribution for this item is examined, it is observed that 60% of the students marked the option "B" which is one of the wrong choices. When the relevant question and the "B" option are examined, it is seen that 60% of the students turn towards the wrong conceptual knowledge that the intensity of an object in swimming state is equal to that of the liquid it is in. This has led to the increase in the item discrimination, and for the item to have a difficult trait.

In the first latent class, the fourth item which was observed to have the highest item difficulty value (b=3.512) was determined to be considerably easy. However, in the second latent class, it had the lowest difficulty value (b=-2.675), which means it was a difficult item. In the third latent class, on the other hand, this item showed no significant fit with the model tested. According to the CTT, the item difficulty of this item is .75. In this case, it can be stated that an item which seems quite easy according to the CTT could be a difficult question for students in some latent groups. More detailed studies should be conducted as to why this problem prepared on the subject of "genetic crossing" has been identified as difficult in the second latent class. Findings can be interpreted for all items similarly. In this study, only a few remarkable items have been interpreted.

4. DISCUSSION AND CONCLUSION

The aim of this study is to determine to which IRT models [(Rasch, 2PL, 3PL, 4PL and mixed-IRT (2 and 3PL)] TEOG 2015 science and technology sub-test conducted at national level fits best. In addition, it is also aimed to predict item parameters for the model that fits best. For this purpose, before analysing the data, the assumptions of IRT (unidimensionality, local independence, monotone increase of the item characteristics curve and whether the test is a speed test) are tested and all assumptions were seen to be met. Predictions are made for the 2-, 3- and 4-PL and the 2- and 3-PL models according to the MixIRT in order to determine the model that fits data best. Then the item parameters are predicted for three latent classes separately according to MixIRT model with two parameters and three latent classes which is also fits the model the best. When the gender distribution of the students in latent classes is examined, it is seen that the ratio of the students based on gender in all the latent classes is very close. When items' fitting to the model is evaluated, difficulty value of one item in the first latent class, three items in the second latent class, and two items in the third latent class do not fit to the data. When the difficulty range of the items and the difficulty averages in all three latent classes are examined, it is seen that the vast majority of the items in the second latent class have negative difficulty value. In this context, it can be stated that the items are easier for the students in this latent class than it is for the first and third class. The difficulty values of the items are seen to increase in the third latent class. For this case, it can be stated that the items are difficult for the students in this group. This finding is consistent with findings obtained from the study of De Ayala and Santiago (2017) in which the MixIRT models are tested with the mathematical abilities of students in the 1-3 latent class according to the 1PL model. According to the fitting model, it is determined that some of the items have been found easier by those in a latent class while harder for some others in this study as well.

When the discriminate values of the items are examined, it is seen that the highest discriminate values are in the latent class one. Considering the difficulty and discrimination averages of the items in the latent classes, it can be expressed that items are of moderate difficulty and discriminative for students in latent class-1; the items are considerably easy and able to distinguish the individuals a little for the students in the latent class-2; the items are difficult to the students in the third latent class and they can distinguish the students in this group a little. When the results are evaluated in general, students who have the lowest science literacy are most probably in the LC-3 (students with the lowest science achievements). Students who have science literacies at the highest level are most probably in LC-2 (students with the highest science achievements). Furthermore, students who have science literacies at the moderate level are most probably in LC-1 (students with the intermediate science achievements). In this context, it is recommended to carry out studies in which many variables such as school, teacher and student characteristics are discussed together in order to be able to put forward the profiles of the students in the emerged latent classes.

The item with the lowest discrimination value in all three latent classes is item 17, which is the last item in the test. This item seems to be an easy item for students in the second latent class when the difficulty values of this item in the latent classes are examined. This is also an example of the change of the item parameters according to MixIRT in latent groups. The item 15, is the item that has the highest discrimination value in all three latent classes. This item also has the lowest difficulty value in all three latent classes. It is observed that this item, which seems quite easy according to the CTT, could be a difficult item for students in some latent groups.

MixIRT is based on the assumption that the sample consists of latent subclasses. Different from IRT, MixIRT does not assume that the item parameters remain invariant among the groups. It is flexible on this subject and it allows the change of item parameters between the latent classes (De Ayala & Santiago, 2017). Separating students' ability into latent classes allows researchers to obtain more reliable thus more valid information about item and group characteristics. In addition, MixIRT approach also enables modelling both continuous and categorical data at the same time and this makes it possible to gather more information (De Ayala & Santiago, 2017).

In order for the estimates made to be less inaccurate, different models based on theories must be discussed in the analysis of the data in the exams that are conducted at the national level and for the purpose of selection and placing of the students to a secondary education institution. This research is the first study to compare the model fit data according to MixIRT models for a national test in Turkey. In this context, it is important to support the results obtained with the studies to be made on this subject with different subtests in different years.

The conducted study also has some limitations. First of all, only the TEOG 2015 application science and technology subtest has been dealt with in the study. Interested researchers can test the model-data fit for the data of different subtests in different years. Moreover, dichotomous data were studied in this study. Interested researchers can compare the results by using MixIRT in polytomous items with traditional models. Finally, no constraints have been identified while making parameter predictions for the IRT. Researchers who are interested in studying on this subject can examine the model data fit by setting constraints for item parameters.

ORCID

Seher Yalçın https://orcid.org/0000-0003-0177-6727

5. REFERENCES

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin*, 81-20.

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.

Can, S. (2003). *The analyses of secondary education institutions student selection and placement test's verbal section with respect to item response theory models* (Unpublished Master's thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. doi: 10.1111/j.1745-3984.2005.00007

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

De Ayala, R. J. & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, 60, 25-40. doi: 10.1016/j.jsp.2016.01.002

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Egberink, I. J., Meijer, R. R. & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, 44, 232–244.

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

Erdemir, A. (2015). *Bir, iki, üç ve dört parametreli lojistik madde tepki kuramı modellerinin karşılaştırılması (Comparison of 1PL, 2PL, 3PL and 4PL item response theory models)* (Unpublished Master's thesis). Gazi University, Graduate School of Educational Sciences, Ankara.

Finch, W. H. & French, B. F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 167-178. doi: 10.22237/jmasm/1335845580

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and application.* Boston: Kluwer Academic Publishers Group.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications Inc.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*(4), 307–327.

Kılıç, İ. (1999). *The fit of one- two- and three- parameter models of item response theory to the student selection test of the student selection and placement center* (Unpublished doctoral dissertation). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373. doi: 10.1177/0146621608326422.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525. doi:10.1348/000711009X474502

Maij-de Meij, A. M., Kelderman, H., & Van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, 32, 611–631.

Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975-999. doi:10.1080/00273171.2010.533047

Muthen, B. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, 31, 1050-1066.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Partchev, I. (2017). *Simple interface to the estimation and plotting of IRT Models*. R-project, Package 'irtoys' manual. Retrieved from https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Robitzsch, A. (2018). *Supplementary item response theory models*. Retrieved from https://cran.r-project.org/web/packages/sirt/sirt.pdf

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

Skrondal, A. & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi: 10.1111/j.1467-9469.2007.00573.x

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars, & A. L. McCutcheon, *Applied latent class analysis* (p. 89-107). New York: Cambridge University Press.

Vermunt, J. K., & Madigson, J. (2004). Local independence. In A. B. M. S. Lewis Beck (Ed.), *Encyclopedia of social sciences research methods* (pp. 732-733). Thousand Oaks: Sage Publications.

von Davier, M. & Rost, J. (2017). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (p. 393-406). Boca Raton: Chapman and Hall/CRC.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard item response theory models: Fitting the four-parameter model to the minnesota multiphasic personality inventory. In Embretson, S. (Ed), *New directions in psychological measurement with model-based approaches* (p. 147-173). Washington DC: American Psychological Association.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 314–324

DOI: 10.21449/ijate.421167

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

The Use of Three-Option Multiple Choice Items for Classroom Assessment

Erkan Hasan Atalmış [1*]

[1]Kahraman Sutcu Imam University, Faculty of Education, Department of Educational Measurement and Evaluation, Kahramanmaras, Turkey

Abstract: Although multiple-choice items (MCIs) are widely used for classroom assessment, designing MCIs with sufficient number of plausible distracters is very challenging for teachers. In this regard, previous empirical studies reveal that using three-option MCIs provides various advantages when compared to four-option MCIs due to less preparation and administration time. This study examines how different elimination methods; namely, the least selected and the random methods, influence item difficulty, item discrimination and test reliability on decreasing the number of options in MCIs from four to three. The research findings have revealed that the concerning methods did not affect item difficulty, item discrimination, and test reliability negatively. Results are discussed in relation to promoting quality classroom assessment.

ARTICLE HISTORY

Received: 01 January 2018

Revised: 23 April 2018

Accepted: 30 April 2018

KEYWORDS

Classroom Assessment,
Item-Writing Guidelines,
Number of Options,
Multiple-Choice Items,
Test Quality

1. INTRODUCTION

Classroom assessment is an indispensable period of education and training. To what extent the goals and behaviors that students need to gain during the semester has been determined and how much teachers teach what they think they are teaching has been presented through classroom assessment. Therefore, it is of high importance for teachers to carry out an effective in-class assessment, and teachers are required to spend a significant part of their professional work life in classroom assessment studies (Darling-Hammond & Youngs, 2002; Stiggins, 1991). Upon examining the related literature, the significance of in-class assessment was revealed and various recommendations were presented in this context. Among these recommendations are that paper-pencil tests which are the mostly used method of classroom assessment should be prepared by the teachers themselves (Frey & Schmitt, 2010). This allows the assessment tool be consistent and compatible with the class activities as the measurement tool.

Multiple-choice items (MCIs) are one of the most commonly used item type in classroom assessment (Haladyna & Rodriguez, 2013). When previous studies were analyzed, both theoretical and empirical studies regarding reliability and validity of these item types were

CONTACT: Erkan Hasan Atalmış ✉ erkanatalmis@gmail.com Kahraman Sutcu Imam University, Faculty of Education, Department of Educational Measurement and Evaluation, Kahramanmaras, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

conducted and these were determined to be more reliable and valid than particularly open-ended items (Collins, 2006; Tarrant, Knierim, Hayes, & Ware, 2006; Thorndike, 2005). However, the studies emphasized the challenges of preparing the appropriate number of rational choices for MCIs, so they developed alternative ways related to MCIs.

One of these alternative methods has been considered as a reduction of the number of options. Although various studies revealed that reducing the number of options from 4 to 3 does not have a negative effect upon test reliability and item discrimination (Atalmis & Kingston, 2017; Delgado & Prieto, 1998), no consensus has been reached so far on the comparison of three-option and four-option items in terms of item difficulty. That is, it could not be exclusively argued that one type is more difficult than the other in all circumstances. Even though Rodriguez (2005) suggests that the number of options in MCIs may result from different methods used to reduce the number of options from 4 to 3, this is not revealed empirically.

In this regard, whether different methods used in reducing the number of options from 4 to 3 has an impact upon test reliability, item discrimination and item difficulty will be empirically examined and thus the use of 3 option items in the classroom assessment is thought to provide a new path.

1.1. Classroom assessment activities (Assessment Criteria)

The quality of classroom activities was discussed by educators and researchers as classroom assessment activities play a significant role in improving the outputs of the training. In this sense, researchers emphasized that classroom assessment activities should aim at increasing the quality of learning in the classroom, rather than largely through the traditional sense of passing and failing the exams (Chappuis & Stiggings, 2002; Leahy, Lyon, Thompson, & Wiliam, 2005). Hence, classroom assessment must have the ability to answer questions such as how well learners are learning and how effectively teachers teach (Angelo & Cross, 2001). The most important way to achieve this is to use classroom assessment methods that provide accurate and descriptive feedback to students and teachers about learning and teaching activities in the classroom. This is only possible with reliable, valid and useful measuring tools.

Reliability is defined as the accuracy or precision of measurement procedure and so it is the degree to which measurement are free from error (AERA, APA, & NCME, 2014; Thorndike, 2005). Errors can arise either from the measurement tool, the measured characteristic, and the person who measure or from the environment. In this context, test reliability is negatively influenced by such factors as incorrectly responded questions whose answers are known to the students, involvement of guessing factor, subjective evaluation of teachers, testing environment, and cheating. Thus, the fact that tests used in the classroom are mostly composed of more questions, objectively scored and sensitive in selecting the test environment will increase the test reliability.

Validity is the test quality that indicates the degree to which a measuring instrument measures the desired property (AERA, APA, & NCME, 2014; Haladyna & Rodriguez, 2013). Hence, the validity of a measurement tool is measured through different features, such as content-related validity, construct validity and criterion-related validity. Content-related validity is about how much the test covers the features desired to measure (Thorndike, 2005). To illustrate, the extent to which a test prepared in a mathematics class covers the acquisitions of the unit that is to be measured relates to content-related validity. In this respect, more question-based testing also increases content-related validity just as test reliability. The construct validity refers to the fact that the construct to be measured is measured without any other mixing (Messick, 1989). For instance, if a test would only measure students' mathematical skills, this test would violate the validity of the test when it involves such skills that include mathematics questions including reading and attention skills. Criterion-related validity is the relationship between a test and another test (Thorndike, 2005). To exemplify, if the paper-pencil test to measure students' math skills helps solve the mathematical problems experienced by the person in real life, then the paper-pencil test's criterion-referenced validity might be strong.

Along with reliability and validity, another feature of the measurement tool is practicality which is defined as the economical and easy development, application and scoring of a test (Thorndike, 2005). For example, in a 20-person class, when a test type is evaluated for each student's lesson time, paper-pencil tests seem to be more useful than performance tests in both applying and scoring process.

1.2. Multiple-Choice Items

Multiple-choice items (MCIs) are widely used paper-pencil test to construct objective tests, which are considered as quickly and unambiguously scored, and minimize test administration and scoring time (Haladyna, Downing, & Rodriguez, 2002; Haladyna & Rodriguez, 2013). However, constructing these items consisting of plausible distractors and measuring desired objectives is challenging for item writers (Collins, 2006; Haladyna et al., 2002). Item preparation is also called as "a creative art". (Rodriguez, 1997). In this regard, item-writing guidelines which are supposed to help design items systematically with the aim of increasing validity evidence for the test have been reported in previous studies. Validity evidence is a scientific notion used to describe how to develop tests accurately and how to predict, evaluate, and interpret the test scores (AERA, APA, & NCME, 2014).

To date, a limited number of studies focused on item-writing guidelines for item and test construction. One of the pioneering ones in this concern was conducted by Haladyna and Downing (1989) who proposed 43 item-writing guidelines after reviewing textbooks published in the field of measurement and evaluation. Approximately two decades later, Haladyna et al. (2002) redesigned the existing version identifying 31 valid item-writing guidelines mainly for classroom assessment, and classified them into five categories: content, formatting, style, forming the stem, and forming the choices. Several years later, Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) evaluated twenty classroom assessment textbooks and identified 40 most commonly used item writing guidelines. They also classified them depending on validity concerns, (i.e., potentially confusing wording or ambiguous requirements, guessing, rules addressing test-taking efficiency, and rules designed to control for test wiseness). Moreno, Martínez, and Muñiz (2006) designed a condensed version of the existing item writing guidelines and classified them into three groups with a focus on foundations, the expression of the domain and context in each item and test, and on response options. Likewise, the same researchers decreased the number of the guidelines to 9 and evaluated them according to the definition of validity (Moreno, Martinez, & Muniz, 2014). One of the common characteristics of item writing guidelines proposed by previous studies was that each study emphasized the construction of a sufficient number of options with plausible distractors for each item since one of the challenging part of the item-writing process of MCIs is to construct a sufficient number of plausible distracters (Haladyna et al., 2002). Plausible distractors are developed based on students' particular errors at some point in analyzing and solving the problem (Thorndike, 2005). Thus, MCI writers should have deep pedagogical content knowledge and teaching experience. This results in decreasing the probability of using more options in MCIs, such as the use of three options rather than four options.

1.3. Comparing Four Option with Three Option MCIs

Extant empirical studies concluded that using three-option MCIs provides various advantages compared to four-option MCIs due to less preparation and administration time (Balta & Eryılmaz, 2017; Haladyna & Downing, 1989; Haladyna et al., 2002; Rich & Johanson, 1990). Empirical studies have examined how item (test) difficulty, item discrimination and test reliability vary across 4-choice items and 3 choice items. Item difficulty is defined as the proportion of students who choose the correct answer while item discrimination is defined as how well the item differentiates students with high ability in the construct of interest from students with low ability. Test reliability, as mentioned in the introduction section, is defined as the consistency of test results.

The studies have found opposite results regarding item difficulty. Some studies found that item difficulty was not statistically different between four-option items and three-option items (Abad, Olea & Ponsoda, 2001; Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Shizuka, Takeuchi, Yashima & Yoshizawa, 2006), whereas others concluded that MCIs with three options were statistically more difficult than MCIs with four options, which is counterintuitive (Landrum, Cashin, & Theis, 1993; Rogers & Harley, 1999). Rodriguez (2005) conducted a meta-analysis and examined 48 empirical studies from 1925 to 1999 in order to uncover the effect of the number of options upon psychometric characteristics of MCIs. Of these 48 studies related to achievement and aptitude tests, 27 studies included pertinent results. The results supported that three-option items were slightly easier than four-option items.

Considering studies on item discrimination, item discrimination between MCIs with four options and items with three options was not statistically different in most studies (Atalmis & Kingston, 2017; Cizek & O'Day, 1994; Crehan, Haladyna, & Brewer, 1993; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Rogers & Harley, 1999; Shizuka et al., 2006; Tarrant & Ware, 2010). Yet, some studies provided statistically significant evidence that item discrimination for MCIs with three options was higher than that of MCIs with four options (Baghei & Amrahi, 2011; Landrum et al., 1993; Rodriguez, 2005; Trevisan, Sax, & Michael, 1991). Consequently, the literature reveals that three-option items do not affect negatively item discrimination.

A limited number of studies on test reliability have indicated that the number of options did not have a statistically significant impact on test reliability (Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1999) while some found that test reliability increased when the forms with three options were employed (Rodriquez, 2005; Tarrant & Ware, 2010).

1.4. Significance of the Study and Research Questions

Although previous studies revealed that reducing number of options from 4 to 3 did not have a negative effect upon test reliability and item discrimination, no consensus has been reached so far on the comparison of three-option and four-option items in terms of item difficulty. Even though Rodriguez (2005) suggests that the number of options in MCIs may result from different methods used to reduce the number of options from 4 to 3, this has not been revealed empirically. Given previous studies, these used different traditional methods to eliminate one of four-option to construct three-option MCIs, such as eliminating the least selected option or a random option. However, they did not consistently investigate the impact of elimination method on item and test characteristics. Therefore, this research aims to examine how item and test psychometric characteristics vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs. More specifically, we have examined three research questions as follows:

- Does item difficulty for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?
- Does test reliability for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?

- Does item discrimination for mathematics items vary when different elimination methods are applied to reduce the number of options for four-option MCIs to three-option MCIs?

2. METHOD

This section covers data collection, participants, instrument development, and data analysis.

2.1. Data Collection

Data collection procedure includes several phases. First, test forms including parallel four-option MCIs were designed, and each form was administered to 7th grade students in state primary schools located in Turkey for the pilot study. After calculating item psychometrics characteristics of each item, 20 of them were selected to be used in the final version of the instrument. Subsequently, two forms (Form B1 and B2) were designed with parallel items on each, one of the options of MCIs in each form was eliminated by using the least selected option in Form B1 and random option in Form B2. After two forms were administered to 7th and 8th grade students in Turkey, data analysis was conducted.

2.2. Participants

This research was carried out with 7th and 8th grade students in the pilot study and the main study in Turkey. Convenience sampling method was applied for piloting and the ultimate phase. The pilot test was administered to 1130 students attending sixteen state primary schools in the province of Manisa, Turkey. The final test was administered to 847 students who enrolled in eight schools in the provinces of Manisa and Kahramanmaras, Turkey. In both phases, only students’ responses to mathematics items were collected without their academic proficiency and demographic features.

2.3. Instrument Development

Instrument development procedure of this study includes several phases. First, we developed an item pool including a large number of items representing the full range of the objectives of the mathematics topic, “equation and expression”. Hence, after 58 MC mathematics items were developed, two forms (Form A1 and Form A2) composing of 29 items on each were constructed so that the examinees could answer all of them during a class period. After Form A1 and Form A2 was respectively administered to 474 and 656 students attending sixteen public schools in the province of Manisa, Turkey, item difficulty and item discrimination were computed for each item in Form A1 and Form A2. Item difficulty is defined as the proportion of the students choosing correct answers while item discrimination shows how well the item discriminates between students with high ability and low ability (Thorndike, 2005). Item-total correlation index, one of most widely used method, was used to calculate item discrimination for each item (Downing, 2005). The findings showed that item difficulty indexes for the items range between .19 and .74, while item discrimination indexes range between .25 and .65.

Second, 20 out of 58 MCIs which were higher quality (higher discrimination index and middle item difficulty) were selected and designed to be used in the final version of the instrument. Namely, items with discrimination index of .30 or greater than .30 and three functioning distractors were selected (Field, 2009). Previous studies proposed that functioning distractors were plausible distractors chosen by at least 5% of examinees (Haladyna & Dowing, 1993; Rodriguez, Kettler, & Elliott, 2014).

Third, two forms (Form B1 and B2) were redesigned with 10 parallel items measuring the same specific learning standards and objectives on each form. Moreover, parallel items were constructed with the same content and rationale of distractors, but the numbers were different. Due to a small number of items in each form (*i.e.*10 items in each group), nonparametric tests were preferred by using item difficulty and item discrimination values of items on each form to show the equality of Form B1 and Form B2. A Mann-Whithey U Test was conducted; accordingly, it was found that average rank of item difficulty did not statistically differ between Form B1 (*z*=-1.17, *p*=.24). To test two item discrimination index, item-total correlation was used and then Fisher-Z transformation method is carried out to make correlation values normalized, as follows:

$$z = \frac{1}{2}\ln(\frac{1+r}{1-r}) \qquad (1)$$

where r is item-total correlation for each item and z is the transformed value of r. The findings showed item discrimination did not statistically vary across Form B1 and B2 (*z*=-.19, *p*=.85).

Fourth, one of the options of each of the MCIs in each form was eliminated via the least selected option in Form B1 and random method in Form B2. Table 1 reveals distractor selection frequencies, proportion of students choosing options of each item, and eliminated distractors in Form B1 and Form B2.

Table 1. *Distractor Selection Frequencies in Form B1 and Form B2*

Item ID	Form B1				Item ID	Form B2			
	a	b	c	d		a	b	c	d
1	.25	.15	.53*	.07**	11	.27	.14**	.54*	.05
2	.12	.10**	.42	.36*	12	.52*	.17	.17	.13**
3	.11**	.23	.19	.47*	13	.13**	.07	.44	.36*
4	.40*	.27	.21	.12**	14	.16	.53*	.17	.14**
5	.19	.50*	.18	.13**	15	.15	.20**	.53*	.11
6	.54*	.15	.11**	.20	16	.55*	.13	.12	.20**
7	.54*	.12**	.15	.19	17	.10	.13	.44**	.33*
8	.26	.31*	.09**	.34	18	.27**	.34*	.14	.25
9	.23	.39*	.15**	.23	19	.22**	.25	.34*	.19
10	.26	.27	.28*	.18**	20	.23	.19**	.30*	.28

* key; **eliminated distractor

Finally, the test consisting of all items was designed to counterbalance the order of MCIs in order to avoid systematic errors, entailing that some students were supposed to take the test with Form B1 followed by Form B2, while others take the tests in concern in the reverse order.

2.4. Data Analysis

In order to examine whether different elimination methods have a significant impact on psychometrics characteristics, item difficulty and discrimination were calculated for each item, and test reliability of forms (Form B1 and Form B2) was measured individually and together by using item response theory (IRT). Regarding item difficulty and item discrimination, two parameter logistic (2PL) model and three-parameter logistic (3PL) model data fit statistics were calculated for 20 items by employing IRTPRO 4.2. After likelihood ratio test was conducted to compare the models, it was found that the values of the -2 log likelihoods (i.e., -2ln*L*) for the 2 PL and 3 PL models were 19577.35 and 19306.99, respectively. The difference between these values was found as 270.36 with 19 degree of freedom, which is statistically significant.

Therefore, using 3PL model represents a statistically significant improvement in fit over the 2PL model.

Item difficulty, *b* parameter, shows the position of the Item Curve Characteristics (ICC) regarding the ability scale and item difficulty index ranges between -3 and +3. Easy items are located somewhat below 0 while difficult items are located somewhat above 0 (De Ayala, 2013). Item discrimination, *a* parameter, reveals the proportion to the slope of the ICC at the *b* point on the ability scale (Hambleton, Swaminathan, & Rogers, 1991). Items with higher *a* values are demanded because these items discriminate well among examinees. In this study, after item difficulty and item discrimination values of items on each form were calculated, nonparametric tests were conducted to compare Form B1 and Form B2.

Test reliability is defined as internal consistency of the test and it can be calculated using coefficient alpha (α) estimation method in Classical Test Theory (CTT). Subsequent to calculating the standard error of estimate (SEE) for each reliability coefficient value to obtain the 95% confidence interval (Duhachek & Iacobucci, 2004; Van Zyl, Neudecker, & Nel, 2000), it is examined whether the reliability coefficient is statistically different from one sub-test to another. However, calculating test reliability depends on the particular set of items in CTT. This is a disadvantage of CTT over IRT since each item contribute test reliability individually and independently. Therefore, test reliability is also calculated using IRT as well as CTT in this study.

In summary, item difficulty, item discrimination, and test reliability are important criteria to calculate psychometrics characteristics. The following section shows the results of item difficulty and item discrimination indexes for each item, and test reliability indexes for each form.

3. FINDINGS

This section provides the results of item difficulty and item discrimination, and test reliability, respectively.

3.1. Item Difficulty and Item Discrimination

Table 2 reveals item difficulty and item discrimination indexes for each item on Form 1 and Form 2. Item difficulty index of items ranges from -.22 to 1.14 in Form B1 with the median value of .43 while item difficulty index of items varies from -.86 and .77 in Form B2 with the median of .21. Item discrimination index of items ranges between 1.43 and 7.73 in Form B1 while these ranges are observed with 1.20 and 7.99 values in Form B2. The median of Form B1 and Form B2 are 2.21 and 2.26, respectively.

Table 2. *Item Difficulty and Discrimination Indexes across Forms*

Item ID	Form B1 (The Least Selected Method)		Item ID	Form B2 (Random Method)	
	Item dif. (b)	*Item disc. (a)*		*Item dif. (b)*	*Item disc. (a)*
1	0.63	7.73	11	-0.21	3.21
2	-0.22	3.06	12	0.52	1.92
3	0.82	2.98	13	-0.79	1.54
4	0.83	2.14	14	-0.86	1.20
5	0.19	2.22	15	-0.10	1.71
6	0.22	1.43	16	0.63	2.59
7	-0.21	3.27	17	0.66	7.99
8	0.05	1.84	18	0.76	1.37
9	1.14	1.92	19	0.77	3.23
10	0.78	2.20	20	-0.26	4.19

To show the equality of item difficulty and item discrimination of Form B1 and Form B2, a nonparametric test; Mann-Whithey U Test, was run and it was found that average rank of item difficulty and item discrimination did not statistically vary across Forms B1 and B2 (item difficulty: *z* = -1.36, *p* = .17; item discrimination: *z* = -.34, *p* = .73).

3.2. Test Reliability

The reliability coefficient for each form was calculated through coefficient alpha estimation method in CTT. All forms had good internal consistency values ($\alpha_{\text{whole test}}$ = .84; $\alpha_{\text{form B1}}$ = .72; $\alpha_{\text{form B2}}$ = .73), which were greater than .70 (Thorndike, 2005). The 95% confidence interval for each alpha yielded great similarity in Form B1 and Form B2, which do not statistically differ from each other since their intervals are overlapped (Form $\text{B1}_{(.70,\ .74)}$; Form $\text{B2}_{(.71,.75)}$). When test marginal reliability is calculated using IRT 3PL model, the findings showed .70 for Form B1 and Form B2 and .84 for the whole test, which indicated similar results to that of CTT's. It allows us to infer that applying different elimination methods to reduce four-options to three-options does not statistically influence the reliability of the test.

In summary, the findings showed that different elimination methods; the least selected method and the random method, did not affect item difficulty, item discrimination, and test reliability negatively.

