



Volume 8

Issue 1

2021

***International Journal of
Assessment Tools in Education***

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

<http://www.ijate.net>

e-ISSN: 2148-7456

© IJATE 2021





e-ISSN 2148-7456

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

Volume 8

Issue 1

2021

Dr. İzzet KARA

Publisher

International Journal of Assessment Tools in Education

&

Pamukkale University,

Education Faculty,

Department of Mathematic and Science Education,

20070, Denizli, Turkey

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : ijate.editor@gmail.com

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

Online ISSN: 2148-7456

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

Design & Graphic: IJATE

Support Contact

Dr. İzzet KARA

Journal Manager & Founding Editor

Phone : +90 258 296 1036

Fax : +90 258 296 1200

E-mail : ikara@pau.edu.tr

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

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



International Journal of Assessment Tools in Education

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

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

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

Starting from this issue, the abbreviation for *International Journal of Assessment Tools in Education* is "*Int. J. Assess. Tools Educ.*" has been changed.

Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- ERIH PLUS,
- EBSCO,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib

Editor

[Dr. Ozen YILDIRIM](#), Pamukkale University, Turkey

Editorial Board

[Dr. Eren Can AYBEK](#), Pamukkale University, Turkey

[Dr. Beyza AKSU DUNYA](#), Bartın University, Turkey

[Dr. Selahattin GELBAL](#), Hacettepe University, Turkey

[Dr. Stanislav AVSEC](#), University of Ljubljana, Slovenia

[Dr. Murat BALKIS](#), Pamukkale University, Turkey

[Dr. Gulsah BASOL](#), Gaziosmanpaşa University, Turkey

[Dr. Bengu BORKAN](#), Boğaziçi University, Turkey

[Dr. Kelly D. BRADLEY](#), University of Kentucky, United States

[Dr. Okan BULUT](#), University of Alberta, Canada

[Dr. Javier Fombona CADAVIECO](#), University of Oviedo, Spain

[Dr. William W. COBERN](#), Western Michigan University, United States

[Dr. R. Nukhet CIKRIKCI](#), İstanbul Aydın University, Turkey

[Dr. Safiye Bilican DEMİR](#), Kocaeli University, Turkey

[Dr. Nuri DOGAN](#), Hacettepe University, Turkey

[Dr. R. Sahin ARSLAN](#), Pamukkale University, Turkey

[Dr. Anne Corinne HUGGINS-MANLEY](#), University of Florida, United States

[Dr. Francisco Andres JIMENEZ](#), Shadow Health, Inc., United States

[Dr. Nicole KAMINSKI-OZTURK](#), The University of Illinois at Chicago, United States

[Dr. Orhan KARAMUSTAFAOGLU](#), Amasya University, Turkey

[Dr. Yasemin KAYA](#), Atatürk University, Turkey

[Dr. Hulya KELECIOGLU](#), Hacettepe University, Turkey

[Dr. Hakan KOGAR](#), Akdeniz University, Turkey

[Dr. Omer KUTLU](#), Ankara University, Turkey

[Dr. Seongyong LEE](#), BNU-HKBU United International College, China

[Dr. Sunbok LEE](#), University of Houston, United States

[Dr. Froilan D. MOBO](#), Ama University, Philippines

[Dr. Hamzeh MORADI](#), Sun Yat-sen University, China

[Dr. Nesrin OZTURK](#), Izmir Democracy University, Turkey

[Dr. Turan PAKER](#), Pamukkale University, Turkey

[Dr. Murat Dogan SAHİN](#), Anadolu University, Turkey

[Dr. Ragıp TERZİ](#), Harran University, Turkey

[Dr. Hakan TURKMEN](#), Ege University, Turkey

[Dr. Hossein SALARIAN](#), University of Tehran, Iran

English Language Editors

[Dr. Hatice ALTUN](#) - Pamukkale University, Turkey

[Dr. Arzu KANAT MUTLUOGLU](#) - Ted University, Turkey

Editorial Assistant

[Anil KANDEMİR](#) - Middle East Technical University, Turkey

TABLE OF CONTENTS

Research Articles

[Analysis of Students Satisfaction with Virtual Education in Medical Science University during the Pandemic Outbreak of COVID-19](#), Pages 1 – 8

Freshteh OSMANİ

[An Evaluation of Pass/Fail Decisions through Norm- and Criterion-Referenced Assessments](#), Pages 9 – 20

Ismail CUHADAR, Selahattin GELBAL

[Comparison of confirmatory factor analysis estimation methods on mixed-format data](#), Pages 21 – 37