4. CONCLUSION AND DISCUSSION

This study aimed to examine how psychometric properties of items and test vary when different elimination methods were used to reduce the number of options of MCIs by applying the least selected method and random method. Research results showed that item difficulty, item difficulty, and test reliability did not statistically differ across the elimination methods administered to the items.

The results of this study could contribute to the growing body of research focusing on the impact of the number of options on psychometric characteristics. Overall, earlier research on item and test characteristics generally put great emphasis on comparing three-option MCIs with four-option MCIs rather than option elimination in these items. Namely, most of the empirical studies usually employed a particular traditional elimination method to reduce number of options of MCIs (i.e. least frequently chosen option, least discriminating option, and a random option). The findings of the present study are consistent with those in the some of the previous studies which showed that three-option MCIs perform equally well as four-option MCIs in terms of item discrimination and test reliability regardless of elimination methods (Atalmis & Kingston, 2017; Baghei & Amrahi, 2011; Delgado & Prieto, 1998; Rogers & Harley, 1994; Sidick, Barett, & Doverspike, 1994; Tarrant & Ware, 2010). Prior research reported contradictory results for item difficulty; for instance, three-option MCIs have been determined to be more difficult than four-option MCIs in Crehan et al. (1993) whereas Rodriguez (2005) also found a small change in item difficulty (.04) between four-option MCIs and three-option MCIs. However, this change was found to be statistically significant, meaning that four-option makes items more difficult. Atalmis & Kingston (2017), Baghei & Amrahi (2011), Delgado & Prieto (1998), Shizuka et al. (2006) and Tarrant & Ware (2010) found both types of MCIs were found to be equally difficult. On the other hand, none of the previous studies examined the impact of elimination method on psychometric characteristics of the mathematics items and/or tests. Keeping this in mind, the current study addressed how two elimination methods differentially affect psychometric characteristics in concern.

The results of this study could not only provide empirical support for test development studies but also make several recommendations to the test designers and classroom teachers. The use of different elimination methods does not significantly influence item difficulty, item discrimination, and reliability. In other words, psychometric characteristics did not vary when the least selected option and random option were deleted. Therefore, reducing the number of options of four-option MCIs to three options increases the efficiency of item-writing and administering test regardless of elimination methods. For instance, more three-option MCIs could be constructed in relatively shorter period of time as opposed to four-option MCIs (Aamodt & McSahne, 1992) since construction of a rationale distractor is widely considered as time consuming and one of the most challenging part of item writing (Haladyna et al., 2002). Besides, administering a test including three-option MCIs is expected to increase test reliability in certain ways as opposed to that composed of four-option MCIs. First, administering a test composed of three-option MCIs takes less time than that including four-option MCIs, and shorter tests are likely to decrease students' fatigue and test anxiety, which increases the reliability of the test. Second, the same amount of time is allotted to the implementation of the test including three-option MCIs and to that including four-options MCIs, the former is expected to contain a relatively higher number of items than the latter, which increases the test reliability.

Despite reporting findings on the use of elimination methods which are administered to MCIs, this study has certain limitations. Data in this study were obtained from the seventh and eighth grade students. So, different findings could be driven when the test is administered to the students attending lower or higher grades. For instance, relatively different findings are expected to be obtained when it is applied to those attending higher grades since they are considered to be more test-wise. It is also confined to the content of this test; namely, the items were constructed only in the scope of mathematics. So, tests prepared in different disciplines are expected to yield different results in terms of psychometric characteristics of MCIs and the test. The other limitation of this study is that convenience sampling method was applied for piloting and the ultimate phase from only two cities in Turkey. Although applying this sampling method is plausible as it is fast, inexpensive and easy, it might be implausible to generalize the findings for entire population. In addition, the items used in this study were at the middle difficulty level due to the fact that their item difficulty indexes ranged generally from -1 to +1. This could limit us to discriminate students in the upper and lower levels. The final limitation is that there are only 10 items used per test form. Although the reliability coefficient for each test composed of 10 items was found good internal consistency values in existing study, a small sample for statistical tests limits the generalizability.

Acknowledgments:

An earlier version of this paper was presented at the National Council on Measurement in Education in Chicago, US in 2015. The researcher would like to thank Dr. Marianne Perie, the director at Center for Educational Testing and Evaluation at Kansas, and Dr. Neal Kingston, the director of the Achievement and Assessment Institute at University of Kansas in the US, for their invaluable encouragement that made it possible to conduct this study.

ORCID

Erkan Hasan Atalmış https://orcid.org/0000-0001-9610-491X

5. REFERENCES

Aamodt, M. G., & McShane, T. D. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*(2), 151–160.

Abad, F., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema, 13*(1), 152-158.

AERA, APA, & NCME (2014). *Standards for educational and psychological tests*. Washington DC: American Psychological Association, American Educational Research Association, National Council on Measurement in Education.

Angelo, T. A., & Cross, K. P. (1993). Classroom assessment techniques: A handbook for college teachers. *San Francisco: Jossey-Bass*.

Atalmis, E. H., & Kingston, N. M. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*(4), 143-157.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*(2), 192-211.

Balta, N., &Eryılmaz, A. (2017). Counterintuitive dynamics test. *International Journal of Science and Mathematics Education*, *15*(3), 411-431.

Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational leadership*, *60*(1), 40-44.

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, *54*(4), 861-872.

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions 63 for continuing medical education activities and self-assessment modules. *RadioGraphics*, *26*(2), 543-551.

Crehan, K.D., Haladyna, T.M., & Brewer B.W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*(1), 241-247.

Darling-Hammond, L., & Youngs, P. (2002). Defining “highly qualified teachers”: What does “scientifically-based research” actually tell us?. *Educational researcher*, *31*(9), 13-25.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, *14*(3), 197-201.

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three-and four-option multiple choice questions. *Procedia-Social and Behavioral Sciences*, *98*, 398-403.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143.

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792 - 808.

Field, A. (2009). *Discovering statistics using SPSS*. London, England: Sage.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357-364.

Frey, B. B., & Schmitt, V. L. (2010). Teachers’ classroom assessment practices. *Middle Grades Research Journal*, *5*(3), 107-117.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hambleton, R.K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.

Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53*(3), 771–778.

Leahy, S., Lyon, C, Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Messick, S. (1989).Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, *2*(2), 65-72.

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, *27*(4), 388-394.

Rich, C. E., & Johanson, G. A. (1990, April). *An item-level analysis of "none of the above."* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Rodriguez, M. C. (1997). The art & science of item writing: A meta-analysis of multiple choice item format effects. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *Sage Open*, *4*(4), 1-10.

Rogers, W.T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: susceptibility to test wiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*(2), 234-247.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*(1), 35-57.

Sidick, J.T., Barrett, G.V., & Doverspike, D. (1994). Three-alternative multiple choice14 tests: An attractive option. *Personnel Psychology, 47*(4), 829-835.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, *10*(1), 7–12.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*(6), 354-363.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. *Nurse education today*, *30*(6), 539-543.

Thordike, R.M. (2005). *Measurement and Evaluation in Psychology and* Education (7th Ed.). Upper Saddle River, NJ: Pearson Education.

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, *51*(4), 829-837.

Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*(1), 271–280.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 325-338

DOI: 10.21449/ijate.379508

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

Measurement Invariance of Science Self-Efficacy Scale in PISA

Nermin Kıbrıslıoğlu Uysal [1,*], **Çiğdem Akın Arıkan** [1]

[1]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation, Beytepe, Ankara - Turkey

Abstract: The aim of this study is to find out whether the science self-efficacy scale in PISA 2006 and PISA 2015 applications ensure measurement invariance. Sample of the study consists of 4791 students in PISA 2006 and 5071 students in PISA 2015 implementation. Multi-group Confirmatory Factor Analysis (MGCFI) was performed to determine invariance of the science self-efficacy scale across year and gender. Invariance stages were examined by means of the comparison of fit indexes. The results of the study indicated that the science self-efficacy scale satisfied all stages of invariance by gender both in 2006 and 2015. Structural and metric invariance was provided for both gender across years and total group across years.

ARTICLE HISTORY

Received: 16 January 2018
Revised: 11 April 2018
Accepted: 14 April 2018

KEYWORDS

PISA, Self-efficacy, Gender, Measurement invariance

1. INTRODUCTION

In recent years, interaction and competition among countries has been growing with development in technology and ease of communication. This situation is accompanied by comparisons made among countries in many areas (OECDa). Therefore, international organizations carry out studies in different areas. These works not only allow countries to assess themselves in the international platform but also constitute feedback regarding their policies and education levels.

A particular work done by the OECD in the field of education is the PISA (Program for International Student Assessment). PISA has been applied every three years since 2000, and its main goal is to measure whether 15-year old students, who are expected to complete mandatory education use the information obtained from their education life in real life situations independent from countries' educational curricula (OECDb). Turkey has participated in PISA since 2003. In Turkey, while PISA was given as a pencil and paper exam until 2015, it was turned into computer-based application in 2015 (OECD, 2015).

PISA evaluations include reading, mathematics and science test in the cognitive domain and each year one of these areas is taken as the main focus. Besides these tests, within the scope of PISA applications, surveys are given to students, parents and school administrators

CONTACT: Nermin Kıbrıslıoğlu Uysal ✉ nkibrislioglu@hacettepe.edu.tr Hacettepe Üniversitesi, Beytepe Yerleşkesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı 06800, Beytepe, Ankara

ISSN-e: 2148-7456 /© IJATE 2018

in order to learn more broadly about students' backgrounds and learning experiences and school system and learning environments. Such surveys also focus on the core area covered in each cycle (OECDb). Thus, the same basic area is measured every 9 years. Apart from that, in order to reveal the trend by years and associate the findings with previous tests in PISA applications; common items are used in the cognitive domain and common scales in the affective domain. Thus, students are compared by years in both cognitive and affective domains (OECDb).

PISA 2006 and 2015 implementations focus on science. Though revised partially, the students' surveys applied in 2006 and 2015 contain similar scales. Therefore, it is possible to make a comparison on affective characteristics of students across years. In addition, the literature provides abundant samples of comparison of affective characteristics by gender. However, in order to derive the correct results from the comparisons made, the examinees who are equivalent in terms of the trait measured must get the same score from the tests or scales (Schimit and Kuljanin, 2008). To put it differently, the same trait must be associated identically with the group of variables observed the same in all groups (Borsboom, 2006). In other words, the scores obtained from the scales can be used for comparison across different groups provided that the scales ensure the measurement invariance between the groups concerned. In fact, measurement invariance is an assumption that must be checked before comparisons are made between groups because the traits measured where measurement invariance is not met, may not be identical across the groups measured (Vandenberg & Lance, 2000). Nonetheless, measurement invariance is rarely tested in studies. This makes the validity of the results obtained questionable because comparisons are made without having information about construct validity of scales and equality of the validity across groups (Gregorich, 2006).

1.1. Measurement Invariance

In general terms, measurement invariance refers to examining whether the scores measuring a particular construct have the same meaning under different circumstances. Different conditions could include different populations, different measurement times or different methods of application (such as paper-pencil and computer-based). Consistency of the construct across years is referred to as longitudinal measurement invariance and deals with whether factor structure varies in longitudinal pattern by years. Invariance between populations is related to structural bias and investigates whether the measured trait is the same or not across groups (Kline, 2011).

A widely used method for measuring invariance is the multi-group confirmatory factor analysis (MGCFI) method (Widaman and Rice, 1997; Vandenberg and Lance, 2000; Kline, 2011). In the MGCFI method, different stages of measurement invariance can be tested for different purposes. Those stages can be listed as invariance of covariance matrice, configural invariance, metric invariance, scalar invariance, strict factorial invariance, invariance of factor covariance and invariance of factor averages (Vandenberg & Lance, 2000). As a matter of fact, the comparability of observed scores between groups can be provided with configural, metric, scalar and strict factorial invariance (Widaman & Rice, 1997). In this study, measurement invariance is investigated in relation with these four type of invariance. The stages of measurement invariance are summarized in Table 1**.**

As seen in Table 1, the first stage of measurement invariance is configural invariance. Configural invariance only requires the identical measurement pattern across groups. If configural invariance is not provided, measurement invariance will not be ensured at any stage (Kline, 2011). Secondly, metric invariance requires identical factor loadings across groups as well as configural invariance. Metric invariance is also called weak factorial invariance. Once metric invariance is ensured, it can be argued that the covariance differences in the variable measured across groups arise from the common factors; leaving the root of observed score differences between groups unexplored (Millsap & Olivera-Aguilar, 2012). On the other hand, when metric invariance is not ensured, it could be argued that the factors do not have the same meaning across groups (Gregorich, 2006). The following stage, scalar invariance is a strong level of invariance and requires equality of factor variance and covariances between groups as well as metric invariance. When scalar invariance is ensured, comparison of differences between averages of the groups yields significant outcomes (Millsap and Olivera-Aguilar, 2012). Finally, strict factorial invariance requires equality of item residual variances between groups in addition to scalar invariance. Provision of strict factorial invariance leads the way for comparing not only observed variable averages but also factor variance and covariances between groups (Gregorich, 2006). However, as variance of the latent variable increases, item residual variance also increases, strict factorial invariance is often not achieved in practice. The stages of measurement invariance are hierarchical. Therefore, the stages are evaluated respectively, and if invariance is not provided at any stage, there is no need to examine the following stage.

Table 1. Measurement Invariance Stages

Degree of Invariance	Condition of Invariance	Group Comparison
Configural invariance	Item/Factor groups	---
Metric invariance	Item/Factor groups and factor loads	Factor variance and covariances
Scalar Invariance	Item/Factor groups, factor loads and item constants	Factor variance and covariances, factor and observed variable averages
Strict factorial invariance	Item/Factor groups, factor loads, item constants, and item residual variances	Factor variance and covariances, factor and observed variable averages, observed variance and covariances

1.2. Self-efficacy

According to Bandura (1982), self-efficacy is the self-judgment about how well an individual can do a behavior. Self-efficacy perception affects behaviour and performance, as well as beliefs of individuals. So even if individuals have an idea about the result of a behavior, they tend to avoid conducting that behaviour as long as they have a low level of self-efficacy related to that particular behaviour (Bandura, 1977).

Self-efficacy is not a personal trait by nature; rather, it focuses on performance capabilities targeting specific objects (Zimmerman, 2000). In this case, science self-efficacy can be defined as the extent at which students believe in their own abilities to succeed in science-related tasks. Self-efficacy has an impact on future-oriented behaviours of individuals. In other words, before individuals perform any behavior, they evaluate their self-efficacy towards that behavior. In this regard, students' self-efficacy perception in a particular subject area affects their desire to underatake activities related to that field, the efforts they would show for these activities and continuity of the efforts, and thus their performance in that area (Zimmerman, 2000).

Students' science self-efficacy affects their desire to undertake science related tasks, science achivement as well as their future preferences (Lent, Brown, & Larkin, 1986; Post, Stewart & Smith, 1991; Andrew, 1998; Scott & Mallinckrodt, 2005; Zedlin, Britner & Pajares, 2007). For example, Scott ve Mallinckrodt (2005) examined female high school students' science self-efficacy and their career preferences related to science. They reported that between strudents who prefereed science related major and the ones do not, differ significantly with respect to their science self-efficacy. Moreover, there are plenty of studies related to comparison of male and female students’ self-efficacy in the literature (Post, Stewart & Smith, 1991; Britner & Pajares, 2001; Zedlin, Britner & Pajares, 2007:) Britner and Pajares (2001) investigated possible gender differences on high school students’ science self-efficacy and motivation. They reported that girls have stronger science self-efficacy beliefs and higher grades while boys have stonger performance-approach goals. Zedlin, Britner and Pajares (2007) examined the self-efficacy beliefs of men and women who selected career in science and mathematics majors. The results of the study indicated that women and men have different sourses of self-efficacy beliefs.

Hence, science self-efficacy is an important construct and it is studied in the literature frequently. Moreover, comparisons of science self-efficacy between gender groups is also very common. In order to provide correct implications from these comprasions it is very important to test the invariance of these scales (Vandenberg & Lance, 2000).

In order to determine scientific literacy in PISA applications, not only achievement tests but also surveys on students' affective traits related to academic achievement are applied. In both 2006 and 2015 tests, students' science self-efficacy was measured in the scope of the affective domain surveys. The scale items were same in 2006 and 2015. In relation with science self-efficacy, a number of tasks were listed in the students' survey and students were asked how easy they find it to do these tasks on their own (MEB, 2010). The scale items are shown in Table 2.

Table 2. PISA 2006-2015 science field self-efficacy scale items

1. Recognizing the question underlying the newspaper article on a health problem
2. Explaining why earthquakes take place more often in some areas
3. Explaining the role of antibiotics in treatment of diseases
4. Identifying the problem regarding proper collection and treatment of wastes
5. Predicting how the changes in the environment could affect survival of certain living species
6. Interpreting the scientific information on labels on foodstuffs
7. Discussing how new evidence can change the understanding that there is life on the Mars
8. Deciding which of the two views about acid rains is better

*Taken from 2006 and 2015 PISA National Reports.

1.3. Aim of the Study

The aim of this study is to find out whether or not the science self-efficacy scale, which was given as a common scale in 2006 and 2015, satisfies measurement invariance across years and gender groups in Turkey sample.

In the literature, there are many studies related to measurement invariance for different groups on scales used in international tests such as PISA and TIMSS (Ercikan and Koh, 2005; Marsh et al., 2006; Wu, Lin & Zumbo, 2007; Lee, 2009; Akyıldız, 2009; Uzun & Öğretmen, 2010; Güzeller, 2011; Asil & Gelbal, 2012; Uyar & Doğan, 2014; Başusta and Gelbal, 2015; Bulut, Palma, Rodrigez and Stanke, 2015; Kıbrıslıoğlu, 2015; Karakoç and Alatlı, 2016; Ölçüoğlu & Çetin, 2016; Gülleroğlu, 2017). Ercikan and Koh (2005) examined invariance of English and French forms in TIMSS 1999 implementation. They analysed the invariance with both MGCFA and differential item functioning (DIF) analysis. They reported that both mathematics and science forms did not ensure measurement invariance. Lee (2009) examined whether math self-concept, math self-efficacy, and math anxiety scales in PISA 2003 implementation provide one consistent factor structure between 41 countires. For this purpose, he conducted exploratory, confirmatory and multi group confirmatory factor analysis. The results of the study indicated that structure of these constructs differ between countries. Bulut, Palma, Rodrigez and Stanke (2015) investigated measurement invariance of support and positive identity scales among White and Latin American students across years. The results of the study indicated that subgroup-year interaction has a significant effect on parameter shift. Hence, the invariance of scale parameters between two different groups of students differs across years did not provided.

Uyar and Doğan (2014) investigated measurement invariance of the model for learning strategies in the PISA 2009 students' survey across gender, school type and statistical region group. It was found out that only configural invariance and metric invariance were met in gender and school type groups, while all of the invariance conditions were fulfilled in relation with regions. In another study, Uzun and Öğretmen (2010) investigated whether variables affecting students' success in science such as self-efficacy, attitude, significance and in-class student activities satisfy measurement invariance in the Turkish participants in 1999 TIMMS-R. They found out that self-efficacy, significance and in-class student activities satisfy metric invariance; while attitude satisfies scalar invariance between gender groups. Kıbrıslıoğlu (2015) investigated invariance of the items in mathematics subscale of PISA 2012 was investigated across gender and cultures. It was found out that intercultural invariance meets invariance only in configural level while gender groups meets strict factorial invariance. Furthermore, Başusta and Gelbal (2015) examined measurement invariance of the science and technology items in the PISA 2009 students' survey against gender. They reported that scalar invariance ensured between gender groups. Also, Gülleroğlu (2017) investigated measurement invariance of affective traits such as interest, anxiety, self-efficacy and sense of self regarding mathematics against gender in the PISA 2012 implementation. It was noted that mathematics self-efficacy scale does not satisfy configural invariance. On the other hand, sense of self-regarding mathematics scale ensured configural invariance and anxiety and interest towards mathematics satisfy scalar invariance.

Different from the literature, present study investigates not only measurement invariance across years and genders but also whether or not the scale items satisfy invariance for each gender between the years 2006 and 2015. Due to the non-longitudinal nature of PISA data, the analyses targeted measurement invariance in different groups across years. In the PISA applications, population defined as the representative population and the sample is selected at random. Present study was carried out assuming that both applications consisted of samples with similar individuals. Therefore, the study is expected to demonstrate whether or not the construct varied in gender subgroups between 2015 and 2006. Moreover, investigation of bias in subgroups allow for unearthing the probable bias that can not be revealed as a result of bias analysis for the whole group but affects subgroups (Huggings-Manley, 2016). So, the study is considered significant as it attempts to additionally reveal the change between genders over the years.

Hence, this study was intended to find out whether science self-efficacy scale in PISA 2006 and 2015, which were given as common tests, satisfy measurement invariance depending on year and gender, and also invariance against year in gender subgroups. Answer was sought for the following questions in the study:

(1) Does science self-efficacy scale satisfy measurement invariance between the years 2006 and 2015?

(2) Does science self-efficacy scale satisfy measurement invariance between gender subgroups?

2. METHOD

2.1. Research Method

This study is a descriptive study as it aims identifying whether the science self-efficacy scale in the students' survey in the PISA 2006 and PISA 2015 implementations is invariant by years and gender.

2.2. Population and Sample

A total of 4942 students from 160 schools participated in the PISA 2006 application held in Turkey. The participants were selected from 7 geographic regions, 51 provinces in a random manner by two-step stratification of regions and schools. As for PISA 2015, participants were selected by means of two-stage random sampling method. At step one; schools and students were identified by means of stratified random sampling with respect to the strata of Statistical Region Units Classification (SRUC) Level 1, education type, school type, the place of schools and administrative form of schools. PISA 2015 was administered to 5895 students from 187 schools in 61 provinces to represent 12 different regions according to the SRUC Level 1 (MEB, 2017).

2.3. Data Analysis

Measurement invariance between the groups was examined by means of MGCFI method. Before analysis, the assumptions of missing data, extreme values, multivariate normality, and multicollinearity were tested (Çokluk, Şekercioğlu & Büyüköztürk, 2012). The assumptions are elaborated below.

For missing data, first of all, the examinees who responded to none of items in the PISA 2006 and PISA 2015 applications were removed from the data. Then, missing data analysis was performed for both data sets and missing data rates were examined. The analysis yielded missing data rate below 5% distributed randomly. Kline (2011) stated that in the case of large samples, missing data rate below 5% with random distribution, such data could be omitted. For this reason, the missing data were removed from both data sets under listwise deletion condition. Analyses were performed for 4814 and 5235 respondents for the PISA 2006 and PISA 2015 implementations, respectively.

Secondly, univariate and multivariate outliers were examined. One way to determine univariate extreme values is to convert variables into standard variables. In large samples (n> 100), z-scores outside the range of -3 to +3 are regarded extreme values (Tabachnick & Fidell, 2007). On the other hand, multivariate extreme values can be computed from the Mahalanobis distance. The Mahalanobis distance exhibits the chi-square distribution and degree of freedom is equal to the number of variables in the data set. The values smaller than the chi square value at 0.001 significance level were identified as outliers (Tabachnick and Fidell, 2007). For univariate extreme values, z values were examined indicating no case outside the specified range. Mahalanobis distances were then investigated for multivariate extreme values. As a result, 23 respondents were removed from analysis as extreme values for PISA 2006 and 164 respondents for PISA 2016, respectively. Frequency table for the remaining respondents are given in Table 3.

Table 3. Gender-Related Frequency

Gender	2006		2015	
	f	%	*f*	%
Female	2229	46.5	2559	50.5
Male	2562	51.5	2512	49.5
Total	4791	100	5071	100

It can be seen in Table 3 that 4791 participants in year 2006 included 2229 females and 2562 males; whereas 5071 participants in 2015 consisted of 2559 females and 2512 males. The distribution of gender groups by years is balanced.

Multivariate normality is achieved provided that univariate normality is provided and linearity and residuals are covariant (Kline, 2011). For univariate normality assumption, skewness and kurtosis coefficients of the variables were examined. It was found out that the skewness and kurtosis coefficients for PISA 2015 data are in the range of -1 and +1. As for the PISA 2006 data, only the coefficient of item 7 was found to be -1.064 the other variables falling in the specified range. The skewness and kurtosis coefficients in the specified range suggest that the variables satisfy normality assumption (Büyüköztürk, 2002). As for linearity, residual graphs were examined indicating that linearity assumption is met. For homoscedasticity, Durbin Watson values were examined with resulting values in the range of 0 - 4 fulfilling the homoscedasticity assumption (Tabachnick & Fidell, 2007). In addition, multivariate normality was checked with Bartlett sphericity Test yielding significant results for all subgroups. This shows suitability of the data for multivariate normal distribution (Çokluk et al., 2012).

Multicollinearity assumption was tested with tolerance value, conditional index and variance inflation factor values (VIF). Multicollineraity does not exist when tolerance values are greater than 0.10 for the absence of transactions, VIF values are smaller than 10, and conditional index values are (CI) smaller than 30 (Tabachnick & Fidell, 2007). In this study, tolerans values are between 0.4-0.8, VIF values are between 1.5-2.4 and CI values are between 1-12.5. Analyses showed no problem of multicollinearity in data. The investigations revealed that the data meet the required assumptions. Prior to MGCFI, a confirmatory factor analysis (CFA) was performed and goodness of fit statistics were examined to test the fit of the model. The CFA model for PISA 2006 is shown in Figure 1.

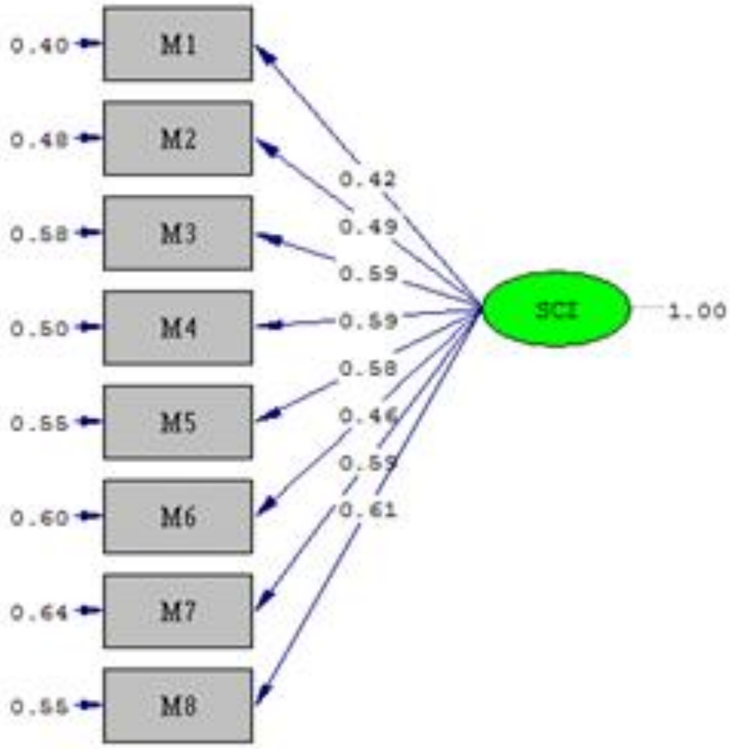


Figure 1. Configural model of the science self-efficacy scale

The values in the model above shows that factor loadings fall in the range of .40 and .64. It corresponds to compliance indices at acceptable levels except from χ2/df ratio (χ2= 427.35, df= 20, CFI=0.98, TLI= 0.97 RMSEA= 0.06, GFI=0.97). The chi-square statistic is affected by sample size and usually significant in large samples (Kline, 2011) and therefore, χ2 value does not taken as a basis for rejection or acceptance of the models (Schermelleh-Engel, Moosbrugger & Müller, 2003). So, the model was evaluated according to other indices.

MGCFA analysis was conducted in four stage. At first; configural invariance, which has free factor loads, factor correlations and error variances was tested. At the following stage, metric invariance was tested, which has free factor correlations and error variances under the condition of equal factor loads. Then, scalar invariance was tested with equal factor loads and factor correlations but free error variances. Lastly, strict factorial invariance was tested which has equal factor loads, factor correlations and error variances. At each stage, the difference values of the comparative fit index, (ΔCFI) were examined to decide whether invariance is satisfied or not. ΔCFI values smaller than or equal to -0.01; indicates invariance is achieved; otherwise it is not satisfied (Cheung & Rensvold, 2002).

3. RESULTS

This study was carried out to investigate measurement invariance of science self-efficacy scale over years as well as gender groups. The findings are reported in a way to discuss the research problems one by one.

3.1. MGCFI Results by Years

The goodness of fit indexes obtained at each invariance stage of the MGCFI are displayed in Table 4, which implies whether the science self-efficacy scale consists of eight items are equivalent between the years 2006 and 2015.

Table 4. Goodness of fit indexes by levels of invariance for 2006 and 2015

	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	1595.4	40	0.092	0.039	0.96	0.97	
Metric	1693.2	47	0.087	0.043	0.97	0.97	0
Scalar	2734.0	63	0.094	0.096	0.96	0.96	-0.01
Strict	2596.6	71	0.087	0.055	0.98	0.97	0.01

The goodness of fit indexes in Table 4 show that the RMSEA value was outside the acceptable interval at the configural invariance stage, while the other statistics of concordance fall in the acceptable range (RMSEA>.08; SRMR<.1; TLI>.95; CFI>.95). This means that the structure of the model has remained the same over the years. After providing configural invariance, metric invariance was tested.

For metric invariance, the goodness of fit indexes in Table 4 show that the RMSEA value is outside the acceptable range, but the other statistics indicate model fit. Metric invariance is ensured as ΔCFI value is in acceptable range (ΔCFI ≤ 0.01). This finding implies that the relations between the measured traits and self-efficacy dimension have remained similar across years. As metric invariance ensured, scalar invariance is tested.

The values in Table 4 show that RMSEA values are outside the acceptable range, while the rest of the statistics fall within the acceptable range. ΔCFI value indicate scalar invariance was met, (ΔCFI ≤ 0.01). Hence we concluded that sclalar invariance was ensured. This finding suggests that item factor loads and factor correlations are similar in both years. After ensuring scalar invariance, the last stage, strict factorial invariance, was implemented.

While checking the indices in Table 4 for strict invariance, the differences between CFI were obtained for scalar invariance and strict factorial invariance, respectively. The ΔCFI values reveal that strict factorial invariance is satisfied in this case.

The analysis of the invariance between 2006 and 2015 in whole group indicated that invariance is ensured in scalar level. Hence, item factor loads and factor correlations are similar in both years while item residual variances are different.

3.2. MGCFI Results by Gender

Measurement invariance between genders was checked separately for years 2006 and 2015 in whole group. The resulting goodness of fit indexes are given in Table 5.

Table 5. Goodness of fit indexes by gender in 2006 and 2015

2006	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	435,23	40	0.064	0,031	0,97	0,98	
Metric	457,77	47	0.061	0,035	0,97	0,98	0
Scalar	651,07	63	0,063	0,051	0,97	0,97	-0,01
Strict	662,48	71	0,059	0,051	0,97	0,96	-0,01
2015	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	1249,37	40	0,11	0,043	0,96	0,97	
Metric	1267,83	47	0,11	0,047	0,97	0,97	0
Scalar	1482,44	63	0,098	0,057	0,97	0,97	0
Strict	1544,86	71	0,093	0,058	0,97	0,97	0

For configural invariance; the goodness of fit indexes in Table 5 demonstrate that all indices fall within the acceptable range in 2006; whereas the RMSEA values are outside such range in 2015. This result suggests that the structure of the model remained unchanged for genders across years. Once configural invariance was ensured as prerequisite of metric invariance, the latter was checked.

In relation with metric invariance, Table 5 indicates acceptable limits for all statistics for year 2006 and 2015; while the statistics except for RMSEA fall within acceptable limits for 2015 (RMSEA<.08; SRMR<.1; TLI>.95; CFI>.95). Examination of ΔCFI refers to positive metric invariance ($\Delta CFI \leq 0.01$). This finding implies that the relationships between the measured traits and science self-efficacy dimension are similar in both genders. After metric invariance was ensured, the next phase was implemented.

Examination of the values in Table 5 for scalar invariance reveals that the RMSEA, value is outside the acceptable limits for 2015 but the other values are acceptable. When ΔCFI values are examined, it is seen that scalar invariance is provided ($\Delta CFI \leq 0.01$). The finding reflects invaried item factor loads and item constants across genders. When scalar invariance was deemed acceptable, the last stage was implemented.

With respect to strict factorial invariance, goodness of fit indexes in Table 4 refer to acceptable levels for year 2006. However, in 2015, the statistics except for RMSEA are seen at acceptable limits. The ΔCFI value is found to be within the acceptable range for both 2006 and 2015 ($\Delta CFI \leq 0.01$). The finding suggests that error variances did not vary between genders in 2006 and 2015.

The analysis of the invariance between gender groups indicated that invariance is ensured in strict invariance level in both 2006 and 2015 implementation with respect to ΔCFI values. Hence, item factor loads, factor correlations and error variances are similar between gender groups in both years. However, the model fit of 2015 seem problematic as RMSEA values are really high in al stages. This may implies that the model in 2015 may be revised.

3.3. MGCFI results in gender subgroups across years

As for the third sub-problem of the research, measurement invariance analyses across years were performed separately in gender subgroups. Indeed model invariance was tested between female students in 2006 and female students in 2015; male students in 20016 and male students in 2015 respectively. The resulting coefficients are given in Table 6.

Table 6. Goodness of fit indexes by gender subgroups between years 2006 and 2015

Female	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	691,09	40	0,084	0,035	0,97	0,98	
Metric	728,01	47	0,079	0,04	0,97	0,98	0
Scalar	1323,66	63	0,092	0,1	0,96	0,95	-0,03
Strict	2333,52	71	0,11	0,084	0,94	0,92	-0,03
Male	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	993,51	40	0,1	0,043	0,96	0,97	
Metric	1063,86	47	0,095	0,049	0,96	0,97	0
Scalar	1571,52	63	0,099	0,094	0,96	0,95	-0,02
Strict	2903,2	71	0,12	0,085	0,93	0,91	-0,04

To start with, configural invariance was tested through invariance of the model, factor and items in gender subgroups across years. Examination of the indices in Table 6 show that both subgroups have acceptable values except for RMSEA. This result suggests that the model structure remained unchanged in both male and female subgroups across years.

When the indices in Table 6 are examined in relation with metric invariance, female participants meet the acceptable limits for all statistics, while males are seen to be within the acceptable interval for the values except for RMSEA. Examination ΔCFI reveals that metric invariance is satisfied between 2006 and 2015 both in females and males ($\Delta CFI \leq 0.01$). This finding refers to similar relationship between the measured traits in 2006 and 2015 science self-efficacy dimension in female group as well as the male group. Once metric invariance ensured, the next stage of scalar invariance is checked.

Considering scalar invariance, the fit indices in Table 6 reveal that RMSEA values are outside the acceptable interval both for males and females. ΔCFI values show that scalar invariance is not met between 2006 and 2015 both for females and males ($\Delta CFI > 0.01$). The finding implies that item factor loads were unchaged in gender groups across years, but item constants varied. Due to the lack of scalar invariance, it is questionable to compare gender averages across years. Likewise, strict factorial invariance was not tested.

The analysis of the invariance between implementations within each gender groups indicated that invariance is ensured only in metric level in both male and female group. Hence, item factor loads, factor correlations and error variances are different between 2006 and 20165 implementations in female group as well as male group.

4. DISCUSSION

In this study, it was intended to find out whether science self-efficacy scale of Turkish students taking PISA 2006 and PISA 2015 exams changes across years and gender.

In the model for the self-efficacy scale towards science; it was observed that configural metric and scalar invariance are satisfied for the total group across years; while strict factorial invariance is not satisfied. It was found out that the variables in the model manifest similar factor loads and factor correlations but different error variances across years. On the other hand, separate investigation of invariance by gender subgroups in years 2006 and 2015 revealed ensuring of configural invariance, metric invariance, scalar invariance, and strict factorial invariance. The finding of this study does not seem to be in parallel with findings by Uyar and Doğan (2014) and Gülleroğlu (2017). In neither study above, strict factorial invariance was not provided according to gender group, but it was satisfied in present study. In addition, Uzun and Öğretmen (2010) found out that science self-eficacy variable meets metric invariance according to gender. The finding of Uzun and Öğretmen seems to be at odss with our findings. On the other hand, the results are parallel with findings obtained by Başusta and Gelbal (2015) from a study conducted on PISA 2009 focusing on invariance of the items in science and technology scales across gender. It was found out that the variables in the model revealed similar factor loads, factor correlations and error variances between different gender sub-groups for both years. This suggests that averages of the variables in the science self-efficacy scale can be comparible across gender subgroups. In studies by Saracaloğlu, Yenice and Özden (2013) and Balbağ and Balbağ (2016), it was found out that self-efficacy perception regarding Science and Technology Literacy does not show a significant variance across gender. In this regard, it can be argued that the measurements obtained from the model established with science self-efficacy in the PISA students' survey can be generalized for gender.

In gender subgroups, only configural and metric invariance were satisfied between 2006 and 2015 in both male and female groups, but scalar and strict factorial invariance could not be met. It was found out that the model varibles show similar factor loads between gender groups across years. In other words, the structure was found to be constant across genders. On the other hand, the respondents of 2006 and 2015 applications have divergent interpretations could have an influence on the lack of scalar and strict factorial invariance. During the 9-year period between the two applications, the self-efficacy of different genders towards science could have differed. Still, bearing in mind the probability that the modelled structure cannot remain unchanged, one reason for the lack of invariance could be because the existing structure might have changed. We think that it is needed to write items conforming to current conditions by revising the items in students' questionnaires covered under international tests like PISA guiding national educational policies in the light of application experience. For example, life on Mars did not raise as heated debate as now in 2006. However, today research at NASA is underway for trip to Mars. Considering these developments, we believe that it is no longer possible to interpret item 7 in the questionnaire, which reads as “Discuss how new evidence could change the debate whether life exists on the Mars” in the same way in practice in the course of time. In addition, while the questionnaire was given as a paper-pencil test in 2006, the application became computer-based in 2015. So, computer skills of students may affect their responses, this also could account for the lack of invariance.

Lastly, Bulut, Palma, Rodrigez and Stanke (2015) studied parameter invariance of support and positive identity scales among White and Latin American students across years. They found out that subgroup-year interaction has a significant effect on parameter shift. Hence, the variance of scale parameters between two different groups of students differs across years. Huggings-Manley (2016) stated that item parameters remain unchanged in applications performed at different times; yet, parameter difference could appear between subgroups in the course of time. Thus, further examination of the lower values of goodness of fit indexes in male group across years could shed light onto cases, which are not explained by studies on group invariance and bias. This can also be a sign that self-efficacy pattern varies differently among girls and boys across years. Moreover, recent social responsibility projects promoting sciences for female students may also have had an effect on their self-efficacy. Determining possible causes of these results remain out of the scope of our research. Still, it is advisable to carry out more in-depth research to this end in future studies.

ORCID

Nermin Kıbrıslıoğlu Uysal https://orcid.org/0000-0002-9592-469X
Çiğdem Akın Arıkan https://orcid.org/0000-0001-5255-8792

5. REFERENCES

Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6(1)*, 18- 47.

Andrew, S. (1998). Self-efficacy as a predictor of academic performance in science. *Journal of Advanced Nursing, 1998, 27*, 596–603, doi: 10.1046/j.1365-2648.1998.00550.x.

Asil, M. &Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Eğitim ve Bilim, 37(166)*, 236-249.

Balbağ, M. Z., & Balbağ, N. L. (2016). Öğretmen adaylarının fen ve teknoloji okuryazarlığına ilişkin özyeterlik algıları ile bilgi okuryazarlıkları arasındaki ilişkinin incelenmesi. *Pegem Atıf İndeksi*, 429-446.

Bandura, A., (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84 (2)*, 191-215.

Bandura, A., (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37 (2), 122-147.

Başusta, N. B., & Gelbal, S. (2015). Gruplararası Karşılaştırmalarda Ölçme Değişmezliğinin Test Edilmesi: PISA Öğrenci Anketi Örneği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30(4)*, 80-90.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care, 44*, 176-181.

Britner, S.L. & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering, 7 (4)*, doi: 10.1615/JWomenMinorScienEng.v7.i4.10.

Bulut, O., Palma, J., Rodrigez, M. C. & Stanke,L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *SAGE Open*, 1–18, doi: 10.1177/2158244015586238.

Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi elkitabı*. Ankara: Pegem Yayıncılık.

Cheung, G., W., & Rensvold, R., B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9(2)*, 233–255, doi: 10.1207/S15328007SEM0902_5

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları.* Pegem Akademi.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*, 23-35.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*, 78-94.

Gülleroğlu, H. D. (2017). PISA 2012 Matematik Uygulamasına Katılan Türk Öğrencilerin Duyuşsal Özeliklerinin Cinsiyete Göre Ölçme Değişmezliğinin İncelenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 37(1),* 151-175.

Güzeller, C. (2011). PISA 2009 Öğrenci Anketinde Yer Alan Bilgisayar Tutum Boyutunun Kültürlerarası Eşitliğinin İncelenmesi. *Eğitim ve Bilim, 36(162),*320-327.

Huggings-Manley, A.C. (2016). Psychometric Consequences of Subpopulation Item Parameter Drift. *Educational and Psychological Measurement, 77(1),* 143–164. doi: 10.1177/0013164416643369

Karakoc Alatli, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research, 66,* 389-406. http://dx.doi.org/10.14689/ejer.2016.66.22

Kıbrıslıoğlu, N. (2015). *PISA 2012 Matematik Öğrenme Modelinin Kültürlere ve Cinsiyete 476 Göre Ölçme Değişmezliğinin İncelenmesi: Türkiye -Çin (Şangay) -Endonezya Örneği.* Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi

Kline, R.B., (2011). *Principles and Practices of Structural Equation Modelling*. New York, The Guilford Press.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries *Learning and Individual Differences 19,* 355–365.

Lent, R. W., Brown, S. D., & Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology, 33(3),* 265-269.

OECDa (Organisation for Economic Co-operation and Development). http://www.oecd.org/

OECDb (Organisation for Economic Co-operation and Development). http://www.oecd.org/education/

OECD (Organisation for Economic Co-operation and Development) (2015). *PISA 2015 Technical Report.* http://www.oecd.org/pisa/data/2015-technical-report/

Marsh, H. W., Hau, K. T., Artelt, C., Boument, J., & Peschar, J. (2006). OECD's brief selfreport measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6 (4),* 311-360.

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380– 392). New York, NY: Guilford Press.

Ölçüoğlu, R., & Çetin, S. (2016). TIMSS 2011 Sekizinci Sınıf Öğrencilerinin Matematik Başarısını Etkileyen Değişkenlerin Bölgelere Göre İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7(1), 202-220.*

Post, P., Stewart, M. A., & Smith, P. L. (1991). Self-efficacy, interest, and consideration of math/science and non-math/science occupations among Black freshmen. *Journal of Vocational Behavior, 38(2),* 179-186. Doi: https://doi.org/10.1016/0001-8791(91)90025-H

Saracaloğlu, A. S., Yenice, N., & Özden, B. (2013). Fen bilgisi, sosyal bilgiler ve sınıf öğretmeni adaylarının öğretmen öz-yeterlik algılarının ve akademik kontrol odaklarının incelenmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 34(2),* 227-250.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: review of practice and implication. *Human resources management review, 18,* 210-222.

Scott, A. B., & Mallinckrodt, B. (2004). Parental Emotional Support, Science Self-Efficacy, and Choice of Science Major in Undergraduate Women. *The Career Development Quarterly, 53 (3),* 263-273, doi: 10.1002/j.2161-0045.2005.tb00995.x

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003), Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodnessof-Fit Measures, *Methods of Psychological Research Online, 8* (2), pp.23-74.

Tabachnick, B. G., & Fidell, L. S. (2007*). Using Multivariate Statistics* (5. Eds). Boston: Pearson Education.

Uyar. Ş. ve Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi*, *2,* 30-43.

Uzun, B., Öğretmen, T. (2010). Fen basarisi ile ilgili bazı değiskenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değismezliğinin değerlendirilmesi. *Eğitim ve Bilim, 35(155),* 26-35.

Vandenberg, R. J., & Lance, C. E., (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions Practices, and Recommendations for Organizational Research. *Organizational Research Methods 3 (4),* 4-70.

Widaman, K. F., & Reise, S. P., (1997). Exploring the measurement invariance of psychological instruments: Applications in substance use domain. The science of prevention: *Methodological advances from alcohol and substance abuse research,* 281-324.

Zedlin, A., L., Britner, S., L., & Pajares, F. (2007). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, 45(9),1036–1058.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25,* 82-91.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 339-369

DOI: 10.21449/ijate.416261

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

High School Students' Performances on Proof Comprehension Tests

Bahattin İnam[1], Işıkhan Uğurel [2], Burçak Boz Yaman[3*]

[1]Şehit Aydın Berber Mesleki ve Teknik Anadolu Lisesi, Çaycuma, Zonguldak, Turkey
[2]Mathematics and Science Education Department, Dokuz Eylül University, İzmir, Turkey
[3]Mathematics and Science Education Department, Muğla Sıtkı Koçman University, Muğla, Turkey

Abstract: This study is a part of a large scale project in which an action research design is used to teach proof to 11th grade students. This part of the project aims to identify students' comprehension level through five proof comprehension tests developed by the researchers based on the National Geometry Curriculum. Data were analyzed by considering the framework of Yang and Lin's (2008) multilevel model. Results showed none of the students were successful at the most sophisticated level of the proof comprehension tests which requires conducting a proof in various ways or proving different theorems by using the same proof methods. Moreover, the highest proof comprehension was obtained from the level containing knowledge about definition, properties, and meanings of symbols. Achievement and comprehension decreased for components of a proof needing higher level mathematical skills. Based on the study's results, suggestions about teaching proof are provided.

ARTICLE HISTORY

Received: 17 January 2018
Revised: 10 April 2018
Accepted: 13 April 2018

KEYWORDS

Proof, proving,
Comprehension of proof,
Proof comprehension test,
High school student,
Mathematics education

1. INTRODUCTION

A mathematical proof is used to verify a result; inform and convince others; discover a result; and arrange results into a deductive system (Almeida, 2003). It is a concept containing mental processes like identifying mathematical structures and invariants, exploring, proposing assumptions, and organizing logical arguments (Ball, Hoyles, Jahnke, & Movshovitz-Hadar, 2002). Proof includes not only understanding a concept and the mental processes, but also realizing how and why the concept definition and mental processes work (Tall, 1992). Mathematical proof and proving is central to improving mathematical thinking also advanced mathematical thinking and performing mathematics, comprehending structure and the nature of mathematical knowledge. Moreover, it is important for understanding historical evolvement and type of mathematical objects, the way of developing and sharing them with society and as an individual (Uğurel & Moralı, 2010). When we think about all these properties, proof and proving is important not only for providing justification for mathematical knowledge, but also

CONTACT: Burçak Boz Yaman ✉ burcakboz@gmail.com Mathematics and Science Education Department, Muğla Sıtkı Koçman University, Muğla, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

for doing and understanding mathematics. Therefore, it is necessary for constructing and developing mathematical knowledge and communicating mathematically (Stylianides, 2007). Learning proof is an important topic for all levels of education in order for mathematics education to be effective; however, research shows students have difficulties conducting and understanding proof. Sarı (2011) asserted that many researchers focused on existing problems and their elimination. Sarı (2011, p.19) listed the following problems researched in the literature [Note: citations in the following bullet points are 'as cited by Sarı (2011)'].

- Perceptions of proof and inadequate understanding of concept of proof, meaning of proof, role of proof, aim of proof, and necessity of proof (Alibert & Thomas, 1991; Almeida, 2000; Harel & Sowder, 2007; Knapp, 2005; Knuth & Elliot, 1997; Martin & Harel, 1989; Weber, 2006).
- Not knowing how to start a proof (Atwood, 2001; Baker & Campbell, 2004; Moore, 1994; Selden & Selden, 2007a).
- Inadequate knowledge about mathematical definitions, role and importance of definition in mathematics, and how to use them (Atwood, 2001; Edwards & Ward, 2004; Knapp, 2006).
- Insufficient information about a theorem or concept (Dreyfus, 1999; Hart, 1994; Ko & Knuth, 2009; Moore, 1994; Weber, 2006).
- Even with knowledge of theorem and concept, could not use them properly (Pedemonte, 2007; Selden & Selden, 2007a; Weber, 2001).
- Deficiencies about logic, and inadequacy of using quantifiers (Atwood, 2001; Baker & Campbell, 2004; Epp, 2003; Harel & Sowder, 2007; Selden & Selden, 2007a).
- Could not reach maturity and proficiency logically; could not follow chain of reasoning (Harel & Sowder, 2007; Knapp, 2005; Selden & Selden, 1995; Weber, 2001).
- Inadequately knowing mathematical proof methods and techniques, and not applying them correctly (Antonini & Mariotti, 2007, 2008; Goetting, 1995; Stylianides et al., 2004, 2007; Thompson, 1996; Wu Yu et al., 2003).
- Inability with mathematical language (differences between daily and mathematical language); make it difficult to understand mathematical language (Baker & Campbell, 2004; Epp, 2003; Ferrari, 2004; Selden & Selden, 2007a).
- Inability to write mathematical proof or explain thoughts (Dreyfus, 1999; Dubinsky, 2000; Ko & Knuth, 2009; Weber & Alcock, 2009).

One main point of the findings listed above, and also from other studies (Di Martino & Maracci, 2009; Hemmi, 2008; Remillard, 2010), are the knowledge and skill deficiencies in general mechanism and stages of a proof. This reveals the importance of understanding a mechanism of proof and its components. Consequently, the process of understanding/comprehending a proof and the process dynamics are fundamental to teaching proof.

1.1. Comprehending Proof

To evaluate understanding of a proof, students are usually asked to repeat the given proof or apply within a similar theorem (Weber & Mejia-Ramos, 2011). This evaluation approach makes the form of proof more important than the meaning of proof (Lin & Yang, 2007), and it depends more on memorizing than comprehending the proof. However, new learning approaches focus on conceptual learning (National Council of Teachers of Mathematics [NCTM], 2000; Ministry of National Mathematics Education [MoNE], 2013). How a proof is comprehended is essential, yet researchers have different ideas about understanding and comprehending a proof.

One model about how a proof is better comprehended is suggested by Leron (1983), who presented mathematical proofs in a step-by-step, one directional linear style, from hypothesis to conclusion. Leron emphasized the method's appropriateness to hold validity of proof, yet inadequate for communicating mathematical knowledge. He claimed that proofs restructured as short, independent modules emphasizing specific knowledge/ideas are comprehended better, and he introduced the structural model of teaching.

Selden and Selden (1995) stated that before comprehending a proof as a whole, comprehending expression of a proof is more important. Mejia-Ramos (2008) divides reading proof activities into "understanding proof" and "evaluating proof"; illustrating mathematics textbook proof reading activities as an example for understanding proof, and teacher assessment of proof for evaluating proof. Mejia-Ramos (2008) stated that proof reading activities should not only control proof validity, but also focus on understanding the context of that proof. In understanding a proof, Weber and Mejia-Ramos (2011) expressed just knowing the proof steps is inadequate; understanding a proof logically is central to comprehending a proof. Duval (2002) stated three kinds of learning occur when comprehending a proof. First one is learning the meaning of terms, symbols or shapes used in a proof. The second knowledge is inserting expressions in proof steps; deciding which statements are preliminary, definition, or conclusion. Before deciding the required statement of proof, students cannot decide where to start or end. The last knowledge is to be able to explain transition among proof steps.

Stylianides (2007) defined mathematical argument as "Proof is a mathematical argument, a connected sequence of assertions for or against a mathematical claim" (p.291), stating that comprehending an argument as a proof requires a four-way evaluation:

Basic: Comprehending statements (like definition, axiom) that constitute a proof and understanding the roles in a proof.

Formulation: Comprehending proof development and what generalization could be logically conducted in proof steps.

Presentation: Comprehending language used in expressing a proof. A comprehended proof can be expressed in a student's own words.

Social dimension: Satisfying the truthfulness of a proof for each individual. Each presented proof should be appropriate for a group's academic level, with each group member convinced of the proof's truthfulness.

Another holistic approach on proof comprehension is presented by Yang and Lin (2008) and Yang, Lin, and Wang (2008). Lin and Yang (2007) suggest a model for reading comprehension of geometry proof, including learning to comprehend a proof, comprehending levels generated in such learning, and different question types to identify levels of comprehension. They explain that "reading is not only recognition of words and recall of their meaning, but also an active and constructive process between readers, media and contents" (Yang et al., 2008, p.80). However, comprehending a proof is explained as "reading comprehension of proofs means understanding proofs from the essential elements of knowing how a proof operates and why a proof is right, in addition to knowing what a proof can prove" (Yang & Lin, 2008, p.60). According to this model, students should first recognize premises, then use premises to construct a connection between results, and finally combine premises and results to construct new comprehension. Based on this theoretical structure, four levels identify how a proof is read by comprehending. Among all the other proof comprehending models Yang and Lin (2008) give a well-designed and multi-dimensional structure which is easy to evaluate and follow students' proving processes. The details of the model will be presented below.

1.2. Model for Reading Proof by Comprehending

Yang and Lin (2008) constructed a four-level model (Figure 1). The first level (Surface) is to grasp the meaning of mathematical terms, symbols, or figures in a proof. The second level (Recognizing Elements (pieces)) defines the logical state of the expressions (obvious or latent), and includes recognition of premises, conclusions. The next level (Chaining Elements (relations)) is comprehending and combining logical arguments in a proof. The final level (Encapsulation) is deciding how to conduct a proof in another situation and internalizing propositions of a proof. They define encapsulation as "a developmental situation without end" (p.71), stating their model for comprehending a proof is aimed at identifying students who reach this last level. In their multidimensional model, Yang and Lin (2008) construct "facets" to organize the necessary learning in switching between the four levels of Figure 2.

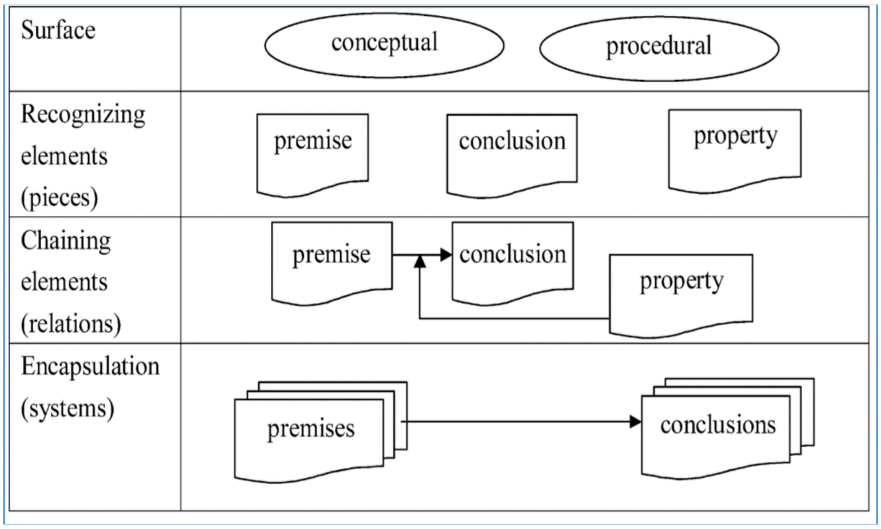


Figure 1. Proof reading comprehension model (Yang & Lin, 2008, p.63)

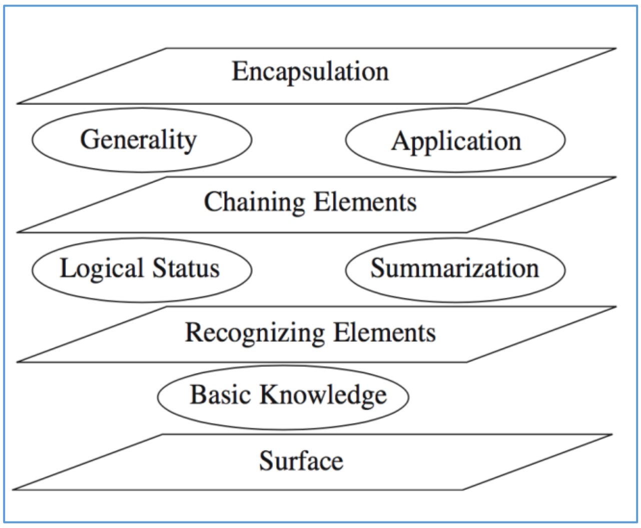


Figure 2. Proof reading comprehension theoretical model (Yang & Lin, 2008, p.71)

The model explains a five-faceted structured. The facets are pretending as a passage between two levels. A person who hold the knowledge of the related facet can move on the next level. For instance, the first facet "Basic Knowledge" is needed to move up to the Recognizing Elements level. This facet measures understanding of mathematical terms, figures, and symbols in premises and proofs. The second facet (Logical Status) and third one (Summarization) are needed to switch from Recognizing Elements to Chaining Elements. Logical Status requires the recognition of arguments as premises, conclusion, or applied properties in a proof. Summarization defines the core or critical idea in a proof. The fourth (Generality) and fifth (Application) facets are necessary to switch from Chaining Elements to Encapsulation. Generality identifies the accuracy of a proposition and understanding what a proof will prove. Application requires the application of proven proposition in another situation. These five facets and four levels construct a model for comprehending a proof.

In the current study, a Proof Comprehending Test (PCT) conducted based on comprehending a proof model is used to identify the degree of students' comprehension of the five facets. Table 1 explains the learning objects used and which learning behavior occurs in the constructed PCT to reveal component-level comprehension of Yang and Lin's (2008) multidimensional model.

Table 1. Structure of reading geometric proof by comprehending (Yang & Lin, 2008)

Facet	Object of comprehension	Operational definition
Basic Knowledge	Content of premise or conclusion	Recognizing the meaning of a symbol Explaining the meaning of a property Recognizing the meaning of a property
Logical Status	Status of premise	Cognizing a condition applied directly
	Logical relation between premise and conclusion	Judging logical order of statements
	Property applied to derive conclusion from premise	Recognizing which properties apply
Summarization	Multiple arguments and critical ideas	Identifying critical procedures, premises, or conclusions Identifying critical ideas of a proof
Generality	Proposition or proof	Judging correctness
	All arguments and attached figure	Identifying what a proof validates
Application	Application in same premise	Application in same premise
	Identifying different premises	Identifying different premises

1.3. Proof Comprehension Tests

It should be noted that, although many research studies prefer "comprehension test" over "proof comprehension test", PCT was chosen for the current study to narrow down its usability. Houston (1993a, 1993b) was the pioneer whose research studies directly conducted comprehension tests. Houston (1993a) used writing comprehension test in a college mathematics course to develop and evaluate student understanding of reading and writing ability for mathematical modelling texts. Houston was inspired by a comprehension test used in an English course with students given a text and asked questions about it. Then he conducted comprehension tests with specific text from a mathematics curriculum. He let students work on the prepared text individually or as a group in order to understand the text, and then applied questions he had prepared to understand their text comprehension.

Conradie and Frith's (2000) long-term research studied comprehension testing at Cape Town University, South Africa, for freshman to senior-year students. The researchers applied similar comprehension tests to Houston (1993a, 1993b), but their tests used proof as a text. The basic properties of Conradie and Frith's (2000) comprehension test study was that theorems were presented with their proofs and students questioned on properties of the proofs. Comprehension test philosophy is that during application, students' understanding can be deeply investigated, and that learning with memorizing is prevented. Conradie and Frith (2000) specified comprehension test advantages as;

- It encourages to understand theorems and proofs rather than memorization.
- A comprehension test gives a far more precise evaluation of a student's understanding at all levels.
- Improves the quality of feedback of both teacher and student.
- According to classical methods it is less frustrating.
- Improves mathematical communication skills.

According to them, comprehension test uses testing to understand; special steps in a proof, structure of a proof, concepts used in proofs, results of the assumptions, and critical perspectives of a proof. On the other hand, Conradie and Frith (2000) list some disadvantages of using PCT;

- Need more time comparing with the traditional methods.
- It may prevent some students' interest in theoretical part of the lesson.
- Students may think they cannot prepare for comprehension tests.

Besides Conradie and Frith (2000), Yang and Lin (2008) developed an instrument for measuring Reading Comprehension of Geometry Proof (RCGP) based on multidimensional model of comprehending a proof with four levels. Whilst Conradie and Frith (2000) do not specify criteria in the context of the questions asked to students, the PCT (we called RCGP as Proof Comprehension Test –PCT) developed by Yang and Lin (2008) has questions matched to each level, and their model is also appropriate for evaluating learning behaviour.

Although PCTs are functional tools for comprehending proofs, there has been limited research to date on the teaching and learning of proofs. The current study is aimed at bridging part of this knowledge gap.

1.4. Basis for the Study

Yang and Lin are pioneers who used PCT effectively for understanding/ comprehending proofs by students and led more people to use this tool for teaching proofs. They suggested a four-level model for comprehending proofs in a series of studies (Lin & Yang, 2007; Yang, 2012; Yang & Lin, 2008), and produced a proof comprehension test based on these levels. Yang and Lin (2008) used only one PCT (one geometry proof with 16 questions) to conduct their model. They then (Yang, 2012; Yang & Lin, 2008) developed PCT to investigate the functionality of models and analyze the relationship of students' geometrical knowledge and logical reasoning. They did not specify the tool as a (proof) comprehending test, but when the format of the proof and the related questions are considered, it can be seen as a PCT that has been systematically elaborated and applied to the developed model.

Roy, Alcock, and Inglis (2010) used Yang and Lin's (2008) proof comprehending tests at the undergraduate level to investigate the effects of presenting proofs in different forms on comprehension. Alcock and Wilkinson (2011) designed an electronic (e-)proof to support mathematical proofs for undergraduates, based on Yang and Lin's (2008) levels of comprehension. Similarly, Mejia-Ramos, Fuller, Weber, Rhoads, and Samkoff (2012) developed another model involving different comprehension levels for undergraduates based on the model developed by Yang and Lin (2008). In a study conducted with middle school preservice teachers, Zazkis and Zazkis (2014) tried to identify how preservice teachers evaluate students' understanding of proof. Preservice teachers were asked to construct probable "proof scenarios", with realistic dialogue between students and teacher based on proof comprehension levels prepared by Mejia-Ramos et al. (2012). In Zazkis and Zazkis' (2014) study, the analysis was conducted based on the third level of the model, "Justification of claims". In Turkey, Yıldız (2006) conducted a proof comprehension test with four preservice mathematics teachers and analyzed their thoughts regarding the test. Another study on proof comprehension test was conducted by İnam & Uğurel (2016). The researchers investigated difficulties of teachers who conducted a PCT-based secondary school mathematics course, examining teachers' interventions and their effectiveness.

Although the area of proof teaching has been much researched, there are limited studies on the comprehension of proof based on PCT. Most studies have been at college level and based on a single proof comprehension test. Most studies about PCT (e.g., Conradie & Frith, 2000; Houston, 1993a, 1993b; Yang & Lin, 2008) are not process-centered, but mostly focus on situational explanation. The purpose of the current study is examining the 11th grade students' proof comprehension levels based on Yang and Lin's (2008) model. The corresponding research question for this current study is, "What is the 11th grade students' performance in proof comprehension tests based on quadrilateral?"

Therefore, the current study will contribute to the literature as a qualitative study with a teaching application, and as an action study applied within a secondary school. We believe that the current research will address a gap in the proof comprehension test literature and aid new research.

2. METHOD

This study forms part of a comprehensive qualitative research project. The main project is planned within a qualitative paradigm and constructed as an action research design. According to Koshy (2005), the first action research step is identifying the topic of study, then the group to be studied is identified. The basic concepts/knowledge should be constructed. The tools that help follow the process are clarified and the plan conducted. Once the collected data is examined, the plan is redesigned according to the results. This cycle continues until the aim of the study is reached.

In the main large scale project, PCT-based teaching is conducted in a secondary school classroom for five weeks and the process is evaluated from many perspectives (evaluating teaching process, opinions on PCT, performance on PCT, etc.). The current study, as part of the overall project, investigated the performance of students on 5 PCTs conducted for five weeks.

2.1. Participants

The participants were selected based on typical case sampling method, which involves identifying "typical" among a series of cases which helps to introduce a new application or novelty (Yıldırım & Şimşek, 2013). In this method, "the critical part is selecting average and typical, not extraordinary" (Büyüköztürk, Kılıç-Çakmak, Akgün, Karadeniz, & Demirel, 2014, p.91). The current study was conducted with 20 students (6 males, 14 females) aged 16 years attending 11th grade at a state high school in Zonguldak province, Turkey, and one mathematics teacher (one of the researchers). While selecting the study site, perspectives considered were; convenience of reaching the participants, a school teaching the reformed geometry curriculum, and voluntary student participation. Selection from among the teacher's geometry classes was dependent on heterogeneity of geometry achievement and volunteering. Before conducting the study, the teacher briefed the students about PCT and the research process. No PCT application or investigation was conducted at this stage. Participants had some basic knowledge about proof and proving based on their 10th grade mathematics curriculum, but not proof comprehension testing. Participants' cumulative Grade Point Average (GPA) was 3.79 (out of 5) at 9th grade and 3.71 at 10th grade. All students had selected mostly science and mathematics courses after 9th grade.

The teacher accompanied the students throughout the study process. He taught mathematics and geometry to the participants since 9th grade and is their classroom teacher, so therefore knowledgeable about the students' academic development. Since the teacher knows each student well, it is perceived that during the PCT process its effect can be easily monitored. The teacher is knowledgeable about PCT, having studied PCT for two years by reviewing the literature, analyzing PCT examples, developing, implementing and evaluating PCT, and hosting a seminar for other teachers. Application sessions were organized based on student school schedules. Students were assured recordings were for academic purposes only, with real names replaced by pseudonyms and gathered data not used for purposes beyond academic aims.

2.2. Data Gathering Tools (PCT)

In the big scale study there are five different types of data gathering tools; pre and post free writings of students, teacher-researcher's reflective journals, students' interviews and constructed PCTs. On the other hand, in the current study the constructed PCTs are considered as data gathering tool.

A literature review was conducted while producing the data gathering tools and other proof comprehension tests examined. PCT which consists of all secondary school geometry topics, prepared from examples in the literature and the national mathematics curriculum. The researchers intensely examined 9-12th grade Geometry Curriculum, textbooks, other resources to identify theorems and premises appropriate and functional for PCT. Based on teaching experience, the teacher-researcher selected proof problems according to difficulty level, background information, and classroom applicability in a reasonable timeframe. Eight theorem were initially identified based on these properties. The selected eight theorems were examined by another content specialist for applicability and transferability as a PCT item, and consequently five of the theorems selected. The selected theorems are about "quadrilateral" in the 11th grade Geometry Curriculum of the Ministry of National Education (MoNE, 2010).

There are three process standards are in the related unit, with two selected for the study: "Process standard 2: Prove the theorems about quadrilaterals and conduct applications" (MoNE, 2010, p.32), and "Process standard 3: Calculate the circumstance of quadrilateral and prove theorems about area of quadrilaterals and conduct applications." (MoNE, 2010, p.33). In the curriculum explanation, the learning behaviors mentioned in the process standards, their scope and limits are described. According to this framework, content of the produced PCTs, related process standards and curriculum explanations are as follows;

PCT-1 is prepared for conclusion of "the sum of interior angles of quadrilateral is 360 degrees" (process standard 2) (PCT-1 is given at Appendix).

PCT-2 is prepared for the theorem "the angle produced by two angle bisectors from adjacent two interior angles of a quadrilateral is equal to half of the summation of other two quadrilateral's angles." (process standard 2).

PCT-3 is prepared for the explanation "In any ABCD quadrilateral, if the diagonals are intersecting perpendicularly then the addition of square of opposite sides are equal to each other." (MoNE, 2010, p.32) (process standard 2).

PCT-4 is prepared for the explanation "Prove that the area of convex quadrilateral region is equal to half of the multiplication of diagonals' length and sine of an angle between diagonals" (MoNE, 2010, p.33) (process standard 3).

PCT-5 is prepared for the explanation "An area of a quadrilateral whose corners are the mid points of sides of a quadrilateral is half of the quadrilateral" (MoNE, 2010, p.32) (process standard 3).

After these stages, the selected five theorems were reconstructed according to PCT format. The draft PCT forms were applied to ten 11th grade students not participating in the study. According to both written answers and informal interviews, the PCT were reviewed again for understandability, difficulty, and practicability. Additionally, three teachers were asked to examine the PCTs. Prior to their examination, the teachers were informed about PCTs, given examples, and the aim of the study and problems explained. Afterwards, based on feedback, ideas, and suggestions, the PCTs were reconstructed to become the original forms. These original forms were then re-examined by the three teachers and the researchers of the current study. Final changes were then applied to form the final version for application.

2.3. Application of PCT

Five PCTs were applied to 11th grade students for a period of five weeks in two class hours (40 minutes each). In the first application session students were given a worksheet containing PCT-1. They were briefly informed about PCT before the application. They started to answer questions in a small group, but they had many questions about PCT since this was their first experience attempting to answer such proof questions. In the second class hour, it was realized that the students were experiencing difficulties in answering the questions; therefore, PCT-1 was shown on the projection screen and a classroom discussion held for each question. In the following week, PCT-2 was given to the students as a worksheet, and this time they were tasked with answering questions by themselves. In this session the students pointed out that when they could not answer one of the steps in a question, they were unable to move on to the next. This criticism was taken into consideration and the next PCTs were redesigned accordingly. In PCT-3 the application procedure changed back to being a group study followed by individual study, with restricted time allowed and options given for group discussions. If the students wanted to work as a group they were permitted to discuss the PCT within the group for 10 minutes, after which they had to answers the questions by themselves or continue studying by themselves during the remainder of the class hour. PCT-4 application was also conducted in the same way, with 10 minute-group-discussion, followed by 30 minutes for individual question answering. In the final application, the questions were answered by individuals during the first class hour and then in the second class hour the teacher/researcher presented another theorem from the textbook. The application was ended with this last PCT.

2.4. Data Analysis

A mixed data analysis is used for the current study (Creswell, 2003). As a qualitative analysis part, Yang and Lin's (2008) analysis methods were replicated to analyze the PCT. Each question in the comprehending test is related to facets of Yang and Lin's (2008) model, with learning goals identified for each question. Therefore, students' written answers gathered from tests are first examined and classified through the model.

As a quantitative part of the data analysis process, questions were coded as 0, 1, or 2 according to the degree of reaching the determined learning goals. In Table 2, Evaluation

Criteria for PCT-1 is given as an example. If the answer fully meets the determined learning goal, it was coded as '2'; with '1' for partially meeting the determined learning goal; and '0' for not meeting any of the learning goals. Each comprehension test was examined and graded by the researcher-teacher, and a teacher working at the same school to achieve grading reliability. Afterwards, the teachers compared their grading and a final grading agreed. Finally, random selection of eight PCT coded by the second author and the results were compared for coding reliability.

Table 2. PCT-1 Evaluation Criteria

Facet	Learning Goal	Question	Grades
Basic Knowledge	Defining terms in a proof	1	0,1,2
	Questioning truthiness of properties in a proof	2	0,1,2
	Explaining applied property	3	0,1,2
Logical status	Verifying logical orders in a proof	4	0,1,2
	Verifying logical orders in a proof	5	0,1,2
Summarization	Identifying critical step(s) in a proof	6	0,1,2
Generality	Questioning the truthiness of a proof	7	0,1,2
	Explaining truthiness of a proof	8	0,1,2
Application	Conducting a proof in a different way	9	0,1,2
	Applying a proof to different situations	10	0,1,2

Next, the performance percentages for levels of proof comprehension were identified. The highest grades achievable for questions regarding a comprehension facet were determined, and then the grades achieved were used to calculate the percentage. Finally, as with Yang and Lin's (2008) analysis, the calculated performance percentages were evaluated in three groups. Students with performance percentages of 0-33% from comprehension levels were classified as the 'low group'; with 34-66% as the 'medium group'; and 67-100% as the 'high group'.

As an example, Questions 9 and 10 in PCT-1 (Appendix) are designed concerning the Application facet. Student S6 could not answer Question 9 and scored 0, but scored 2 by correctly answering Question 10 (see Figure 3). The highest score achievable from both questions was 4 points. The comprehension level of this student is therefore 50%, with 2 points scored out of 4. Finally, since the percentage scored is 33-66%, the student is placed in the 'medium group' for the Application facet.

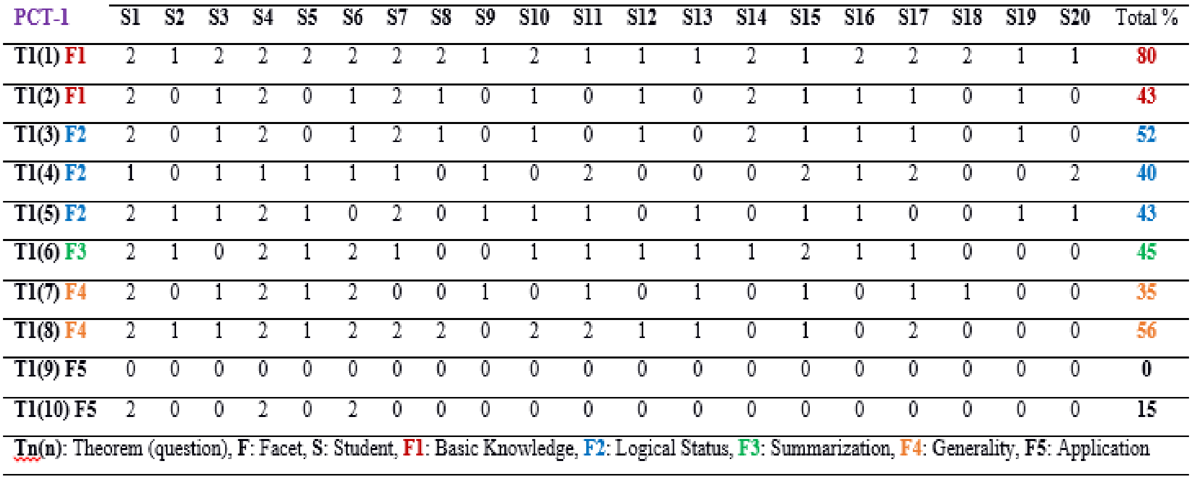
3. FINDINGS

3.1. Analysis Results: PCT-1

PCT-1 was prepared for 'process standard 2' of quadrilateral unit, taken from the 11th grade geometry curriculum (MoNE, 2010). The theorem is the "sum of interior angles of a quadrilateral is 360 degrees", which is from the explanation part of the related process standard. Students' answers for each questions were graded as 0-1-2 (see Data Analysis section) and performance percentages then calculated. Ten questions were prepared for the proof presented in PCT-1 (see Appendix). Answers to these questions were evaluated according to determined facets. Results are presented in Table 3.

When Table 3 is examined for Questions 1 and 2 on the Basic Knowledge (F1) facet, regarding symbols and statements of proof, participant performance is medium (approximately (80+43)/2=62%). Performances are at the medium level (45%) for three facets; Logical Status (F2) on comprehension of passes among proof steps, Summarization (F3) on comprehension of critical ideas in a proof, and Generality (F4) in which accuracy of proof is questioned.

Table 3. PCT-1 scores and total percentages by facet



PCT-1	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Total %
T1(1) F1	2	1	2	2	2	2	2	2	1	2	1	1	1	2	1	2	2	2	1	1	80
T1(2) F1	2	0	1	2	0	1	2	1	0	1	0	1	0	2	1	1	1	0	1	0	43
T1(3) F2	2	0	1	2	0	1	2	1	0	1	0	1	0	2	1	1	1	0	1	0	52
T1(4) F2	1	0	1	1	1	1	1	0	1	0	2	0	0	0	2	1	2	0	0	2	40
T1(5) F2	2	1	1	2	1	0	2	0	1	1	1	0	1	0	1	1	0	0	1	1	43
T1(6) F3	2	1	0	2	1	2	1	0	0	1	1	1	1	1	2	1	1	0	0	0	45
T1(7) F4	2	0	1	2	1	2	0	0	1	0	1	0	1	0	1	0	1	1	0	0	35
T1(8) F4	2	1	1	2	1	2	2	2	0	2	2	1	1	0	1	0	2	0	0	0	56
T1(9) F5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T1(10) F5	2	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15

Tn(n): Theorem (question), F: Facet, S: Student, F1: Basic Knowledge, F2: Logical Status, F3: Summarization, F4: Generality, F5: Application

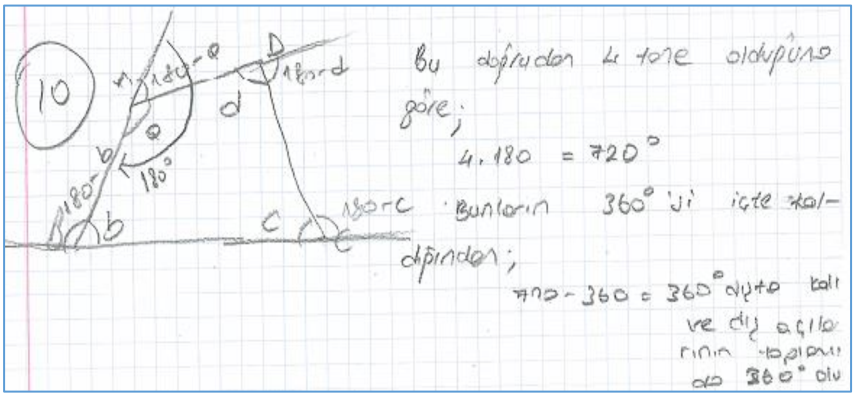
However, performance for the Application (F5) facet, about applicability of a proof in different situations, is low level (approximately 7%). It is observed that the higher the comprehension level, the lower the students' comprehension percentages.

When each question is examined, the highest comprehension percentages occurred in Question 1 (80%), regarding knowledge of terms used in a proof. The lowest comprehension level is for Question 9 (0%), regarding conducting proofs in different ways. According to these results, it is observed that students comprehend prerequisite knowledge like definition, figures, and symbols, but are poorer at comprehending conducting proofs in different ways.

In Questions 3-5, in which transitions of logical relationships among proof steps are questioned, although students reached similar comprehension percentages, they are unsatisfactory. In each question, different comprehension percentages are seen. Accordingly, students do not have the same comprehension performance for all proof steps, and may comprehend one step transition, but not the next. Since the structure of proofs and each proof step has different functionality, this result is accepted as natural.

Question 6 in PCT-1 is about identifying critical steps on which proof is based. The performance percentage for Question 6 is also unsatisfactory (45%). According to this result, although students comprehend basic information like definitions and symbols, they performed poorly on identifying the basic foundation of a proof. Question 6 asked, “According to you, which steps are the critical steps for this proof?”. S17 answered; “3rd and 4th steps; because if we do not know the sum of interior angles of a triangle we cannot conduct the proof”, showing that S17 understood the critical idea of the proof.

For Question 10, by using the given proof in PCT-1, students were asked to show the sum of exterior angles of a quadrilateral. Only three students could answer this (S1, S4, S6). After the classroom intervention, during student interviews they confessed that they saw this proof before and could therefore answer it. S6's answer is presented in Figure 3 (Student's written answer in Figure 3: Since there are four of this line 4.180=720^o. Among 360^o of it remains inside so that 720-360=360 belongs to exterior angles and the sum of exterior angles is 360^o).


10
180-a
a
180-d
D
d
b
180°
180-b
b
c
180-c

Bu doğrudan 4 tane olduğuna göre;
4.180 = 720°
Bunların 360°'ni içte kal-
dığından;
720-360 = 360° dışta kalı
ve dış açıla
rının toplamı
da 360° olu

Figure 3. Answer by S6 for PCT-1 Question 10

Table 4 presents the proof comprehension levels for each students according to according to determined percentages. As explained before students are labelled as low with performance percentages of 0-33%, medium with 34-66% and high with 67-100%.

Table 4. PCT-1 participant evaluation results

Level	Comprehending Degree	Student	Frequency	Percentage
Surface	Low	S2,S9,S11,S13,S20	5	25
	Medium	S5,S12,S15,S18,S19	5	25
	High	S1,S3,S4,S6,S7,S8,S10,S14,S16,S17	10	50
Recognizing Elements	Low	S2,S5,S8,S10,S12,S13,S14,S18,S19	9	45
	Medium	S1,S4,S6,S9,S16,S17	6	30
	High	S3,S7,S11,S15,S20	5	25
Chaining Elements	Low	S2,S3,S7,S8,S9,S10,S13,S18,S19,S20	10	50
	Medium	S5,S11,S12,S14,S16,S17	6	30
	High	S1,S4,S6,S15	4	20
Encapsulation	Low	S2,S3,S5,S7,S8,S9,S10,S11,S12,S13, S14,S15,S16,S17,S18,S19,S20	17	85
	Medium	S1,S4,S6	3	15
	High	-	0	0

Table 4 shows that student percentages at high comprehension levels are mostly at the Surface level, and low comprehension levels found mostly at the Encapsulation level which involves conducting proofs in different ways. Aligned with this result, student comprehension performance descends from Surface level to Encapsulation. It is observed that for PCT-1's theorems and proofs, students comprehend definitions and symbols in proofs, but inadequately performed in levels involving high degrees of comprehension.

3.2. Analysis Results: PCT-2

PCT-2 was prepared for the quadrilateral unit given in 'process standard 2' of the 11th grade geometry curriculum (MoNE, 2010). The theorem is "the measure of the angle of an intersection of bisectors belonging to two adjacent interior angles of a quadrilateral is equal to half the sum of the other two angles", and is given in the explanation part of the curriculum. Ten questions were written for the proof in PCT-2. Students' answers are evaluated according to the related facets and presented in Table 5.

When Table 5 is examined, participant performance percentage is seen as medium (64%) for Questions 1-3 on the Basic Knowledge (F1) facet, regarding symbols and statements of proof. The performances for the Logical Status (F2) facet, on comprehension of transition among proof steps, are of medium level (approximately 37% where (30+35+45)/3), but low level for the Summarization (F3) facet (8%), on comprehension of critical ideas in a proof, the Generality (F4) facet (11%), on the accuracy of proof, and the Application (F5) facet (0%), on the applicability of a proof in different situations. Noteworthy is the Application facet where all students presented unsatisfactory performance, hence the percentage is zero.

Other findings reached from Table 5 are that when the comprehension level progresses, there is no identifiable pattern of movement, increasing or decreasing. A decrease is observed in the transition from Basic Knowledge to Logical Status to Summarization facets, but increases when passing through the Generality facet.

Table 5. PCT-2 scores & total percentages by facet

PCT-2	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Total %
T2(1) F1	1	1	1	0	1	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	20
T2(2) F1	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	2	1	1	83
T2(3) F1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	1	2	1	1	90
T2(4) F2	1	0	1	2	0	0	0	0	0	1	0	2	0	0	2	2	0	0	0	1	30
T2(5) F2	1	0	1	2	0	0	0	0	0	1	0	2	1	0	2	2	0	0	1	1	35
T2(6) F2	1	0	1	2	0	0	1	1	1	1	1	1	2	1	2	2	0	0	1	1	45
T2(7) F3	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8
T2(8) F4	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8
T2(9) F4	1	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	0	0	0	28
T2(10) F5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tn(n): Theorem (question), F: Facet, S: Student, F1: Basic Knowledge, F2: Logical Status, F3: Summarization, F4: Generality, F5: Application

It is natural to find no linear decrease or increase on comprehension performance towards different facets when the multidimensional structure of proof and proving is considered.

Question 1 in PCT-2 is about defining the term, "bisector". The comprehension level for Question 1 is the lowest (20%) among all questions on the Basic Knowledge facet, and the highest is for Question 3 (90%) regarding the sum of interior angles of quadrilaterals. Since the proof used in PCT-1 is about sum of interior angles of a quadrilateral, the result from Question 3 in PCT-2 may reflect PCT-1. As an example, S5's answer to Question 3 ("Do you agree with the equality $m(\hat{A})+m(\hat{B})+m(\hat{C})+m(\hat{D})=360$ degrees given in the proof? Why?") was "I agree, because the angles $m(\hat{A})$, $m(\hat{B})$, $m(\hat{C})$, $m(\hat{D})$ construct a quadrilateral. Since the sum of interior angles is 360 degrees, then $m(\hat{A})+m(\hat{B})+m(\hat{C})+m(\hat{D})=360$ degrees". As seen, student S5 comprehended the basic knowledge needed to complete the proof.

PCT-2 asks the proof validity when the bisector of angles are drawn from different vertices. The comprehension level percentages are the lowest (30%) for Questions 4-6, which relate to the Logical Status facet. S4 responded “No” for Question 4 (“what if the bisectors of theorem intersect outside of the quadrilateral region, is the theorem still true?”). S4’s explanation is shown in Figure 4.

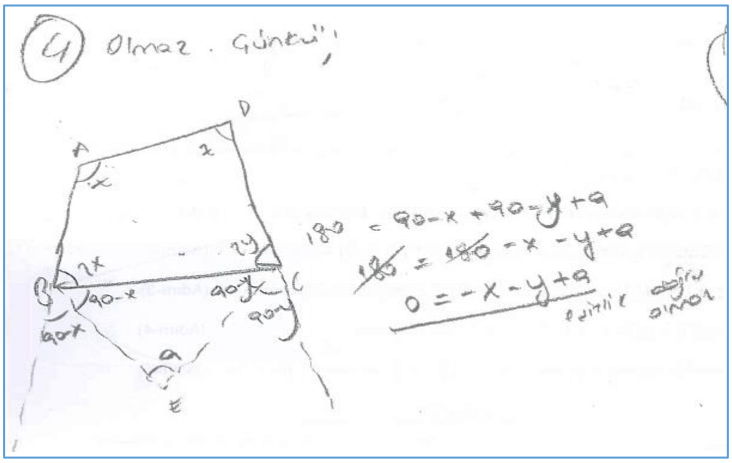


Figure 4. Answer by S4 for PCT-2 Question 4

According to results for the Logical Status facet, it is understood that students could not comprehend the logical relationships of proof steps. Considering the percentages for Question 1, it may be concluded that students lack understanding about bisector which affects the next steps’ comprehension about bisector.

In PCT-2 Question 10, students must prove the given theorem in different ways, but no student could answer this question. S10’s answer (see Figure 5) shows the proof simply conducted in the same way again. Student comprehension levels are individually presented in Table 6 based on the data for each facet and levels, and the table shows each student’s performance level and degree.

Table 6 shows that most students present high comprehension performance for Surface Level; the knowledge of statements and symbols for proofs. However, in other comprehension levels, percentages decreased from 20% to 0%. Different from PCT-1, results for PCT-2 present a decreasing pattern for different comprehension levels. In the Encapsulation level, which involves proving a proof in a different way or conducting another proof depending on previous proof comprehension, no high or medium degree of comprehension occurred.

Another remarkable result is that the same student may present different performances degrees for different comprehension levels (e.g., S17 is medium for Surface level, but low for Recognizing Elements). Accordingly, it can be concluded that students did not perform the same for all stages of a proof.

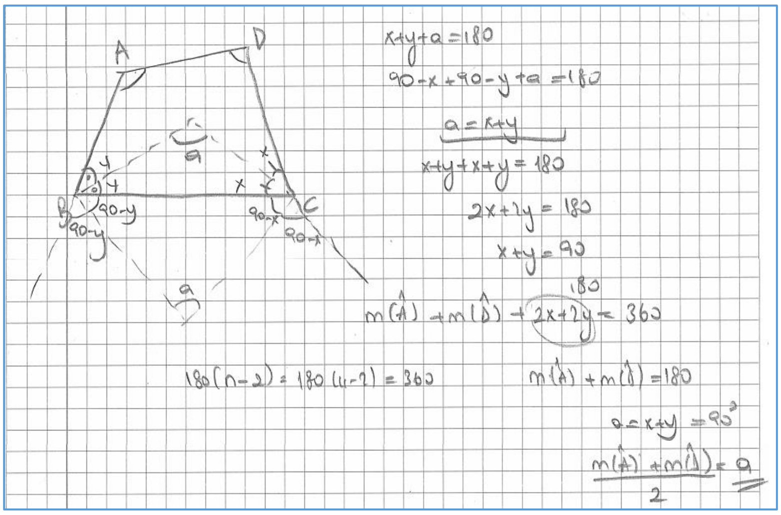

A
D
B
C
y
y
x
x
a
a
90-y
90-y
90-x
90-x
x+y+a=180
90-x+90-y+a=180
a=x+y
x+y+x+y=180
2x+2y=180
x+y=90
180
m(Â)+m(D̂)+2x+2y=360
180(n-2)=180(4-2)=360
m(Â)+m(D̂)=180
a=x+y=90°
m(Â)+m(D̂)/2 = a

Figure 5. Answer by S10 for PCT-2 Question 10

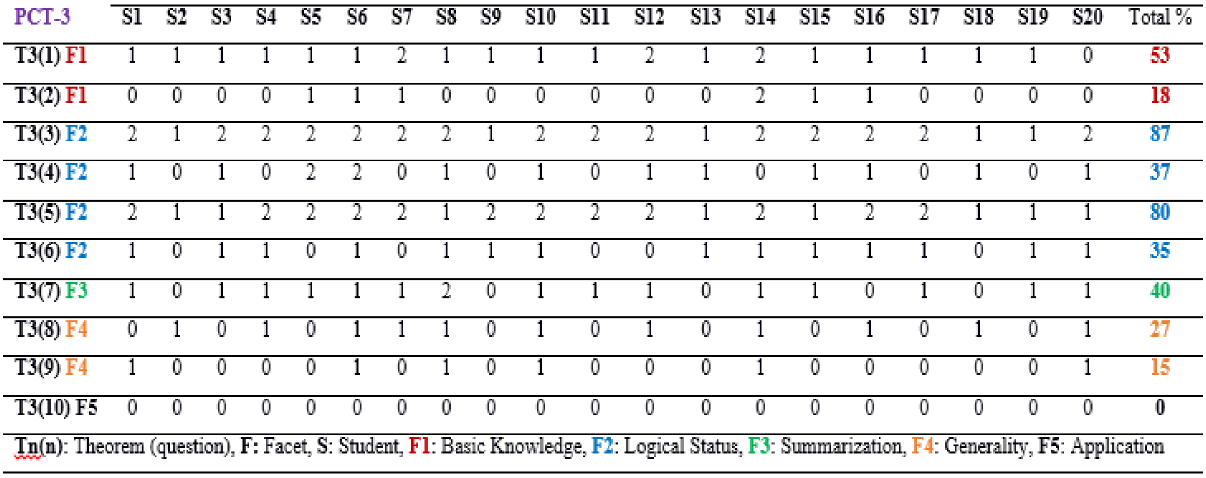
Table 6. PCT-2 participant evaluation results

Level	Comprehending Degree	Student	Frequency	Percentage
Surface	Low	S7,S19,S20	3	15
	Medium	S17	1	5
	High	S1,S2,S3,S4,S5,S6,S8,S9,S10,S11, S12,S13,S14,S16,S18	15	75
Recognizing Elements	Low	S2,S5,S6,S7,S8,S9,S10,S11,S13,S14, S17,S18,S19	12	65
	Medium	S1,S3,S10,S20	4	20
	High	S4,S12,S15,S16	4	20
Chaining Elements	Low	S2,S3,S4,S5,S6,S7,S8,S9,S10,S11, S12,S13,S14,S15,S16,S18,S19,S20	18	90
	Medium	S1,S17	2	10
	High	-	0	0
Encapsulation	Low	S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11 S12,S13,S14,S15,S16,S17,S18,S19,S20	20	100
	Medium	-	0	0
	High	-	0	0

3.3. Analysis Results: PCT-3

PCT-3 was prepared for quadrilateral unit in ‘process standard 2’ of the 11th grade geometry curriculum (MEB, 2010). The theorem is “Prove that in any quadrilateral ABCD, if the diagonals are perpendicular to each other, then the sum of the square of opposite sides of the quadrilateral are equal”, and is given in the geometry curriculum explanation. The 10 questions related with the proof and students' scores according to determined facets are presented in Table 7.

Table 7. PCT-3 scores & total percentages by facet



PCT-3	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Total %
T3(1) F1	1	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	1	1	1	0	53
T3(2) F1	0	0	0	0	1	1	1	0	0	0	0	0	0	2	1	1	0	0	0	0	18
T3(3) F2	2	1	2	2	2	2	2	2	1	2	2	2	1	2	2	2	2	1	1	2	87
T3(4) F2	1	0	1	0	2	2	0	1	0	1	0	1	1	0	1	1	0	1	0	1	37
T3(5) F2	2	1	1	2	2	2	2	1	2	2	2	2	1	2	1	2	2	1	1	1	80
T3(6) F2	1	0	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	1	1	35
T3(7) F3	1	0	1	1	1	1	1	2	0	1	1	1	0	1	1	0	1	0	1	1	40
T3(8) F4	0	1	0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	27
T3(9) F4	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	1	15
T3(10) F5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tn(n): Theorem (question), F: Facet, S: Student, F1: Basic Knowledge, F2: Logical Status, F3: Summarization, F4: Generality, F5: Application

Table 7 shows participant performance percentage as medium (53%) for Question 1 on Basic Knowledge (F1) regarding information about symbols and statements of proof, but very low (18%) for Question 2. The highest performance percentage (approximately 60% where (87+37+80+35)/4) was for the Logical Status (F2) facet, regarding comprehension of passes among proof steps. Low level performance is also seen for Summarization (F3) facet (40%), regarding comprehension of critical ideas in a proof, Generality (F4) facet (21%) on accuracy of proof, and Application (F5) facet (0%), regarding applicability of a proof in different situations. In Table 6, similar to PCT-2, no students present satisfactory performances in the Application facet.

The most remarkable result is that although students presented lower performances in the Basic Knowledge facet, regarding comprehending statements and symbols in proofs, students presented higher comprehending performances for the Logical Status facet which involves comprehending transition among proof steps. This result shows that although students could not comprehend the basic concepts, they could comprehend the next stages of the proof.

From Question 1 and 2 related to Basic Knowledge facet, the lowest comprehending percentage is 18% for Question 2. Only one student answered correctly about whether diagonals of a quadrilateral are always perpendicular to each other. S14 answered "No, not intersect perpendicular. Because it changes from quadrilateral to another type of quadrilateral. In some quadrilaterals they intersect perpendicular, but not all of them".

However, in the same facet a 53% comprehension level was obtained for Question 1, regarding Pythagorean relation which students are familiar with. Accordingly, it can be concluded that students may not perform at the same comprehension level for basic concepts given in a proof.

High levels of comprehension were seen in Question 3 (87%) and Question 5 (80%), regarding transitions among proof steps, but low levels were seen in Question 4 (37%) and Question 6 (35%), with the same logical perspectives which questioned the possibility of conducting a proof in different ways. This result is remarkable because of the distinct variation seen in the same comprehension facet, showing that students can interpret given steps but cannot produce these steps in different ways.

Question 6 asked in cases where diagonals intersect perpendicularly, would it be possible to produce a proof, and Question 7 asked what critical steps a proof depends on. Although the questions are in different comprehension facets, neither had satisfactory comprehension levels (Q6: 35%, Q7: 40%), with results correlated as the questions are related.

Similar to PCT-2, Question 10 involved conducting proof in different ways for the Application facet, and no student could answer it. Students faced difficulties in this facet as it measures the ability of knowing how to apply a proposition in another situation, and making this facet the highest comprehension level.

Table 8 presents students' individual levels of comprehension based on the data obtained for facets needed to transition among the proof comprehension levels. At the Surface level, the percentage of students with high comprehension is very low (10%), with students having difficulties with basic concepts needed for conducting a proof.

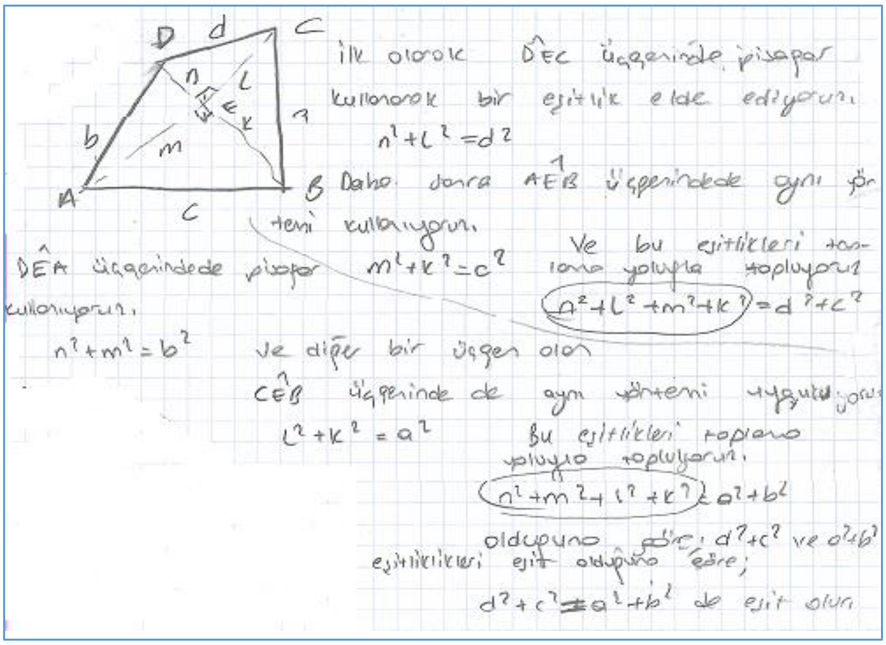
The percentages of high (50%) and medium (40%) level comprehension in Recognizing Elements where there is a logical relationship is more than the percentages of low comprehension (10%). This result reveals the parts of the proofs comprehended.

In PCT-3, the high comprehension degree (50%) is higher than both PCT-1 (25%) and PCT-2 (20%). Different to PCT1 and PCT-2, the highest comprehension degree for PCT-3 occurs in the Recognizing Elements level; however, the comprehension degree was still inadequate at 50%. The reason may be the effect of high percentages of low degree comprehension at the Surface level. It may be considered that knowing the basic concepts is necessary for the advanced comprehension levels, but is not enough.

Table 8. PCT-3 participant evaluation results

Level	Comprehending Degree	Student	Frequency	Percentage
Surface	Low	S1,S2,S3,S4,S8,S9,S10,S11,S12,S13, S17,S18,S19,S20	14	70
	Medium	S5,S6,S15,S16	4	20
	High	S7,S14	2	10
Recognizing Elements	Low	S2,S19	2	10
	Medium	S3,S8,S9,S13,S14,S15,S18,S20	8	40
	High	S1,S4,S5,S6,S7,S10,S11,S12,S16,S17	10	50
Chaining Elements	Low	S1,S2,S3,S9,S10,S11,S13,S15,S16, S17,S18,S19,S20	13	65
	Medium	S4,S5,S6,S7,S12,S14	6	30
	High	S8	1	5
Encapsulation	Low	S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11, S12,S13,S14,S15,S16,S17,S18,S19,S20	20	100
	Medium	-	0	0
	High	-	0	0

In the Chaining Elements level, only one student shows a high comprehension level, whereas six students presented medium level comprehension. In S8's high performance for Question 9 ("Explain the conducted proof in your own words"), S8 explained the conducted proof in his/her own words, but stayed with the given proof; hence, only 1 point was scored (see Figure 6) (student's writings: first of all by using Pythagorean [formula] on triangle DEC we get $n^2+l^2=d^2$ Later I use same presentation in triangle AEB $m^2+k^2=c^2$. These equalities are added and $n^2+l^2+m^2+k^2=d^2+c^2$. Pythagorean [formula] is used in DEA triangle n^2+m^2 $=b^2$ and the same method is applied on the other triangle CEB $l^2+k^2=a^2$. These equalities are added and $n^2+m^2+\ l^2+k^2=a^2+b^2$. Therefore d^2+c^2 and a^2+b^2 are same and $d^2+c^2=a^2+b^2$).


D
d
C
n
l
E
k
a
b
m
A
c
B

ilk olarak $\hat{DEC}$ üçgeninde pisagor kullanarak bir eşitlik elde ediyorum.

$n^2+l^2=d^2$

Daha sonra $\hat{AEB}$ üçgenindede aynı yöntemi kullanıyorum. $m^2+k^2=c^2$

Ve bu eşitlikleri toplama yoluyla topluyorum

$n^2+l^2+m^2+k^2=d^2+c^2$

$\hat{DEA}$ üçgenindede pisagor kullanıyorum.

$n^2+m^2=b^2$

Ve diğer bir üçgen olan $\hat{CEB}$ üçgeninde de aynı yöntemi uyguluyorum

$l^2+k^2=a^2$

Bu eşitlikleri toplama yoluyla topluyorum.

$n^2+m^2+l^2+k^2=a^2+b^2$

olduğuna göre; d^2+c^2 ve a^2+b^2 eşitlikleri eşit olduğuna göre;

$d^2+c^2=a^2+b^2$ de eşit olur

Figure 6. Answer by S8 for PCT-3 Question 9

In the Encapsulation level, regarding proving differently, no student achieves medium or high level comprehension, only low.

3.4. Analysis Results: PCT-4

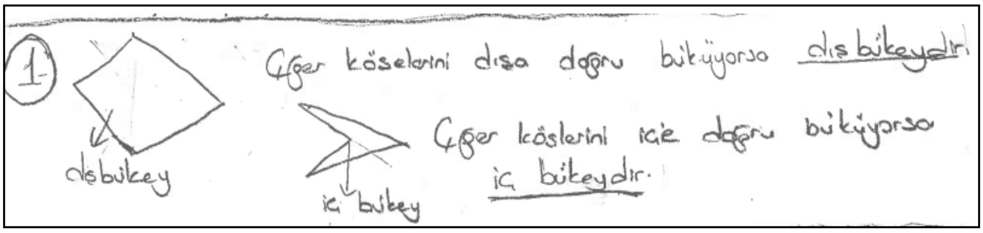
PCT-4 was prepared for the explanation of 'process standard 3' of the 11th grade geometry curriculum (MEB, 2010); that is, "Prove that the area of convex quadrilateral region is equal to half of the multiplication of length of the diagonal by sinus of angle between diagonals". There are 11 questions in PCT-4, evaluated based on the determined facets, with scores presented in Table 9.

Table 9. PCT-4 scores & total percentages by facet

PCT-4	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S 20	Total %
T4(1) F1	0	0	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	0	35
T4(2) F1	1	1	1	0	1	1	0	1	1	1	0	0	0	0	1	1	0	0	0	1	22
T4(3) F2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	100
T4(4) F2	2	2	2	2	2	2	2	1	2	1	0	2	1	1	1	0	1	1	1	1	67
T4(5) F2	2	2	2	2	2	2	2	1	2	2	0	2	1	1	1	0	1	1	1	1	67
T4(6) F2	2	2	2	2	2	2	2	1	2	2	1	2	1	2	1	2	1	2	1	2	85
T4(7) F2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	37
T4(8) F3	1	0	1	1	1	2	0	1	1	1	1	0	1	1	1	1	0	1	0	1	35
T4(9) F4	0	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	33
T4(10) F4	1	1	2	1	1	2	0	2	0	1	1	2	0	0	1	1	1	1	0	2	50
T4(11) F5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tn(n): Theorem (question), F: Facet, S: Student, F1: Basic Knowledge, F2: Logical Status, F3: Summarization, F4: Generality, F5: Application

Questions 1-3 of PCT-4 are on Basic Knowledge, involving statements of proofs and knowledge of symbols such as knowledge of convex, and sinus. However, the percentage of comprehension level is very low for Question 1 (35%) and Question 2 (22%), but for Question 3, all students reached high performance percentages. Question 2 is on sinus which is about trigonometry, and Question 1 asks the definition of "convex". Figure 7 shows student S19 supporting his/her definition with two shapes, identifying one as convex and one as concave. Due to this poor definition, S19 had Question 1 graded at 1 point (student's writing: If the vertices are bending towards outward it is [polygon] convex. If vertices are bending inward it is [polygon] concave.)


1
dışbükey
iç bükey

Eğer köselerini dışa doğru büküyorsa dış bükeydir.

Eğer köşlerini içe doğru büküyorsa iç bükeydir.

Figure 7. Answer by S19 for PCT-4 Question 1

In the Logical Status (F2) facet, where transitions among proof steps are questioned, although adequate comprehension degree is obtained for Questions 4-5 (67%) and Question 6 (85%), a low degree (37%) of comprehension performance is observed for Question 7 (conducting a proof in different ways). This result shows consistency with other PCTs. For Question 7 ("Can you produce area formula for the quadrilateral by using AEB angle?"), one student answered:

> *Yes. |BD| sides is 180 degrees and let AEB angle be x, and also DEA angle y then x+y=180 degrees. From this point we can use this method for the other sides, and combine all of them to find the area formula for the quadrilateral, which is ½. AE. BE. Sin (180-alpha). (S7)*

Another result observed is the relationship between Questions 2 and 7. Both required trigonometry knowledge and students show low comprehension level on both (Q2: 22%, Q7: 37%), suggesting students have poor background trigonometry knowledge (taught in 10th grade). Medium comprehension levels are obtained for the Summarization facet (F3) (64%), in which comprehending critical ideas are explained, and from Generalization facet (F4) in which the certainty of the proof is questioned (approximately 43% where (33+50)/2). Again, similar to PCT-2 and PCT-3, low comprehension performance is seen in the Application facet (F5), in which explained proofs can be applied in varied situations, with no students performing adequately.

As in PCT-3, although in some questions of PCT-4 students present medium or low comprehension, in total the highest comprehension is for the Logical Status facet. Based on the data gathered, each student's level of comprehension is presented in Table 10. At the Surface level, students mostly present medium comprehension level, with no students at low level. For the Recognizing Elements level, half of the students presents high comprehension level. It can therefore be concluded that students can recognize which properties should be applied to the proof or are able to identify logical order of statements.

Table 10. PCT-4 participant evaluation results

Level	Comprehending Degree	Student	Frequency	Percentage
Surface	Low	-	0	0
	Medium	S1,S3,S4,S5,S6,S7,S8,S11,S12,S13, S14,S15,S16,S17,S18,S19,S20	17	85
	High	S2,S9,S10	3	15
Recognizing Elements	Low	S11,S16	2	10
	Medium	S8,S13,S14,S15,S17,S18,S19,S20	8	40
	High	S1,S2,S3,S4,S5,S6,S7,S9,S10,S12	10	50
Chaining Elements	Low	S1,S2,S3,S4,S7,S9,S10,S11,S12, S14,S15,S17,S18,S19,S20	15	75
	Medium	S5,S8,S13,S16	4	20
	High	S6	1	5
Encapsulation	Low	S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11, S12,S13,S14,S15,S16,S17,S18,S19,S20	20	100
	Medium	-	0	0
	High	-	0	0

According to Table 10, among the high comprehension levels, Recognizing Elements, which explains logical relationships of transition among proof steps, is the highest. As with most other PCTs (excluding PCT-1), no students achieves satisfactory comprehension at the Encapsulation level, with all achieving a low degree of comprehension. Although the results of PCT-4 shows no descending or ascending pattern, this is very normal when considering the multilayered structure of proof comprehension.

3.5. Analysis Results: PCT-5

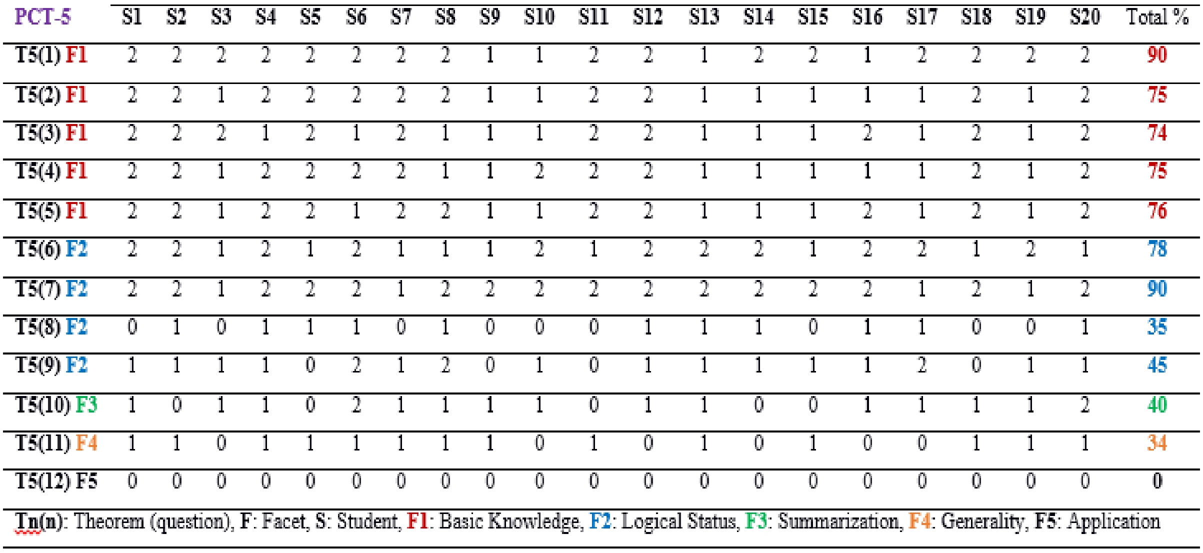
PCT-5 was prepared for the theorem “Prove that the area of quadrilateral is equal to half of the area of a quadrilateral whose midpoints of the edges are accepted as vertices”, which is given as an explanation of ‘process standard 3’ of the 11th grade geometry curriculum (MoNE, 2010). Twelve questions were asked based on the theorem in PCT-5, with answers evaluated according to predetermined facets, and scores presented in Table 11.

Questions 1-5 of PCT-5 are about Basic Knowledge (F1) and the knowledge of statements and symbols of proofs, with a high level (84%) of comprehension performance seen. Question 1 sees the highest level (90%) in which students were asked the definition of “intermediate base”.

One student answered fully (see Figure 8) with:

> *The line drawn from the midpoint of a side of the triangle through to other side that is parallel to the base is called the intermediate base. It is parallel to the base of the triangle so that if the intermediate base is ‘a’ then the base of the triangle is ‘2a’. (S11).*

Table 11. PCT-5 scores & total percentages by facet

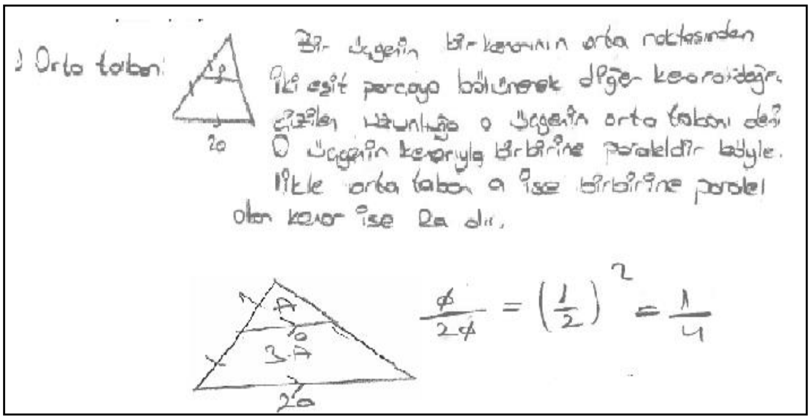


PCT-5	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Total %
T5(1) F1	2	2	2	2	2	2	2	2	1	1	2	2	1	2	2	1	2	2	2	2	90
T5(2) F1	2	2	1	2	2	2	2	2	1	1	2	2	1	1	1	1	1	2	1	2	75
T5(3) F1	2	2	2	1	2	1	2	1	1	1	2	2	1	1	1	2	1	2	1	2	74
T5(4) F1	2	2	1	2	2	2	2	1	1	2	2	2	1	1	1	1	1	2	1	2	75
T5(5) F1	2	2	1	2	2	1	2	2	1	1	2	2	1	1	1	2	1	2	1	2	76
T5(6) F2	2	2	1	2	1	2	1	1	1	2	1	2	2	2	1	2	2	1	2	1	78
T5(7) F2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	2	1	2	1	2	90
T5(8) F2	0	1	0	1	1	1	0	1	0	0	0	1	1	1	0	1	1	0	0	1	35
T5(9) F2	1	1	1	1	0	2	1	2	0	1	0	1	1	1	1	1	2	0	1	1	45
T5(10) F3	1	0	1	1	0	2	1	1	1	1	0	1	1	0	0	1	1	1	1	2	40
T5(11) F4	1	1	0	1	1	1	1	1	1	0	1	0	1	0	1	0	0	1	1	1	34
T5(12) F5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tn(n): Theorem (question), F: Facet, S: Student, F1: Basic Knowledge, F2: Logical Status, F3: Summarization, F4: Generality, F5: Application

Accordingly, it can be concluded that students comprehend intermediate base from the 10th grade. In the Basic Knowledge facet, Questions 2-5 are all similar and results show students’ comprehension levels as close to each other and therefore consistent.

In Questions 6-9 of the Logical Status facet (F2), regarding transition among proof steps, a medium level (62% where (78+90+35+45)/4) of comprehension performance is seen, but upon a question-based examination, various percentages of comprehension performances occurred. For instance, in Question 7 regarding common parenthesis, the highest degree (90%) of comprehension is seen, but in Question 9 regarding premises, only a medium level (45%) of comprehension is obtained.



1) Orta taban: Bir üçgenin bir kenarının orta noktasından iki eşit parçaya bölünerek diğer kenarlardağı çizilen uzunluğa o üçgenin orta tabanı denir. O üçgenin kenarıyla birbirine paraleldir böyle. İlkle orta taban a ise birbirine paralel olan kenar ise 2a dır.

$$\frac{A}{2A} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

x
a
2a
x
x
A
a
3A
x
2a

Figure 8. Answer by S11 PCT-5 Question 1

Accordingly, it can be deduced that students comprehended combining the giving parts of a proof; however, they could not decide which parts are necessary for proving. As an example, a student answered Question 9 regarding identifying which premises are necessary for the conducted proof, as "The equality of the sides should be given" (S9).

As with the other interventions, in PCT-5 no student conducted a proof in a different way, resulting in 0% for the degree of comprehension of the Application facet. Students' comprehension percentages are presented in Table 12 based on the scores in each facet.

Table 12. PCT-5 participant evaluation results

Level	Comprehending Degree	Student	Frequency	Percentage
Surface	Low	-	0	0
	Medium	S3,S9,S10,S11,S13,S14,S15,S16,S17,S19	10	50
	High	S1,S2,S4,S5,S6,S7,S8,S12,S18,S20	10	50
Recognizing Elements	Low	-	0	0
	Medium	S1,S3,S7,S9,S10,S11,S15,S18,S19	9	45
	High	S1,S2,S3,S4,S7,S9,S10,S11,S12,S14,S15,S17,S18,S19,S20	15	75
Chaining Elements	Low	S2,S9,S19,S20	4	20
	Medium	S3,S10,S11,S12,S13,S14,S15,S16,S17	9	45
	High	S1,S4,S5,S6,S7,S8,S20	7	35
Encapsulation	Low	S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S12,S13,S14,S15,S16,S17,S18,S19,S20	20	100
	Medium	-	0	0
	High	-	0	0

When the comprehension levels are considered for PCT-5, students present mostly high levels of comprehension for this proof. For example, in the Surface level, 50% of the students obtain high level comprehension; and similarly for Recognition Elements it is 75%. From this result it can be deduced that most students comprehend the basic concept of this proof and also identify the parts and premises of the proof. Although the Surface and Recognizing Elements levels see a high degree of comprehension increase, in both Chaining Elements and Encapsulation, the high degree of comprehension decreased very fast.

In the Chaining Elements level of PCT-5, 35% of students achieve high level comprehension, which is the best among all the other PCTs. The reason for this result may relate to high comprehension levels observed in Basic Knowledge and Recognizing Elements levels, again where the other PCTs are not. Moreover, when this result is considered against other results obtained from the PCTs and students show high performance on the first two comprehension levels, they also present low performance for the second two; however, in reverse, it is not always true. In summary, Basic Knowledge and Recognizing Elements are necessary, but not sufficient in every case.

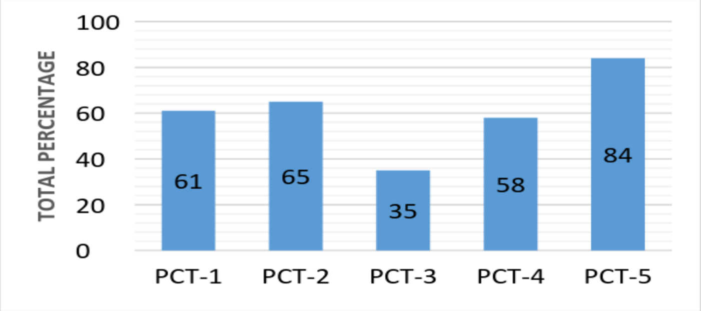
The other result is that students performed poor at comprehension levels when more than one knowledge area is questioned, and the comprehension performance increased for comprehension facets depending on only one area of knowledge. This conclusion matches results obtained from the level of combining the parts.

3.6. Results of Comprehension Level Analyses

In this section, the four levels of proof comprehension are individually investigated to identify any changes or improvements obtainable from the teaching sessions. In the analysis, the students' total grades for each question are noted, and the assessing comprehension level and percentages recorded.

3.6.1. *Analysis Results: Basic Knowledge Level*

Results for the Basic Knowledge level in which the basic terms, statements and symbols in a proof are examined and their results are presented in Graph 1.



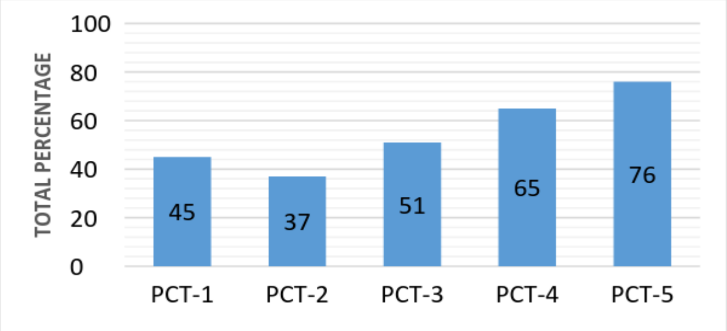
Graph 1. Comparison of Basic Knowledge level comprehension

The lowest comprehension is obtained in PCT-3 and the highest in PCT-5. Comprehension degrees do not present a general increasing or decreasing pattern. This situation relates to previous knowledge about the concepts of the theorems selected for the test. As can be seen from the students' answers, previously acquired knowledge varies even in the same proof.

3.6.2. *Analysis Results: Recognizing Elements Level*

In this level, Logical Relationships of the transitions of proof steps are examined and students' comprehension in test grades gathered and percentages calculated and then compared. Results are shown in Graph 2.

The lowest percentages of comprehension were obtained in PCT-2 and the highest in PCT-5. After the first two PCTs, an increasing pattern is seen. In this Recognizing Elements level, which consists of explaining given proof steps, students seem to comprehend the structure of the logical relationships among the proof steps. However, it may be that these results are because asking students to interpret given proof steps is easier than asking them to construct steps from the beginning.

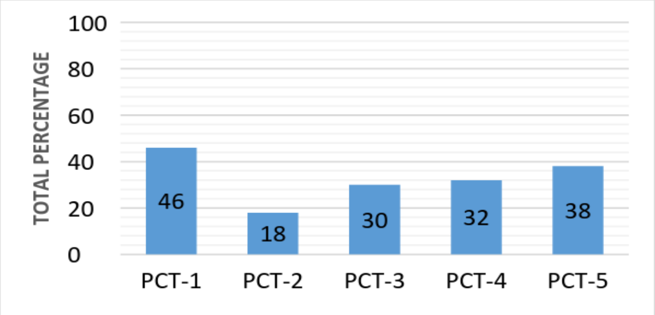


Graph 2. Comparison of Recognizing Elements level comprehension

After PCT-2, an increasing pattern is observed; concluding that recognizing elements of a proof is comprehended. Alternatively, it may be about understanding the PCT's structure, since it is a novel intervention for the students.

3.6.3. *Analysis Results: Chaining Elements*

In comparing the comprehension level of Chaining Elements, which involves combining logical arguments in a proof and defining validation, the combination of element levels for each tests' comprehension percentages are presented in Graph 3.

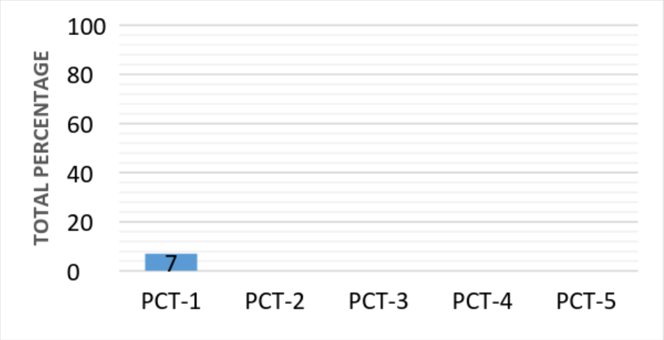


Graph 3. Comparison of Chaining Elements level comprehension

The lowest percentages of comprehension were obtained in PCT-2, and the highest in PCT-1. Generally, very low comprehension was observed in this level. The result for PCT-1 is significant as the scope of knowledge and process of theorem is known by the students. However, although students understand the transitions among proof steps, they are unable to produce the next step by themselves, struggling to determine the critical step the proof is based upon and explaining this step. They also experienced difficulties explaining what the proof is verifying. Except for PCT-1, although comprehension percentages are very low, they still present increasing percentages of comprehension.

3.6.3. *Analysis Results: Encapsulation Level*

Results based on percentages of comprehension at the Encapsulation level, regarding deciding how to conduct a proof in another situation and internalizing propositions of a proof, are presented in Graph 4.

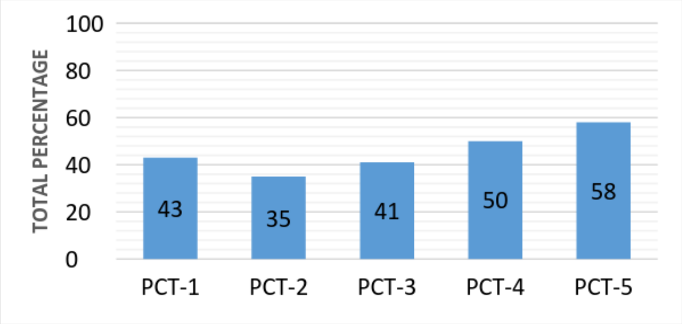


Graph 4. Comparison of Encapsulation level comprehension

Only in PCT-1 students present a degree of comprehension, although actually very low (7%). In PCT-1, the students are given a proof about the sum of interior angles and asked to prove the sum of exterior angles. During the interview conducted with students about the given proof, it was revealed that the strategy used was not their own, but inspired from another proof related to finding the sum of exterior angles. In the other PCTs, no students performed adequate degrees of comprehension for this level.

3.6.4. *Overall evaluation of PCTs*

During analysis conducted across all PCTs, the total grades of each PCT are initially calculated. Then the percentages are calculated by considering the overall grades of the tests. The analysis results are presented in Graph 5.



Graph 5. Overall evaluation of PCTs

The lowest comprehension level is for PCT-2 (35%) which was conducted in the second week of the intervention process, whereas the highest is for PCT-5 (58%) during the final week. Therefore, in general, the comprehension level tends to increase; however, this increment is nonlinear. It is suggested that this is related not only to the PCTs, but to the process of proving.

4. DISCUSSION AND CONCLUSION

The aim of this current study was to present 11th grade high school students' proof comprehension performance according to five prepared PCTs related with the quadrilateral unit of the geometry curriculum. Student performance was analyzed according to Yang and Lin's (2008) multilayered model, whereby a proof is comprehended with five facets among four levels.

The results obtained are discussed in this section within the multilayered model progressions. The model's first step is the Surface Level, where students acquire basic knowledge regarding the meaning of statements and symbols in a proof under the Basic Knowledge facet. Students mostly performed to a medium degree, except for PCT-5. Although the proof in PCT-5 is related to quadrilaterals, for the Basic Knowledge facet questions are all about areas of triangles. Since students' background knowledge about triangles may be better than for quadrilaterals, comprehension for PCT-5 was higher than all other PCTs for the Basic Knowledge facet. Half or more of the students performed high at the Surface level. Only for PCT-3 and PCT-4 did more than half of the students performed low or medium, showing that most comprehend the basic components of proofs, can identify preliminary knowledge and each element of a proof.

According to the literature, the most common proof comprehension problem is "not knowing the definition used in the proof" (Moore, 1994, p. 251) and deficiencies about mathematical definitions, roles, the importance of definition on mathematics, and how those definitions can be used while proving (Atwood, 2001; Edwards & Ward, 2004; Knapp, 2006).

Aligned with the results of the study of Conradie and Frith (2000), in the current study students often fail to understand the meaning of key terms when reading a proof, hindering their ability to comprehend other aspects of a proof, and that less successful students may not try to understand the meaning of key terms and statements (Weber, Brophy, & Lin, 2008).

The second level, Recognizing Elements, contains the Logical Status and Summarization facets, and is where students identify the logical status of statements used either explicitly or implicitly in a proof. Logical Status is explained as "recognizing a condition applied directly, judging the logical order of statements and recognizing which properties are applied" (Lin & Yang, 2007, p.351). In this facet students are expected to identify premises and select logically. Students mostly performed moderately, except for PCT-2 and PCT-5 where they performed outstandingly, meaning they comprehended the sequence of given arguments in a proof. In PCT-5, students must identify critical steps by questioning equalities in the given proof and also some premises. There may be two reasons for outstanding comprehension for this facet of PCT-5. Firstly, since PCT-5 is the final test, students may better understand the structure of these proof questions; and secondly, although the context of the proof is quadrilaterals, PCT-5's questions are also answerable by considering the properties of triangles, which may be better known by students. Moreover, for the Summarization facet, which is defined by Lin and Yang as "identifying critical procedures, premises or conclusions and identifying critical ideas of a proof" (2007, p.751), students presented medium performance for all PCTs, except for PCT-3 which directly asked the critical steps for the given proof. Students did not perform well since they have difficulties decontextualizing the given proof and identifying the necessary steps. In this facet, besides PCT-3, students performed moderate comprehension, but percentages were borderline to low comprehension, suggesting deficiencies with identifying critical procedures of a given proof. This result parallels "deficiencies on knowledge of context and strategies" as reported by Knapp (2006). For the Recognizing Elements level, all PCTs except for PCT-1 and PCT-2 showed high performance for 50% of students. On this topic, Mejia-Ramos et al. (2012) stated that it needs to not only identify the logical status of statements in proofs but also recognize the logical relationship between the statement being proven the assumptions and conclusions of the proof".

The third level of the proof comprehension model is Chaining Elements, in which students comprehend the way in which different statements whose logical status are identified in the previous level, are connected in the proof by identifying their logical relations. This level contains two facets, Generality and Application. Generality facet was introduced by Lin and Yang (2007) as "justifying correctness and identifying what is validated by the proof" (p.751), and students performed moderately for most PCTs, except for PCT2 and PCT3. In these two PCTs, students performed very low when asked to confirm why the given proof is correct. According to Schoenfeld (1994), students mostly focusing on visualization may come away with the misconception that "seeing is believing". For the Application facet, Lin and Yang's (2007) final stage in which creation of new knowledge is sought, except for three students who confessed they had seen the proof before, no others presented a satisfactory performance. Similar to this finding, Heinze, Cheng, and Yang (2004) identified that students performed well in conducting proof if given familiar proof settings. Moreover, since students are introduced to proofs in secondary school they are more challenged in this comprehension level than the others. This finding aligns to results of a study by Hemmi (2008) who stated that having less experiences on proof (in the process of comprehending meaning of proof or learning to construct own proof) then proof is invisible for them based on the condition of transparency.

The final level for the model is Encapsulation; understanding whether students conduct interiorization of the proof as a whole. No students satisfactorily achieved this level, paralleling results of Yang and Lin (2008). In their study, Yang and Lin (2008) stated that Encapsulation, the fourth level of proof comprehension's theoretical framework, is not aimed at secondary school, claiming this comprehension level would occur in advanced mathematics education.

In general, comprehension levels tended to increase through the PCTs since the lowest comprehension occurred in PCT-2 during the second week of application, and the highest occurred in PCT-5 during the final last week. However, this increment is nonlinear. The current study's researchers suggest this is related not only to the PCTs, but also the process of proving. Since proofs require different mental procedures (Ball et al., 2002), obtaining different comprehension levels for each proof is a natural result.

5. IMPLICATIONS

According to the current study's results, it was observed that no students could achieve the Encapsulation level; conducting a proof in various ways or proving different theorems by using the same proof methods. Therefore, no participants reached the NCTM (2000) standards of proof and proving which are to "develop and evaluate mathematical arguments and proofs" and "select and use various types of reasoning and methods of proof" (p.342). Yang and Lin (2008) also obtained similar results and defined Encapsulation level as the "global level". In particular, they indicated that their instrument was not aimed at diagnosing if a student had reached this last level (p.71). This result led them to suggest the wording of the standards written in the mathematics curriculum. In the standards wording, students are asked to "do proof"; however, if they are written as "understand proof" or "interpret proof" it may be easier to reach the aims of teaching and learning about proof, after which, conducting a proof may be constructed on such understanding.

Another result obtained from the current study was no linear increase or decrease of comprehension level and achievement on facets. This reinforces that proof comprehension is a multilayered action and complex concept (Selden & Selden, 1995) involving various mental processes (Ball et al., 2002; Tall, 1992), and that PCTs can be used as a tool for contributing to this complex understanding.

Among the five PCTs, the highest comprehension was obtained from the level containing knowledge about definition, properties, and the meanings of symbols. Achievement and comprehension decreased with identifying the components of a proof needing higher mathematical skills. Özer & Arıkan (2002) obtained similar results with almost no students reaching the necessary level of conducting a proof among 10th graders. However, the current study saw a slight improvement among the PCTs which can be interpreted as affecting students' comprehension of proofs. In summary, although students have low level comprehension from proof comprehension tests, the PCTs positively affected their comprehension therefore PCTs could be used both for teaching proof and evaluating proof comprehension.

Acknowledgements

This study is derived from a part of master thesis of the first author.

ORCID

Bahattin İnam https://orcid.org/0000-0002-6212-8013
Işıkhan Uğurel https://orcid.org/0000-0003-4067-1522
Burçak Boz Yaman https://orcid.org/0000-0002-0922-3652

6. REFERENCES

Alcock, L., & Wilkinson, N. (2011). e-Proofs: Design of a resource to support proof comprehension in mathematics. *Educational Designer*, *1*(4). Retrieved from http://www.educationaldesigner.org/ed/volume1/issue4/article14/index.htm?&popups=off

Almeida, D. (2003). Engendering Proof Attitudes: Can The Genesis of Mathematical Knowledge Teach Us Anything?. *International Journal of Mathematical Education in Science and Technology*, *34*(4), 479-488.

Atwood, P.R. (2001). *Learning to Construct Proofs in a First Course on Mathematical Proof* (Doctoral dissertation). Retrieved from https://scholarworks.wmich.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=2352&context=dissertations

Ball, D. L., Hoyles, C., Jahnke, H. N., & Movshovitz-Hadar, N. (2002). The Teaching of Proof. *ICM*, *3*, 907-920.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel Araştırma Yöntemleri*. (16th ed.). Ankara: Pegem Akademi.

Conradie, J., & Frith, J. (2000). Comprehension Tests in Mathematics. *Educational Studies in Mathematics*, *42*(3), 225-235.

Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 2nd ed. Sage, Thousand Oaks.

Di Martino, P., & Maracci, M. (2009). *The Secondary-Tertiary Transition: Beyond the Purely Cognitive*. In M. Tzekaki, M. Kaldrimidou, & H. Sakonidis (Eds.), *Proceedings of 33rd Conference of the International Group for the Psychology of Mathematics Education* (pp.401-408). Thessaloniki, Greece.

Duval, R. (2002). Proof understanding in mathematics: What ways for students. In F. L. Lin (Ed.), *International conference on mathematics – "Understanding proving and proving to understand"* (pp.61-77). Taipei: National Science Council and National Taiwan Normal University.

Edwards, B.S., & Ward, M.B. (2004). Surprises from mathematics education research: Student (mis)use of mathematical definitions. *American Mathematical Monthly*, *111*, 411-424

Heinze, A., Cheng, Y. H., & Yang, K. L. (2004). Students' performance in reasoning and proof in Taiwan and Germany: Results, paradoxes and open questions. *ZDM*, *36*(5), 162-171. doi:10.1007/BF02655668

Hemmi, K. (2008). Students' encounter with proof: the condition of transparency. *ZDM, The Special Issue on Proof*, *40*(3), 413-426.

Houston, S. K. (1993a). Comprehension Tests in Mathematics. *Teaching Mathematics and its Applications*, *12*(2), 60-73.

Houston, S. K. (1993b). Comprehension Tests in Mathematics: II. *Teaching Mathematics and its Applications*, *12*(2), 113-120.

İnam, B., & Uğurel, I. (2016). İspat kavrama testine dayalı bir öğretim uygulamasında karşılaşılan güçlükler ve sürece müdahale yolları. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *12*(1), 1-21.

Knapp, J. L. (2006). *Students' appropriation of proving practices in advanced calculus* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (Accession Order No. AAT 3241300)

Koshy, V. (2005). *Action Research for Improving Practice*. London: Paul Chapman Publishing.

Leron, U. (1983). Structuring Mathematical Proof. *American Mathematical Monthly*, *90*(3), 174-185.

Lin, F. L., & Yang, K. L. (2007). The Reading Comprehension of Geometric Proofs: The Contribution of Knowledge and Reasoning. *International Journal of Science and Mathematics Education*, *5*(4), 729-754.

MEB [Milli Eğitim Bakanlığı] (2010). *11.Sınıf Geometri Öğretim Programı*. Ankara: MEB. Retrieved from http://ogm.meb.gov.tr/

MEB [Milli Eğitim Bakanlığı] (2013). *Ortaöğretim Matematik Dersi (9, 10, 11 ve 12. Sınıflar) Öğretim Programı*. Ankara: MEB. Retrieved from http://ogm.meb.gov.tr/

Mejia-Ramos, J. P. (2008). *The construction and evaluation of arguments in undergraduate mathematics: A theoretical and a longitudinal multiple-case study*. (Doctoral dissertation). University of Warwick, U.K.

Mejia-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An Assessment Model for Proof Comprehension in Undergraduate Mathematics. *Educational Studies in Mathematics*, *79*(1), 3-18.

Moore, R. C. (1994) Making the Transition to Formal Proof. *Educational Studies in Mathematics*, 27(3), 249-266.

NCTM [National Council of Teachers of Mathematics]. (2000). *NCTM Principles and Standards for School Mathematics*. Reston. VA: NCTM.

Özer, Ö., & Arıkan, A. (2002). Lise matematik derslerinde öğrencilerin ispat yapabilme düzeyleri. In *Proceedings of V. National Science and Mathematics Education Congress*, ODTÜ, Ankara. Retrieved from http://old.fedu.metu.edu.tr/ufbmek-5/b_kitabi/PDF/Matematik/Bildiri/t245d.pdf

Remillard, K. S. (2010). Exploring the learning of mathematical proof by undergraduate mathematics majors through discourse analysis. In *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*. Raleigh, NC: RUME (published online). Retrieved from http://sigmaa.maa.org/rume/crume2010/Archive/Remillard.pdf

Roy, S., Alcock, L., & Inglis, M. (2010). Supporting Proof Comprehension: A Comparative Study of Three Forms of Presentation. In *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*. Raleigh, NC: RUME (published online). Retrieved from http://sigmaa.maa.org/rume/crume2010/Archive/Roy%20et%20al.pdf

Sarı, M. (2011). *Üniversite Öğrencilerinin Matematiksel Kanıt ile İlgili Güçlükleri ve Kanıt Öğretimi*. (Doctoral dissertation). Retrieved from National Thesis Center. (Accession Order No. 299545)

Schoenfeld, A. (1994). Reflections on doing and teaching Mathematics. In A. Schoenfeld (Ed.). *Mathematical Thinking and Problem Solving* (pp.53-69). Hillsdale, NJ: Lawrence Erlbaum Associates.

Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, *29*(2), 123-151.

Stylianides, A.J. (2007). Proof and proving in school mathematics. *Journal of Research in Mathematics Education*, *38*(3), 289-321.

Tall, D. (1992). The Transition to Advanced Mathematical Thinking: Functions, Limits, Infinity and Proof. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp.495-511). Reston, VA: National Council of Teachers of Mathematics/Macmillan.

Uğurel, I., & Moralı, S. (2010). Bir Ortaöğretim Matematik Dersindeki İspat Yapma Etkinliğine Yönelik Sınıf içi Tartışma Sürecine Öğrenci Söylemleri Çerçevesinde Yakından Bakış. *Buca Eğitim Fakültesi Dergisi*, *28*, 135-154.

Weber, K., & Mejia-Ramos, J.P. (2011). Why and how mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics*, *76*(3), 329-344.

Weber, K., Brophy, A., & Lin, K. (2008). Learning about advanced mathematical concepts by reading text. In *Proceedings of the 11th Conference on Research in Undergraduate Mathematics Education*. San Diego, California: RUME (published online). Retrieved from http://sigmaa.maa.org/rume/crume2008/Proceedings/Weber%20LONG.pdf

Yang, K. L. (2012). Structures of Cognitive and Metacognitive Reading Strategy Use for Reading Comprehension of Geometry Proof. *Educational Studies in Mathematics*, *80*(3), 307-326.

Yang, K. L., & Lin, F. L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*(1), 59-76.

Yang, K. L., Lin, F. L., & Wang, Y.T. (2008). The effects of proof features and question probing on understanding geometry proof. *Contemporary Educational Research Quarterly*, *16*(2), 77-100.

Yıldırım, A., & Şimşek, H. (2013). *Sosyal Bilimlerde Nitel Araştırma Yöntemleri* (9th ed.). Ankara: Seçkin Yayıncılık.

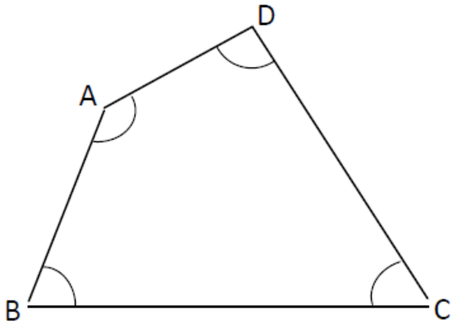
Yıldız, G. (2006). *Lisans Seviyesinde Genel Matematik Dersindeki Teorem ve İspatları Anlamaya Yönelik Kavrama Testinin Hazırlanması Uygulanması ve Öğrenci Görüşlerinin Değerlendirmesi*. (Master's thesis) Retrieved from National Thesis Center. (Accession Order No. 187622)

Zazkis, R., & Zazkis, D. (2014). Proof scripts as a lens for exploring proof comprehension. In T. Fukawa-Connolly, G. Karakok, K. Keene, & M. Zandieh (Eds.), *Proceedings of the 17th Annual Conference for Research in Undergraduate Mathematics Education* (pp.1198-1204), Denver, CO.

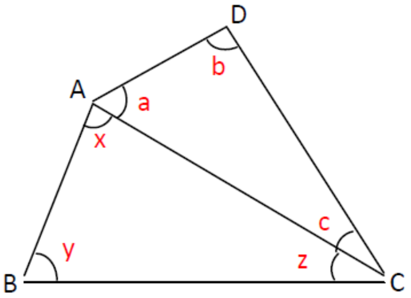
Appendix: PCT-1

Theorem:

Summation of interior angles of a quadrilateral is 360°



Proof:



In the quadrilateral ABCD, draw diagonal [AC] and construct ABC and ACD triangles.
(Step-1)

In figure x, y, z, a, b and c are representing measurement of the related angles.
(Step-2)

In ABC triangle x + y + z = 180°
(Step-3)

In ACD triangle a + b + c = 180° dir.
(Step-4)

Use the given theorem and proof to answer the following questions.

1. Define the terms “quadrilateral”, “triangle”, and “diagonal” used in the proof.
2. Is the equality given in Step 3 true? Explain.
3. Explain how the equality given in Step 5 is obtained?
4. In the proof, if Step 1 and Step 2 change order, is the proof still true?
5. In Step 1, if [BD] is drawn instead of [AC], is the proof still true?
6. According to you, what are the critical step(s) for which proof is based? Explain.
7. When the whole proof is considered, do you think this proof is true? Explain.
8. State the conducted proof with your own words.
9. Can you prove this theorem in a different way?
10. Try to obtain summation of the exterior angles of a quadrilateral by considering the given theorem and proof.

International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 370–388

DOI: 10.21449/ijate.423602

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

Efficiency Measurement With A Three-Stage Hybrid Method

İrfan Ertuğrul **Tayfun Öztaş**[1*]

[1] Pamukkale University, Department of Business Administration, Denizli, Turkey

Abstract: Data Envelopment Analysis (DEA) is one of the most widely used efficiency measurement techniques in the literature. In the method developed by Charnes, Cooper, and Rhodes, the relation between input(s) and output(s) is examined and relative efficiency values are obtained for many decision-making units. In order to be able to accurately measure the efficiency with Data Envelopment Analysis, the selection of input and output variables needs to be done carefully otherwise, the results may be misleading. For this purpose, it is aimed to make an objective selection process by using Grey Relational Analysis (GRA) in the identification of variables in the study. Via this method 17 financial ratios of 20 firms in the BIST Food Index for the period of 2013-2015 categorized into 4 groups, then each category clustered and the ratios which have the highest correlation within each cluster selected as representative indicator. Thus, 3 inputs and 2 output variables were selected so that the number of variables was reduced from 17 to 5. An input-oriented BCC model was established with selected variables to determine the efficiencies of firms in each period. The Malmquist Total Factor Productivity Index was used to analyze the productivity changes between periods. It was concluded that 7 firms were efficient in each year and the productivity of the sector increased between the periods as a result of the analysis.

ARTICLE HISTORY

Received: 10 February 2018
Revised: 24 March 2018
Accepted: 30 April 2018

KEYWORDS

Efficiency measurement, Grey Relational Analysis, Data Envelopment Analysis, Malmquist index, Clustering

1. INTRODUCTION

Efficiency is doing an activity with possibly the shortest time and the lowest cost, taking into consideration the quality (Chorafas, 2015). According to another approach, efficiency is the comparison of the optimal values and the observed values of inputs and outputs. In this approach, optimality is expressed in terms of production possibilities or the behavioral goals of the manufacturer (Fried, Lovell, & Schmidt, 2008). Effectiveness is reaching a goal under various constraints arising from planning including financial plans, timelines and human resources (Chorafas, 2015). If the two definitions are summed up to include both similarities and differences, efficiency is doing things right and effectiveness is doing the right things (Sheth & Sisodia, 2002). Productivity is simply the ratio of output to input. The productivity measure, which includes all factors, is called total factor productivity, while the efficiency of certain features is called partial productivity (Coelli, Rao, O’Donnell, & Battese, 2005).

CONTACT: Tayfun Öztaş ✉ toztas@pau.edu.tr Pamukkale University, Department of Business Administration, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

Economically, efficiency consists of technical and distribution components. The technical efficiency is that only one output is reduced, or an input is increased in order to increase an output (Koopmans, 1951). Technical efficiency is expressed more flexible, as the ability to produce as much output as possible to the extent allowed by technology and input, or the ability to avoid waste during the use of the smallest input allowed by technology for output production. The distribution component refers to the ability to combine inputs and/or outputs at optimal rates considering current prices (Fried et al., 2008).

Efficiency measurement approaches can be grouped under three headings generally. These headings are in the form of ratio analysis, parametric methods and nonparametric methods. These approaches discussed in the following.

Ratio Analysis: Ratio analysis is used with the thought that the performance of the company will be reflected on the balance sheet. With the help of balance sheets, useful information about the company can be obtained and forecasts can be made about the future situation. Although the ratio analysis correctly reflects the situation of companies, there are some limitations. These limitations are: There is no criterion for choosing rates that everyone can accept and added, or simplified ratios may not meet the needs of users (Ho & Zhu, 2004).

Parametric Methods: Parametric methods are based on certain functional form assumptions for the efficient frontier. Parametric approaches are divided into deterministic and stochastic models. In deterministic models, all observations by frontier and existing technology are enveloped as technical inefficiency by determining the difference between observed production and maximum production (Murillo-Zamorano, 2004). The most widely used method in the parametric approach is Stochastic Frontier Analysis.

Nonparametric Methods: Nonparametric methods avoid enforcing the production frontier in a specific functional form (Elyasiani & Mehdian, 1990). Since these approaches do not have parametric constraints, they can easily handle separated inputs and multiple output technologies (Chavas & Aliber, 1993). Nonparametric techniques attract great attention in the literature. The basic reason is that few assumptions are needed, and there is no need to define the functional form of the relationship between inputs and outputs and to specify a form of distribution in terms of inefficiency (Daraio & Simar, 2007). The most commonly used techniques in the literature are Data Envelopment Analysis (DEA) and Free Disposal Hull techniques.

The rest of the study is as follows: Section 2 focuses on Grey Relational Analysis (GRA), DEA, and Malmquist Total Factor Productivity Index, which are used for efficiency measurement. Section 3 gives a literature review about efficiency measurement with GRA and DEA methods. Section 4 presents a three-stage efficiency measurement for 20 food and beverage firms traded in BIST (Borsa İstanbul from Turkey) for the 2013-2015 period. Section 5 gives conclusions of the study.

2. METHOD

Organizations need to determine the correct input and output variables basically in order to accurately measure their efficiency. The main reason of this issue is the generation of large amounts of data during the activities carried out in the organizations. For this purpose, in this study, a three-stage approach has been adopted in the process of measuring the efficiency of BIST food and beverage Index firms between 2013-2015 years. In the first stage, the Grey Relational Analysis was used in the selection of the variables to be used for efficiency measurement. The selected variables were used as inputs and outputs of DEA model in the second stage. In the third and final stage, the Malmquist Total Factor Productivity Index was used to determine the efficiency changes and their causes between the periods.

2.1. Grey Relational Analysis

Grey Relational Analysis is a related concept of Grey System theory. The Grey System is defined as a system containing knowns and unknowns by Ju-Long Deng (1982). Grey systems and its applications have interdisciplinary properties aimed filling gaps between social sciences and natural sciences (Deng, 1989). The word "grey" in Grey System theory or Grey Relational Analysis means a status between black and white. White states certain knowledge, while black states completely missing knowledge. In this case, grey is a mixture of black and white (Ng, 1994).

Grey Relational Analysis suggests a relationship in order that the degree of correlation of factors can be measured. Accordingly, the more similarity between the factors, the more the correlation is to be mentioned. The Grey Relational ratios are used to measure the degree of relationship between the factors (Kung & Wen, 2007).

In order to calculate the correlations between the factors with Grey Relational Analysis, the first step is to perform the normalization process to remove the measurement differences between the factors. Normalization can be done according to whether the factors are benefit or cost attributes. Equation (1) is used for factors with benefit attribute, and Equation (2) is used for cost attribute ones (Wang, 2008). Hereby, $x_i^{(O)}(k)$ is comparability sequence.

$$x_i^*(k) = \frac{x_i^{(O)}(k)}{\sqrt{\sum_{t=1}^{m}\left[x_i^{(O)}(t)\right]^2}} \tag{1}$$

$$x_i^*(k) = \frac{1/x_i^{(O)}(k)}{\sqrt{\sum_{t=1}^{m}\left[1/x_i^{(O)}(t)\right]^2}} \tag{2}$$

After the normalization process is completed, $x_0^*(k)$ reference series that consists of the ideal values are determined (Ertugrul, Oztas, Ozcil, & Oztas, 2016). The Grey Relational coefficients measure the closeness of $x_i^*(k)$ and $x_0^*(k)$ (reference) series. Grey Relational coefficient is calculated as shown in Equation (3) (Kuo, Yang, & Huang, 2008).

$$\gamma\left(x_0^*(k), x_i^*(k)\right) = \frac{\Delta_{min}+\xi\Delta_{max}}{\Delta_{ik}+\xi\Delta_{max}}, i = 1, \dots, m, \;\; k = 1, \dots, n \tag{3}$$

$$\Delta_{ik} = |x_0^*(k) - \; x_i^*(k)|$$

$$\Delta_{min} = Min\{\Delta_{ik}, i = 1, \dots, m, \qquad k = 1, \dots, n\}$$

$$\Delta_{max} = Max\{\Delta_{ik}, i = 1, \dots, m, \qquad k = 1, \dots, n\}$$

In Equation (3), ξ is the distinguishing coefficient in [0, 1] interval, and Δ_{ik} is the deviation sequence of reference sequence and comparability sequence. Grey Relational grade is equal to the weighted average of the Grey Relational coefficients. These values are calculated as shown in Equation (4) (Tzeng, Lin, Yang, & Jeng, 2009).

$$\gamma(x_0^*, x_i^*) = \sum_{k=1}^{n} w_k \gamma\left(x_0^*(k), x_i^*(k)\right), \quad \sum_{k=1}^{n} w_k = 1 \tag{4}$$

2.1.1. *The Selection of Representative Indicator*

Grey Relational Analysis can be used for clustering and determining the factors that represent clusters when many variables exist in efficiency measurement. In the case of *m* decision-making units, *t* periods, and *s* factors the Grey Relational grade is calculated to be similar to Equation (4) (Wang, 2014).

$$r_{0i} = \gamma(x_0^*, x_i^*) = \frac{1}{mt}\sum_{k=1}^{mt} \gamma\left(x_0^*(k), x_i^*(k)\right) \tag{5}$$

Grey Relational matrix $R = \left(r_{ij}\right)$ $(i = 1, \dots, s, j = 1, \dots, s)$ is obtained by Grey Relational analysis. Clustering is done according to the following definitions (Wang, 2014).

Definition 1: If $r_{ij} \geq r$ and $r_{ji} \geq r$, then x_i^* and x_j^* is in the same cluster. Where, r is threshold valued and generally selected as 0.75 in literature.

Definition 2: In case, $r_{ij} \geq r$, $r_{ji} \geq r$, $r_{ik} \geq r$, $r_{ki} \geq r$, but $r_{jk} < r$ or $r_{kj} < r$. If $min\left\{r_{ij}, r_{ji}\right\} \geq min\{r_{ik}, r_{ki}\}$, then x_i^* and x_j^* is in the same cluster.

Definition 3: If x_i^* and x_j^* are in the same cluster, the biggest value of r_{ij} and r_{ji} represents the cluster. If $r_{ij} > r_{ji}$ then factor *i* represents the cluster.

Definition 4: Suppose that x_i^*, x_j^*, and x_k^* are in the same cluster. Representative factor of cluster is determined according to the biggest value of $r_{ij} + r_{ik}$, $r_{ji} + r_{jk}$, and $r_{ki} + r_{kj}$. For instance, if the biggest value is $r_{ij} + r_{ik}$, then representative indicator is factor *i*.

Definition 5: Suppose that T is a cluster consists of four or more elements. The representative factor of cluster will be factor *i*, if $\sum_{j(\neq i)\in T} r_{ij} > \sum_{j(\neq k)\in T} r_{kj}$, $\forall k \in T \; and \; k \neq i$.

2.2. Data Envelopment Analysis

Data Envelopment Analysis is a method introduced by Charnes, Cooper, and Rhodes in 1978. It is based on a methodology that essentially eliminates the assumptions and limitations of classical efficiency measurement approaches (Bowlin, 1998). Data Envelopment Analysis evaluates the relative efficiencies of production units with multiple inputs and multiple outputs. The basic idea of Data Envelopment Analysis is to develop a methodology which determines the decision-making units that have the best function within the set of comparable decision-making units (DMU) and forms an efficiency frontier (Cook & Seiford, 2009). Data Envelopment Analysis can be used to measure the performance of non-profit organizations as well as to measure the performance of profit-oriented organizations (Doyle & Green, 1994).

2.2.1. CCR Model

In the CCR model, the efficiency measurement of any decision-making unit is obtained by maximizing the weighted output to weighted inputs ratio under constraints where the similar rates for each decision-making unit are equal to or less than 1. The model can be expressed mathematically as shown in Equation (6) (Charnes, Cooper, & Rhodes, 1978).

$$\max \theta = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}}$$

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1; j = 1, \dots, n \qquad (6)$$

$$v_r, u_i \geq 0; r = 1, \dots., s; i = 1, \dots, m$$

In the case of the model discussed in Equation (6), if the decision unit having $\theta^* = 1$ and at least one positive optimal value (v^*, u^*) exists, this decision unit is the CCR efficient; otherwise, CCR inefficient. Moreover, since the optimal $\theta = \theta^*$ values are not affected by the measurement unit of the input and output variables, they are called units invariance (Cooper, Seiford, & Tone, 2007).

2.2.2. BCC Model

The BCC model was developed in 1984 by Banker, Charnes, and Cooper. This model is derived from the convexity constraint added to the CCR model, which is based on the assumption of constant returns to scale (Cooper et al., 2007; Banker & Thrall, 1992). The variable associated with this added constraint makes it possible to comment on the returns to scale (increase, decrease, or constant) when evaluating the technical efficiencies (or inefficiencies) of the decision-making units (Ahn, Charnes, & Cooper, 1988). The model is as shown in Equation (7) (Banker, Charnes, & Cooper, 1984).

$$\begin{gathered}
\min \theta - \varepsilon\left(\textstyle\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right); \\
\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{i0};\, i = 1, \dots, m \\
\textstyle\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = \theta y_{r0};\, r = 1, \dots, s \\
\sum_{j=1}^{n} \lambda_j = 1 \\
\lambda_j, s_i^-, s_r^+ \geq 0, \forall i, j, r
\end{gathered} \tag{7}$$

Scale efficiencies of decision-making units can be determined by using efficiency scores of CCR and BCC models. If the CCR efficiency score is considered as technical efficiency and the BCC efficiency score as pure efficiency score, the scale efficiency is calculated as shown in Equation (8) (Cooper et al., 2007).

$$SE = \frac{\theta_{CCR}^*}{\theta_{BCC}^*} \tag{8}$$

2.3. Malmquist Total Factor Productivity Index

The changes in the productivity of decision-making units can be explained by the Malmquist Total Factor Productivity Index in terms of the change in the technical efficiency and the change in the technology over the time (Färe, Grosskopf, Norris, & Zhang, 1994). As the efficiency score for each decision-making unit is being produced with taking reference to the technologies of efficient decision-making units with Data Envelopment Analysis; Productivity changes between t_1 and t_2 periods are determined by the Malmquist productivity index (Berg, Førsund, & Jansen, 1992). The Malmquist index identifies changes in productivity as multiple input or multiple output oriented with the distance functions (Coelli & Rao, 2005). The Malmquist efficiency index, calculated by *x* inputs and *q* outputs between two periods such as *s* and *t* (the reference period) as shown in Equation (9) (Coelli et al., 2005).

$$m_o^t(q_s, q_t, x_s, x_t) = \frac{d_0^t(q_t, x_t)}{d_0^t(q_s, x_s)} \tag{9}$$

Hereby, d_0^t is a distance function that measures the efficiency of the conversion of x_t inputs to q_t outputs in the period *t*. If the m_o value is greater than 1, then it means progress, and if it is less than 1, it means regression.

The performance change between the two periods in the Malmquist productivity index is based on the geometric mean of the calculated index values for both periods.

$$m_o(q_s, q_t, x_s, x_t) = \left[\frac{d_0^s(q_t, x_t)}{d_0^s(q_s, x_s)} \frac{d_0^t(q_t, x_t)}{d_0^t(q_s, x_s)}\right]^{1/2} \tag{10}$$

When Equation (10) is arranged, an index is obtained that has two components that measure efficiency and technology levels and allows inefficiency (Färe, Grosskopf, Lindgren, & Roos 1992).

$$m_o(q_s, q_t, x_s, x_t) = \frac{d_0^t(q_t, x_t)}{d_0^s(q_s, x_s)} \left[\frac{d_0^s(q_t, x_t)}{d_0^t(q_t, x_t)} \frac{d_0^s(q_s, x_s)}{d_0^t(q_s, x_s)}\right]^{1/2} \tag{11}$$

The first part of Equation (11) measures the change in efficiency, while the second part measures the change in technology.

3. LITERATURE REVIEW

This section provides a literature review of studies with similar approaches to efficiency/performance measurement of this paper.

Feng and Wang (2000), used Grey Relational Analysis and TOPSIS methods to measure the performance of airline companies. A total of 63 financial indicators were considered in the study, and with the help of Grey Relational Analysis, fewer indicators were used instead of all the indicators. After the representative indicators were identified, the performance of the 5 airlines was determined by TOPSIS method.

Wang, Ma and Guan (2007), measured the efficiencies of 24 hospitals in China with Grey Relational Analysis and Data Envelopment Analysis. In the first part of the study, 2 inputs and 7 output variables were specified. Using Grey Relational Analysis, the output variables were grouped and the number of variables was reduced to 3 using the representative variables in each group. Then, Data Envelopment Analysis was used to determine efficient hospitals with a model with 2 input-3 output variables.

Wang (2007), utilized the Grey Relational Analysis and Data Envelopment Analysis to evaluate the performance of the TFT-LCD industry in Taiwan. Grey Relational Analysis was used to objectively select variables to be used in Data Envelopment Analysis and to simplify calculations by reducing the number of variables. After the variables were determined, efficient firms were obtained by measuring production efficiency and marketing effectiveness with a two-stage evaluation process with Data Envelopment Analysis.

Chiang-Ku, Shu-Wen and Cheng-Ru (2009), compared the performances of the traditional sales channel, and the bank sales channel which sell policies for an insurance company. The comparison has two stages: Marketability efficiency and profitability efficiency. Variables to be used to measure the efficiency of sales channels were first identified by a Delphi panel consisting of 10 experts, then those with the highest correlation with Grey Relational Analysis were identified as input variables. Data Envelopment models for the two channels were built by using the input and output variables, and the results were analyzed by Mann-Whitney U test. The relationship between the two groups was analyzed by Spearman’s correlation.

Ho (2011), has combined Data Envelopment Analysis and Grey Relational Analysis methods to measure the efficiencies of dot-com companies. In the study, 69 companies that sell via the internet were examined. In the study, firstly 21 inputs and 19 output variables were determined, and the number of variables was reduced by Grey Relational Analysis. A Data Envelopment model was established to measure the efficiencies of dot-com companies with selected 4 input-4 output variables.

Wang (2014), measured the financial performance of container transportation companies using Grey Relational Analysis and fuzzy TOPSIS. In the study, 20 financial ratios were first divided into 4 categories and representing variables were determined with Grey Relational Analysis within each category. Then, the determined variables were used to order the performance of the three firms with the fuzzy TOPSIS method.

Girginer, Köse and Uçkun (2015), measured the efficiency of 10 surgical services in a hospital in Turkey using combined Data Envelopment Analysis and Grey Relational Analysis methods. In the study, efficient decision-making units were determined by performing efficiency measurement by Data Envelopment Analysis using 4 input variables and 2 output variables. Grey Relational Analysis was used to determine the factors that affect the ranking and efficiency of the performance of efficient decision-making units.

İç, Tekin, Pamukoğlu and Yıldırım (2015) compared corporate companies which operate in 24 sectors with the financial performance system that they developed. This model bases on financial ratios and TOPSIS method. In the modeling stage, using the correlation values obtained from TOPSIS, VIKOR, GRA, and MOORA methods, it was found that TOPSIS method is more suitable for this evaluation model.

Tsaur, Chen and Chan (2017), measured the performance of the Taiwan TFT-LCD industry in a four-stage process. In the first stage of the study, efficiency scores were determined with Data Envelopment models for each company between 2009-2012 years. In the second stage, the Malmquist index and the efficiency changes in companies were analyzed. In the third stage, Grey Relational Analysis was performed by determining the weights of input and output variables by entropy method. In the fourth step, the results of the methods were compared, and the results were concluded.

Durga Prasad, Venkata Subbaiah and Prasad (2017) used Data Envelopment Analysis, Analytic Hierarchy Process and Grey Relational Analysis methods together for supplier selection. Efficiency values were computed with Data Envelopment Analysis. The best supplier was selected with Grey Relational Analysis. In this stage, weights of criteria were determined using AHP method.

Pakkar (2017), used Data Envelopment Analysis and Analytic Hierarchy Process methods to develop a Grey Relational Analysis model that have multi-hierarchy. In the method, a multi-featured decision-making process was transformed into a two-level hierarchical structure of attributes and attribute categories. In the first step, the required data were obtained by calculating with simple Grey Relational Analysis and Analytic Hierarchy Process at the attribute level for additive Data Envelopment Analysis model. In the second step, Grey Relational grades of attributes were transformed into Grey Relational coefficients of the categories. For the alternatives, the Grey Relational grades of the categories were calculated by using the Data Envelopment Analysis model and the dissimilarity scores of the categories for the tied alternatives are calculated by the exclusive Data Envelopment Analysis exclusion model.

Pakkar (2018), used Grey Relational Analysis method for multi-attribute decision-making problems which its weights are unknown and in fuzzy number form. Data Envelopment Analysis and Analytic Hierarchy Process methods were used for determination of weights. For this purpose, two sets of weights based on the minimax Data Envelopment Analysis were defined in the framework of Grey Relational Analysis. The first set states weights with the minimum Grey Relational loss; the second set states weights with the maximum Grey Relational loss by using the Analytic Hierarchy Process. The model was exemplified by the selection of a nuclear waste disposal site.

Huang, Dai and Guo (2015) have developed a new Data Envelopment Analysis model for corporate financial failure prediction. The model has two stages and has been developed in order to be able to quickly deal with a large number of inputs and outputs, making use of the hierarchical structure of financial indicators. The Grey Relational Analysis method was used to select the indicators that have a significant correlation among a large number of indicators.

Hsu (2015), has combined Data Envelopment Analysis with the Grey Relational Analysis method, which was developed to examine the activities and performance of semiconductor companies in an increasingly competitive environment. In this regard, two groups of efficient and inefficient semiconductor companies were obtained. Then, efficient and inefficient companies were examined in terms of their operational performance by multi-criteria decision-making method, improved Grey Relational Analysis method and Entropy weight method.

Kaygısız Ertuğ and Girginer (2015) were investigated fiscally metropolitan municipalities in Turkey with Data Envelopment Analysis and Grey Relational Analysis in an integrated manner. Firstly, efficient and inefficient municipalities were determined with Data Envelopment Analysis and then the efficient municipalities ranked with Grey Relational Analysis. Thus, the municipalities with the best and worst performance have been identified.

4. FINDINGS

The main idea of this study is to perform the evaluation process objectively while measuring the efficiency. The number of input and output variables and selection of these variables have a big influence on the quality of the evaluation results. A three-stage hybrid approach has been adopted to study this controversial case in a scientific approach. The approach adopted for the measurement of efficiency has been applied to BIST food and beverage Index firms and the results have been examined. Figure 1 depicts visually the stages of the study.

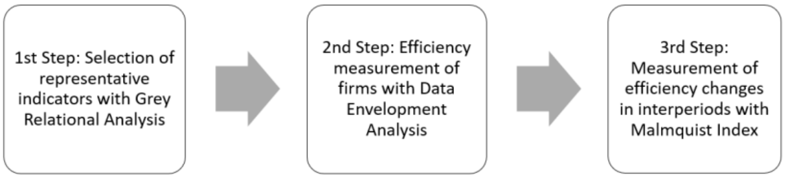

1st Step: Selection of representative indicators with Grey Relational Analysis
2nd Step: Efficiency measurement of firms with Data Envelopment Analysis
3rd Step: Measurement of efficiency changes in interperiods with Malmquist Index

Figure 1. Stages of the analysis

4.1. Material and Method

The financial ratios related to the firms included in the BIST food and beverage index were used as input and output variables in the study. The financial data used in the study covers 3 periods from 2013 to 2015. These ratios were calculated by using Financial Analysis reports of firms which obtained from Bloomberg terminals. The firms included in the scope of the study are listed in Table 1.

At the first stage of the study, 17 financial ratios were chosen to determine the input and output variables to be used for efficiency measurement and these ratios were divided into 4 categories. Three categories related to liquidity ratios, financial structure ratios, and operating ratios were used in determining the representative input variables, and profitability ratios were used in determining the representative output variables. These categories and ratios are as shown in Table 2.

Table 1. Analysed firms

No	Firm	No	Firm	No	Firm	No	Firm
1	AEFES	6	PETUN	11	ULUUN	16	PINSU
2	ULKER	7	TBORG	12	AVOD	17	KENT
3	CCOLA	8	BANVT	13	KERVT	18	ALYAG
4	TATGD	9	KRSTL	14	KNFRT	19	ERSU
5	PNSUT	10	TUKAS	15	PENGD	20	MERKO

Performing efficiency measurement with all 17 ratios in Table 2 makes calculations hard. For this reason, it is necessary to work with fewer ratios. From these ratios in Table 2, it is very important that selection of input/output variables in terms of the efficiency measurement results and the models to be built. For this reason, in order to make the variable selection objectively, the Grey Relational Analysis is used in the first step of the study to divide the ratios within each category into clusters and to determine the ratios that would represent the other ratios in the cluster. To eliminate the measurement differences of the data in the grey relation analysis, normalization was performed according to the benefit and cost information in the attribute column.

After the input and output variables used in the study were determined, the efficiency measurement was performed by Data Envelopment Analysis in the second stage of the study. An input-oriented BCC model was used for the measurement of efficiency. In the third stage of the study after the efficiency scores were obtained, the Malmquist total factor productivity index was used to analyze the changes in the efficiency of the firms and the industry between periods. Microsoft Office Excel and DEAP 2.1 programs were used in calculations.

Table 2. The financial ratios used in the study

	Ratio	Code	Indicator	Formulation	Attribute
Input	Liquidity ratios	L1	Cash ratio	Cash and marketable securities/Current liabilities	Benefit
		L2	Current ratio	Current assets/Current liabilities	Benefit
		L3	Acid-test ratio	(Current assets-inventories)/ Current liabilities	Benefit
	Financial structure ratios	M1	Debt ratio	Total liabilities/total assets	Cost
		M2	Debt to equity ratio	Total debt/ Average shareholders' equity	Cost
		M3	Short-term debt to assets ratio	Short-term debts/Total assets	Cost
		M4	Fixed assets to equity ratio	Fixed Assets/ Average shareholders' equity	Cost
	Operating ratios	F1	Accounts receivable turnover	Net sales/Average net receivables	Benefit
		F2	Inventory turnover	Net sales/Average inventory	Benefit
		F3	Equity turnover	Net sales/Equity	Benefit
		F4	Asset turnover	Net sales/Total assets	Benefit
		F5	Current assets turnover	Net sales /Current assets	Benefit
		F6	Fixed assets turnover	Net sales /Fixed assets	Benefit
Output	Profitabili ty ratios	K1	Gross profit margin	Gross profit/Net sales	Benefit
		K2	Operating margin	Operating Income/ Net sales	Benefit
		K3	Profit margin	Net profit/Net sales	Benefit
		K4	Return on equity	Net income/Average shareholders' equity	Benefit

4.2. Determination of Representative Indicators Using GRA

As variables were determined by Grey Relational Analysis, the measurement values were normalized to the cost or benefit attribute. After the normalization process, the reference series were constructed and the difference series were formed by the comparison series. From the difference series, the Grey Relational coefficients were obtained with the help of Equation (3), and the Grey Relational grades were obtained by taking the averages of these values. Each of the ratios was selected as the reference series to obtain the grey relation matrix consisting of Grey Relational grades and clustering was performed according to this matrix. The following matrices show the Grey Relational matrices and Table 3 shows representative ratios of the clusters obtained for each category group.

$$R_1 = \begin{bmatrix} 1 & 0.756 & 0.804 \\ 0.756 & 1 & 0.870 \\ 0.798 & 0.866 & 1 \end{bmatrix},$$

$$R_2 = \begin{bmatrix} 1 & 0.850 & 0.821 & 0.775 \\ 0.854 & 1 & 0.803 & 0.783 \\ 0.839 & 0.817 & 1 & 0.757 \\ 0.797 & 0.798 & 0.757 & 1 \end{bmatrix},$$

$$R_3 = \begin{bmatrix} 1 & 0.896 & 0.842 & 0.903 & 0.922 & 0.868 \\ 0.895 & 1 & 0.838 & 0.904 & 0.917 & 0.866 \\ 0.843 & 0.841 & 1 & 0.835 & 0.839 & 0.893 \\ 0.904 & 0.906 & 0.834 & 1 & 0.925 & 0.901 \\ 0.923 & 0.919 & 0.839 & 0.926 & 1 & 0.872 \\ 0.864 & 0.864 & 0.889 & 0.897 & 0.867 & 1 \end{bmatrix},$$

$$R_4 = \begin{bmatrix} 1 & 0.789 & 0.723 & 0.727 \\ 0.778 & 1 & 0.805 & 0.781 \\ 0.723 & 0.816 & 1 & 0.890 \\ 0.680 & 0.750 & 0.865 & 1 \end{bmatrix},$$

Table 3. Clusters and their representative indicators

Cluster	Ratios in cluster	Representative indicator
C1	L1, L2, L3	L3 (Acid-test ratio)
C2	M1, M2, M3, M4	M1 (Debt ratio)
C3	F1, F2, F3, F4, F5, F6	F5 (Current assets turnover)
C4	K1, K2	K1 (Gross profit margin)
C5	K3, K4	K3 (Profit margin)

For example, in the Grey Relational matrix for the liquidity ratios in R_1, L1, L2, and L3 are in the same cluster because $r_{12}, r_{13}, r_{21}, r_{23}, r_{31}$, and r_{32} are greater than the threshold value 0.75. The ratio of L3 (acid-test ratio) was chosen because the biggest value of r_{12}+r_{13}, $r_{21} + r_{23}$, and $r_{31} + r_{32}$ is $r_{31} + r_{32} = 1.66$ as mentioned in the second section. Other ratios were determined by a similar approach.

As a result of the clustering process with Grey Relational Analysis, 17 financial ratios were represented with 5 financial ratios. This process provides a reduction of approximately 70% of the number of ratio, which will make the calculations with the Data Envelopment Analysis easier to complete. The input variables consist of acid test ratio (L3), debt ratio (M1) and current assets turnover rate (F5) while output variables are gross profit margin (K1) and profit margin (K3). These ratios and general information are given below respectively.

–*Acid-test ratio:* It may not be easy to take stocks out of hand in the short run because they cannot always be quickly converted into cash. In short-term payments, it helps to determine the liquidity position of the firm by reducing inventories from current assets (Dyson, 2010). It provides a more accurate measure of the payment power than the current ratio (Tayyar, Akcanlı, Genç, & Erem, 2014).

–*Debt ratio:* This rate shows how the firm finances its assets by borrowing in various forms. The higher this rate, the higher the financial risk; the lower the rate, the lower the financial risk (Van Horne & Wachowicz, 2008).

–*Current assets turnover ratio:* It is used to measure the relationship between sales and current asset investments. It expresses firm how many times turns over its current assets in a year.

The higher the rate, the more efficient use of current assets (Wahlen, Baginski, & Bradshaw, 2011). For this reason, it can be used to measure operational performance (Yu, Luo, Feng, & Liu, 2018).

–*Gross profit margin ratio:* Gross profit is the difference between sales revenue and selling cost. Gross profit is, therefore, a measure of the profitability of the procurement (production) and sale of goods or services before other costs are added to the account. Since the cost of sales is a huge expense for many businesses, a change in that location can be a major impact on the profit or loss of the respective year (Atrill, 2012). This ratio is sensitive to pricing, product mix, and unit costs but is not based on sales volume (Isberg & Pitta, 2013).

–*Profit margin*: Net profit margin is a measure of the profitability of sales considering all costs and income of the company. It refers to the net income per unit of money company's sales (Van Horne & Wachowicz, 2008). In a simpler sense, it is the periodic net profit rate that a firm has achieved net sales (Önem & Demir, 2015). The values of the rates selected using Grey Relational Analysis are as shown in Table 4, 5 and 6.

Table 4. Values of representative indicators for the year 2013

Firm	K1	K3	L3	M1	F5
AEFES	0.435	0.047	1.018	0.398	2.321
ULKER	0.230	0.066	0.993	0.599	1.253
CCOLA	0.378	0.094	1.026	0.590	2.410
TATGD	0.209	0.003	0.973	0.609	1.770
PNSUT	0.186	0.083	0.759	0.298	3.544
PETUN	0.173	0.080	0.953	0.245	3.412
TBORG	0.553	0.181	1.030	0.489	1.971
BANVT	0.120	-0.034	0.334	0.881	2.856
KRSTL	0.182	0.054	3.914	0.129	1.116
TUKAS	0.143	-0.291	0.598	0.819	0.835
ULUUN	0.071	0.012	0.687	0.743	2.568
AVOD	0.183	-0.008	0.316	0.500	0.782
KERVT	0.278	-0.165	0.253	1.058	1.742
KNFRT	0.283	0.067	0.432	0.455	0.948
PENGD	0.041	-0.288	0.420	0.628	0.976
PINSU	0.406	-0.079	0.470	0.439	3.541
KENT	0.291	-0.027	0.809	0.371	2.222
ALYAG	0.109	0.053	0.369	0.306	3.664
ERSU	0.097	-0.020	0.758	0.335	1.110
MERKO	0.189	-0.042	0.151	0.797	1.678

Source: Bloomberg

Table 5. Values of representative indicators for the year 2014

Firm	K1	K3	L3	M1	F5
AEFES	0.429	0.004	1.037	0.412	2.119
ULKER	0.210	0.073	2.492	0.614	1.388
CCOLA	0.364	0.053	0.819	0.532	2.370
TATGD	0.211	0.184	1.098	0.473	1.717
PNSUT	0.168	0.093	0.809	0.321	3.672
PETUN	0.149	0.080	0.870	0.226	4.275
TBORG	0.560	0.205	1.294	0.462	1.669
BANVT	0.131	-0.011	0.381	0.907	3.111
KRSTL	0.069	0.006	1.478	0.258	1.095
TUKAS	-0.038	-0.412	0.479	0.660	0.656
ULUUN	0.064	0.015	0.762	0.670	2.527
AVOD	0.125	0.006	0.469	0.539	1.769
KERVT	0.281	-0.070	0.276	1.067	1.783
KNFRT	0.320	0.154	1.569	0.166	1.245
PENGD	0.111	-0.114	0.349	0.676	1.357
PINSU	0.430	0.016	0.636	0.518	3.716
KENT	0.294	0.036	1.230	0.367	2.266
ALYAG	0.048	-0.032	0.299	0.434	2.557
ERSU	0.095	-0.048	1.153	0.263	1.316
MERKO	0.205	0.088	0.761	0.427	2.942

Source: Bloomberg

Table 6. Values of representative indicators for the year 2015

Firm	K1	K3	L3	M1	F5
AEFES	0.410	-0.019	1.155	0.430	2.162
ULKER	0.217	0.084	3.012	0.582	1.380
CCOLA	0.347	0.017	1.025	0.537	2.740
TATGD	0.226	0.074	1.244	0.361	1.834
PNSUT	0.161	0.062	0.577	0.336	3.630
PETUN	0.168	0.113	0.879	0.221	4.547
TBORG	0.548	0.212	1.473	0.439	1.410
BANVT	0.106	-0.050	0.278	0.792	3.729
KRSTL	0.076	0.025	2.116	0.211	1.348
TUKAS	0.203	0.233	0.265	0.553	0.997
ULUUN	0.076	0.008	0.756	0.663	2.366
AVOD	0.203	0.019	0.382	0.465	2.011
KERVT	0.277	-0.222	0.184	0.964	1.512
KNFRT	0.201	0.130	1.735	0.116	0.936
PENGD	0.198	0.036	0.250	0.692	1.328
PINSU	0.476	-0.062	0.324	0.641	3.305
KENT	0.359	0.093	1.232	0.320	1.883
ALYAG	0.051	-0.055	0.092	0.573	2.952
ERSU	0.145	-0.059	0.583	0.259	0.947
MERKO	0.186	0.009	0.263	0.682	1.517

Source: Bloomberg

4.3. Efficiency Measurement with Data Envelopment Analysis

When the values of the financial ratios are examined according to years, it is seen that some of the ratios related to profitability are negative. Data Envelopment Analysis has the constraint that the input and output values are not negative. Since the input-oriented BCC model has the translation invariant property for the output variables, the shift in the output variables will not affect the efficiency result (Lovell & Pastor 1995; Pastor 1996). From this point, if there is more than one negative value in a variable, the sign problem is solved by adding the smallest value to all the variables will make all of them positive. All decision-making units have thus participated in the evaluation process. The results of the calculations made, the efficiency scores according to years are as shown in Table 7.

Table 7. Efficiency scores of firms according to years

	2013			2014			2015		
Firm	BCC	Scale Efficiency	Returns to Scale	BCC	Scale Efficiency	Returns to Scale	BCC	Scale Efficiency	Returns to Scale
AEFES	1	0.963	irs	1	0.927	irs	0.93	0.908	irs
ULKER	0.756	0.999	irs	0.836	0.919	irs	0.694	0.818	irs
CCOLA	0.736	0.953	irs	0.965	0.975	irs	0.76	0.885	irs
TATGD	0.647	0.858	irs	1	1	-	0.779	0.891	irs
PNSUT	0.996	0.989	drs	1	1	-	0.983	0.846	irs
PETUN	1	1	-	1	1	-	1	1	-
TBORG	1	1	-	1	1	-	1	1	-
BANVT	0.704	0.918	irs	1	0.834	drs	0.675	0.623	irs
KRSTL	1	1	-	1	0.84	irs	0.7	0.733	irs
TUKAS	0.937	0.613	irs	1	0.004	irs	1	1	-
ULUUN	0.579	0.885	irs	0.726	0.986	irs	0.569	0.652	irs
AVOD	1	1	-	1	1	-	0.978	0.777	irs
KERVT	1	1	-	1	1	-	1	1	-
KNFRT	1	1	-	1	1	-	1	1	-
PENGD	0.801	0.183	irs	1	0.951	irs	0.939	0.863	irs
PINSU	1	1	-	1	1	-	1	1	-
KENT	0.985	0.726	irs	0.867	0.881	irs	1	0.9	irs
ALYAG	1	1	-	1	1	-	1	1	-
ERSU	1	0.872	irs	1	0.733	irs	1	0.657	irs
MERKO	1	1	-	1	0.973	drs	0.88	0.797	irs
Average	0.907	0.898		0.97	0.901		0.894	0.867	

Then the firms' 2013 efficiency scores are analysed, it is seen that 11 firms are technical efficient according to BCC model. These firms are respectively AEFES, PETUN, TBORG, KRSTL, AVOD, KERVT, KNFRT, PINSU, ALYAG, ERSU and MERKO. Among the 9 technical inefficient firms, 8 firms have increasing returns to scale, but only PNSUT has decreasing returns to scale. ULUUN has shown the lowest performance in terms of technical efficiency among inefficient firms. The average efficiency score of the industry for 2013 was measured as 0.907.

In 2014, AEFES, TATGD, PINSUT, PETUN, TBORG, BANVT, KRSTL, TUKAS, AVOD, KERVT, KNFRT, PENGD, PINSU, ALYAG, ERSU and MERKO firms were determined as technical efficient. All the inefficient firms have increasing returns to scale. The relative lowest performing firm is ULUUN in 2014. In 2014, the average technical efficiency score of the sector was measured as 0.97 and it was observed an increase in efficiency score of the sector according to the previous year.

In 2015, PETUN, TBORG, TUKAS, KERVT, KONFRT, PINSU, KENT, ALYAG and ERSU were found as technical efficient. All inefficient firms have increasing returns to scale and ULUUN has the lowest relative performance. In 2015, the average technical efficiency score of the sector was measured as 0.894, which is lower than the previous year. Among the firms, PETUN, TBORG, KERVT, KNFRT, PINSU, ALYAG and ERSU firms are efficient in all three periods. This shows that the firms manage the inputs and outputs well. ULUUN firm, however, has shown its worst performance in all three periods, so it appears that it cannot use its resources effectively.

4.4. Malmquist Index

The Malmquist index established to determine the inter-period efficiency and technology changes of the firms are as shown in Table 8. In the table if the values are bigger than 1, then progress is discussed; if the values are smaller than 1, then regression discussed otherwise, there is no change.

Table 8. Malmquist index values by periods

	2013-2014					2014-2015				
Firm	effch	techch	pech	sech	tfpch	effch	techch	pech	sech	tfpch
AEFES	0.962	1.089	1	0.962	1.048	0.911	0.859	0.93	0.98	0.783
ULKER	1.018	1.101	1.106	0.92	1.12	0.738	0.923	0.83	0.889	0.681
CCOLA	1.343	1.013	1.312	1.023	1.36	0.715	0.907	0.788	0.908	0.648
TATGD	1.801	1.175	1.545	1.166	2.116	0.694	0.811	0.779	0.891	0.563
PNSUT	1.015	1.223	1.004	1.011	1.241	0.832	0.816	0.983	0.846	0.679
PETUN	1	1.39	1	1	1.39	1	0.708	1	1	0.708
TBORG	1	1.138	1	1	1.138	1	0.905	1	1	0.905
BANVT	1.289	1.08	1.419	0.908	1.393	0.505	1.267	0.675	0.748	0.64
KRSTL	0.84	1.454	1	0.84	1.222	0.61	0.815	0.7	0.872	0.498
TUKAS	0.008	1.152	1.068	0.007	0.009	226.876	1.33	1	226.876	301.805
ULUUN	1.396	1.021	1.253	1.114	1.425	0.518	1.06	0.784	0.661	0.549
AVOD	1	0.875	1	1	0.875	0.76	1.1	0.978	0.777	0.837
KERVT	1	1.122	1	1	1.122	1	1.098	1	1	1.098
KNFRT	1	1.297	1	1	1.297	1	0.749	1	1	0.749
PENGD	6.499	0.904	1.248	5.207	5.872	0.852	1.415	0.939	0.907	1.206
PINSU	1	0.986	1	1	0.986	1	1.145	1	1	1.145
KENT	1.067	1.214	0.88	1.212	1.296	1.179	0.803	1.153	1.022	0.947
ALYAG	1	1.098	1	1	1.098	1	1.191	1	1	1.191
ERSU	0.84	1.453	1	0.84	1.221	0.896	0.79	1	0.896	0.708
MERKO	0.973	0.936	1	0.973	0.91	0.721	1.02	0.88	0.819	0.735
Average	0.912	1.125	1.08	0.844	1.026	1.083	0.966	0.913	1.187	1.047

When Table 8 analyzed in terms of firms, progress or regression in efficiency values can be determined over the periods. For instance, the AEFES firm has regressed in technical efficiency change (effch) and scale efficiency change (sech), progressed in technology change (techch) and remained constant pure technical efficiency change (pech) in the 2013-2014 period. Total factor productivity change (tfpch) of the firm increased by 4.8% in this period. AEFES firm has regressed in terms of all factors between the periods of 2014-2015. In this period, total factor productivity of the firm decreased 21.7%. Although it is possible to make these interpretations for all firms, it is noteworthy that TUKAS changes its efficiency level depending on the production factors. This can be attributed to the company's net losses in 2013 and 2014, its net profit in 2015 and its sale in 2014 (Hürriyet, 2014).

In the 2013-2014 period, the sector regressed in terms of technical efficiency change and scale efficiency change, but it progressed in terms of technology change and pure technical efficiency change between 2013 and 2014. The total productivity of the sector increased by 2.6%. In the 2014-2015 period, the sector progressed in terms of technical efficiency change and scale efficiency change period 2014-2015, it regressed in terms of technology change and pure technical efficiency. The total productivity of the sector increased by 4.7%.

5. CONCLUSION

An organization wants to monitor the process of transforming the inputs to the outputs regardless of its operating purpose. The main purpose of this is determining the problems that can cause inefficiency in the process of converting the scarce resources into goods or services. Data Envelopment Analysis, developed by Charnes, Cooper, and Rhodes in 1978, is a technique frequently used to measure the relative efficiencies of organizations in the literature. The method determines whether the decision-making units are efficient according to the efficiency scores. Inefficient decision-making units can determine how they can become efficient by reducing their inputs or increasing their outputs relative to slack variable values. In this sense, decision-makers can manage resources more effectively.

One of the most crucial factors affecting the results of Data Envelopment Analysis is the determination of input and output variables. In this study, Grey Relational Analysis method was used to make the variable selection process objectively. Grey Relational Analysis is a method of determining correlations between factors by analyzing relations between reference series and comparison series. Since the method is used successfully in systems with known and unknown information, it is suitable for the variable selection process.

In the study, 17 financial ratios are divided into 4 categories at first. These categories are liquidity ratios, financial structure ratios, operating ratios, and profitability ratios. Within each category, similar variables were clustered with the help of Grey Relational Analysis. Then, the correlations were examined and the ratio with the highest correlation was determined as the representative indicator of the clusters. In this view, 17 variables were represented by 5 variables. Liquidity ratio, debt ratio, current asset turnover ratio were determined as input variables, gross profit margin, and profit margin were determined output variables as a result of the process.

An input-oriented BCC model was established after the variables to be used in the efficient measurement were determined. The efficiency values of 20 firms that are traded in the BIST food and beverage index were measured for 2013, 2014, and 2015. As a result of the analysis, PETUN, TBORG, KERVT, KNFRT, PINSU, ALYAG, ERSU firms were found to be relatively efficient in all three periods.

After the measurement of the efficiency, the change of efficiency of the firms between the periods was examined by Malmquist total factor productivity index. As a result of the examinations, 80% of firms for the period of 2013-2014 have progressed in terms of the total productivity factor and 20% have regressed. By contrast, in 2014-2015, 25% of firms have progressed in terms of total factor productivity, while 75% have regressed.

The use of the proposed three-step hybrid method will benefit from various aspects. Firstly, the organizations that want to measure efficiency can determine the variables to be used in the measurement process by analyzing the first step of the proposed method. Thus, the calculations can be made easier by defining the variables that will represent the other variables in the analysis process. With the help of representative indicators, it is possible to perform the efficiency measurement in a shorter time using the easily accessible software. Secondly, firm managers can compare their performance with the performance of their competitors by measuring the efficiency of their firm. If the measurement shows that the firm is efficient, the result is that the firm produces output(s) using the input(s) efficiently. However, if the firm is inefficient, firm managers can compare their firm with reference DMUs and eliminate the inefficiency factor. In this way, firms may become efficient by reducing their input(s) or by increasing their output(s). Thirdly, the proposed method allows firms to monitor changes in total productivity between periods and determine its causes. Thus, it can be determined that the change in total productivity is caused by the progress or regression of the sub-factors. In further studies, the selection process may be completed by using techniques such as the entropy method where there is a priority difference in financial ratios in the selection of representative indicators.

As a result, this proposed three-stage hybrid method can be used for efficient measurement in any sector/industry. The most important contribution of the proposed method to efficiency measurement applications is simplifying calculations and interpretation of findings when there are many variables and the operating periods. In that, firstly due to use of representative indicators it is possible to measure the efficiency with fewer variables. Selection of representative indicators enables to determine the more accurate variables according to properties of data. Secondly, changes in efficiency (progression, regression or remaining constant) and the causes of these changes can be observed between periods. For instance, if there is a regression in efficiency, decision-makers can detect the main reason and they can enhance trouble. In this way, it will be possible to determine permanently whether scarce resources in the economy are being used efficiently. In the further studies, similar efficiency measurements can be applied to other industries or nonprofit organizations. The effects of the numbers of variables and length of the period on the results can be analyzed in detail.

ORCID

İrfan Ertuğrul https://orcid.org/0000-0002-5283-191X
Tayfun Öztaş https://orcid.org/0000-0001-8224-5092

6. REFERENCES

Ahn, T., Charnes, A., & Cooper, W. W. (1988). Efficiency characterizations in different DEA models. *Socio-Economic Planning Sciences, 22*(6), 253-257.

Atrill, P. (2012). *Financial Management for Decision Makers* (6th ed.). Essex: Pearson Education.

Banker, R. D., & Thrall, R. M. (1992). Estimation of returns to scale using data envelopment analysis. *European Journal of Operational Research, 62*(1), 74-84.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078-1092.

Berg, S. A., Førsund, F. R., & Jansen, E. S. (1992). Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980-89. *The Scandinavian Journal of*

Economics, 94, Supplement. Proceedings of a Symposium on Productivity Concepts and Measurement Problems: Welfare, Quality and Productivity in the Service Industries, S211-S228.

Bloomberg. *Financial Analysis Reports*. Retrieved from Bloomberg Terminal.

Bowlin, W. F. (1998). Measuring performance: An introduction to data envelopment analysis (DEA). *The Journal of Cost Analysis, 15*(2), 3-27.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429-444.

Chavas, J. P., & Aliber, M. (1993). An analysis of economic efficiency in agriculture: A nonparametric approach. *Journal of Agricultural and Resource Economics, 18*(1), 1-16.

Chiang-Ku, F., Shu-Wen, C., & Cheng-Ru, W. (2009). Using GRA and DEA to compare efficiency of bancassurance sales with an insurer's own team. *The Journal of Grey System, 21*(4), 395-406.

Chorafas, D. N. (2015). *Business Efficiency and Ethics: Values and Strategic Decision Making*. New York, NY: Palgrave Macmillan.

Coelli, T. J., & Rao, D. S. P. (2005). Total factor productivity growth in agriculture: A Malmquist index analysis of 93 countries, 1980-2000. *Agricultural Economics*, *32*(s1), 115-134.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis* (2nd ed.). New York, NY: Springer Science & Business Media.

Cook, W. D., & Seiford, L. M. (2009). Data envelopment analysis (DEA)-Thirty years on. *European Journal of Operational Research*, *192*(1), 1-17.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007*). Data envelopment analysis: A comprehensive text with models, applications, references and DEA-Solver software* (2nd ed.). New York, NY: Springer Science & Business Media.

Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. New York, NY: Springer Science & Business Media.

Deng, J. L. (1982). Control problems of grey systems. *Systems & Control Letters*, *1*(5), 288-294.

Deng, J. L. (1989). Introduction to grey system theory. *The Journal of Grey System*, *1*(1), 1-24.

Doyle, J., & Green, R. (1994). Efficiency and cross-efficiency in DEA: Derivations, meanings and uses. *Journal of the Operational Research Society, 45*(5), 567-578.

Durga Prasad, K. G., Venkata Subbaiah, K., & Prasad M. V. (2017). Supplier evaluation and selection through DEA-AHP-GRA integrated approach-A case study. *Uncertain Supply Chain Management*, *5*(4), 369-382.

Dyson, J. R. (2010). *Accounting for non-accounting students* (8th ed.). Essex: Pearson Education.

Elyasiani, E., & Mehdian, S. M. (1990). A nonparametric approach to measurement of efficiency and technological change: The case of large U.S. commercial banks. *Journal of Financial Services Research, 4*(2), 157-168.

Ertugrul, I., Oztas, T., Ozcil, A., & Oztas, G. Z. (2016). Grey relational analysis approach in academic performance comparison of university: A case study of Turkish universities. *European Scientific Journal*, June 2016 Special Edition, 128-139.

Färe, R., Grosskopf, S., Lindgren, B., & Roos, P. (1992). Productivity changes in Swedish pharamacies 1980-1989: A non-parametric Malmquist approach. *Journal of Productivity Analysis*, *3*(1-2), 85-101.

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *The American Economic Review, 84*(1), 66-83.

Feng, C. M., & Wang, R. T. (2000). Performance evaluation for airlines including the consideration of financial ratios. *Journal of Air Transport Management, 6*(3), 133-142.

Fried, H. O., Lovell, C. A. K., & Schmidt, S. S. (2008). *Efficiency and productivity*. In H. O. Fried, C. A. K. Lovell, & S. S. Schmidt (Eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 3-91). New York, NY: Oxford University Press.

Girginer, N., Köse, T., & Uçkun, N. (2015). Efficiency analysis of surgical services by combined use of data envelopment analysis and gray relational analysis. *Journal of Medical Systems,39*: 56. DOI: 10.1007/s10916-015-0238-y

Ho, C. T. B. (2011). Measuring dot com efficiency using a combined DEA and GRA approach. *Journal of the Operational Research Society, 62*(4), 776-783.

Ho, C. T., & Zhu, D. S. (2004). Performance measurement of Taiwan's commercial banks. *International Journal of Productivity and Performance Management, 53*(5), 425-434.

Hsu, L. C. (2015). Using a decision-making process to evaluate efficiency and operating performance for listed semiconductor companies. *Technological and Economic Development of Economy, 21*(2), 301-331.

Huang, C., Dai, C., & Guo M. (2015). A hybrid approach using two-level DEA for financial failure prediction and integrated SE-DEA and GCA for indicators selection. *Applied Mathematics and Computation, 251*, 431-441.

Hürriyet. (2014). Türk salça devi Tukaş satıldı işte alıcısı. Retrieved from http://www.hurriyet.com.tr/ekonomi/turk-salca-devi-tukas-satildi-iste-alicisi-26962638

İç, Y. T., Tekin, M., Pamukoğlu, F. Z., & Yıldırım, S. E. (2015). Development of a financial performance benchmarking model for corporate firms. *Journal of the Faculty of Engineering and Architecture of Gazi University, 30*(1), 71-85.

Isberg, S., & Pitta, D. (2013). Using financial analysis to assess brand equity. *Journal of Product & Brand Management, 22*(1), 65-78.

Kaygısız Ertuğ, Z., & Girginer, N. (2015). Bütünleşik VZA ve GİA yöntemleriyle büyükşehir belediyelerinin mali etkinlik analizi: Türkiye örneği [Financial efficiency analysis of metropolitan municipalities with integrated DEA and GRA: The case of Turkey]. *International Journal of Economic and Administrative Studies, 8*(15), 411-428.

Koopmans, T. C. (1951). Analysis of production as an efficient combination of activities. In T. C. Koopmans (Ed.), *Activity Analysis of Production and Allocation: Proceedings of a Conference* (pp. 33-97). New York, NY: John Wiley & Sons.

Kung, C. Y., & Wen, K. L. (2007). Applying grey relational analysis and grey decision-making to evaluate the relationship between company attributes and its financial performance—A case study of venture capital enterprises in Taiwan. *Decision Support Systems, 43*(3), 842-852.

Kuo, Y., Yang, T., & Huang, G. W. (2008). The use of grey relational analysis in solving multiple attribute decision-making problems. *Computers & Industrial Engineering, 55*(1), 80-93.

Lovell, C. A. K., & Pastor, J. T. (1995). Units invariant and translation invariant DEA models. *Operations Research Letters, 18*(3), 147-151.

Murillo-Zamorano, L. R. (2004). Economic efficiency and frontier techniques. *Journal of Economic Surveys, 18*(1), 33-77.

Ng, D. K. W. (1994). Grey system and grey relational model. *ACM SIGICE Bulletin, 20*(2), 2-9.

Önem, H. B., & Demir, Y. (2015). Mülkiyet yapısının firma performansına etkisi: BİST imalat sektörü üzerine bir uygulama [A survey of ownership structure on the performance of the firms: An application on the production sector at BIST]. *Suleyman Demirel University the Journal of Visionary, 6*(13), 31-43.

Pakkar, M. S. (2017). Hierarchy grey relational analysis using DEA and AHP. *PSU Research Review, 1*(2), 150-163.

Pakkar, M. S. (2018). Fuzzy multi-attribute grey relational analysis Using DEA and AHP. In J. Xu, M. Gen, A. Hajiyev, & F. L. Cooke (Eds.), *Proceedings of the Eleventh International Conference on Management Science and Engineering Management. ICMSEM 2017. Lecture Notes on Multidisciplinary Industrial Engineering* (pp. 695-707). Cham, CH: Springer International Publishing.

Pastor, J. T. (1996). Translation invariance in data envelopment analysis: A generalization. *Annals of Operations Research, 66*(2), 91-102.

Sheth, J. N., & Sisodia, R. S. (2002). Marketing productivity: Issues and analysis. *Journal of Business Research, 55*(5), 349-362.

Tayyar, N., Akcanlı, F., Genç, E., & Erem, I. (2014). BİST'e kayıtlı bilişim ve teknoloji alanında faaliyet gösteren işletmelerin finansal performanslarının analitik hiyerarşi prosesi (AHP) ve gri ilişkisel analiz (GİA) yöntemiyle değerlendirilmesi [Financial performance evaluation of technology companies quoted in BIST with Analytic Hierarchy Process (AHP) and Grey Relational Analysis]. *Journal of Accounting and Finance, Ocak/2014*(61), 19-40.

Tsaur, R. C., Chen, I. F., & Chan, Y. S. (2017). TFT-LCD industry performance analysis and evaluation using GRA and DEA models. *International Journal of Production Research, 55*(15), 4378-4391.

Tzeng, C. J., Lin, Y. H., Yang, Y. K., & Jeng, M. C. (2009). Optimization of turning operations with multiple performance characteristics using the Taguchi method and grey relational analysis. *Journal of Materials Processing Technology, 209*(6), 2753-2759.

Van Horne, J. C., & Wachowicz, J. M. (2008). *Fundamentals of Financial Management* (13th ed.). Essex: Pearson Education.

Wahlen, J. M., Baginski, S. P., & Bradshaw, M. T. (2011) *Financial reporting, financial statement analysis, and valuation: A strategic perspective* (7th ed.). Mason, OH: South Western, Cengage Learning.

Wang, R. T. (2007). Performance evaluation of Taiwan's TFT-LCD industry. *International Journal of Value Chain Management, 1*(4), 372-386.

Wang, S., Ma, Q., & Guan, Z. (2007). Measuring hospital efficiency in China using grey relational analysis and data envelopment analysis. In *Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, 18-20 November 2007, Nanjing, China* (pp. 135-139), IEEE.

Wang, Y. J. (2008). Applying FMCDM to evaluate financial performance of domestic airlines in Taiwan. *Expert Systems with Applications, 34*(3), 1837-1845.

Wang, Y. J. (2014). The evaluation of financial performance for Taiwan container shipping companies by fuzzy TOPSIS. *Applied Soft Computing, 22*, 28-35.

Yu, K., Luo, B. N., Feng, X., & Liu J. (2018). Supply chain information integration, flexibility, and operational performance: An archival search and content analysis. *The International Journal of Logistics Management, 29*(1), 340-364.