Abdullah Faruk KILIÇ, Nuri DOĞAN

[Developing and validating a computerized oral proficiency test of English as a foreign language \(Coptefl\)](#), Pages 38 – 66

Cemre ISLER, Belgin AYDIN

[Examining the Measurement Invariance of TIMSS 2015 Mathematics Liking Scale through Different Methods](#), Pages 67 – 89

Zafer ERTÜRK, Esra OYAR

[Investigation of measurement invariance of PISA 2015 collaborative problem solving skills: Turkey, Norway and Singapore](#), Pages 90 – 105

Yusuf Taner TEKİN, Derya ÇOBANOĞLU AKTAN

[Teachers' Attitudes and Opinions about Design and Skill Workshops and Ranking of Workshops by Teachers](#), Pages 106 – 119

Ayşegül BAYRAKTAR, Seher YALÇIN

[Homework Process in Higher Education Scale \(HPHES\): A Validity and Reliability Study](#), Pages 120 – 134

Veda YAR YILDIRIM

[The Effect of Item Pools of Different Strengths on the Test Results of Computerized-Adaptive Testing](#), Pages 145 – 155

Fatih KEZER

[Defining Cut Point for Kullback-Leibler Divergence to Detect Answer Copying](#), Pages 156 – 166

Arzu UÇAR, Celal DOĞAN

[The Problem of Measurement Equivalence or Invariance in Instruments](#), Pages 167 – 180

Tülin (OTBİÇER) ACAR

Review Articles

[Kirkpatrick Model: Its Limitations as Used in Higher Education Evaluation](#), Pages 135 – 144

Michael CAHAPAY

Analysis of Students Satisfaction with Virtual Education in Medical Science University during the Pandemic Outbreak of COVID-19

Freshteh Osmani ^{1,*}

¹Infectious Disease Research center, Birjand University of Medical Sciences, Birjand, Iran

ARTICLE HISTORY

Received: Oct. 26, 2020

Revised: Dec. 17, 2020

Accepted: Jan. 05, 2020

Keywords:

Distance learning,
Online learning,
Virtual education,
COVID-19,
Students.

Abstract: Nowadays, owing to the failure of the Traditional Educational System, the only option left is the Virtual Educational, which will change the educational system at 180 degrees. The aim of this study was to investigate and evaluate the relationship between different factors associated with the level of satisfaction amongst students of Medical Science University during the pandemic outbreak of COVID-19. This cross-sectional study was performed among students of Birjand University of Medical Sciences in 2020. They completed the questionnaire was created using a Google platform and their answers was collected online. Satisfaction towards virtual educational learning plus total evaluation scores for various dimension of questionnaire was analyzed. A total of 320 out of 2700 students participated in the study voluntarily. Students' satisfaction with blended method in teaching style was higher and significant than two separate styles ($p<0.05$), but there was no significant relationship between satisfaction level and some demographic characteristics. Also, the majority of participants (41.7%) have a medium level of Satisfaction. There was significant relationship between the amount of computer skills, Semester and sex with overall satisfaction ($p<0.05$). Students demonstrated a moderate satisfaction and positive attitude towards VR educational system which comprises of a "Virtual Learning Room" at home for both the teacher and student. To be able to implement education in medical universities in the coronavirus crisis, electronic and internet infrastructures need to be completed quickly, and officials should take steps to empower students and teachers to take advantage of this opportunity.

1. INTRODUCTION

As the coronavirus (COVID-19) pandemic becomes widespread, medical science universities have suspended all classes with the hopes of mitigating viral transmission. The mechanism of spreading the virus is mainly dependent on direct contact and airborne droplets, even from asymptomatic carriers, and the rate of transmission is highly increased in crowded places such

CONTACT: Freshteh Osmani ✉ dr.osmani68@gmail.com 📄 Dentistry Clinical Research Development Center, Birjand University of Medical Sciences, Birjand, Iran

ISSN-e: 2148-7456 /© IJATE 2021

as universities. Accordingly, Organization such as UNESCO[†]- OHCHR[‡]- IFRC[§] and WHO^{**} urged countries to provide well-prepared, adaptable and accessible education settings that would be to all universities following closures during this pandemic(Chiao et al., 2018; Wilbur, 2016).

So, for more than 850 million students worldwide, disrupting the original teaching plans of universities. Soon later, many countries started to offer online teaching to students. While the corona virus quickly circulating in many countries, they had required decisive and drastic steps to avert an overflowing-blown contagion that evaluating: teaching, education content, participants' attitude towards to the course and its difficulty, students' perception and final judgment). Answers were presented on a Likert scale. In the present study, Cronbach's alpha coefficient of the questionnaire was obtained 0.83. This questionnaire was given to 7 experts in the field of medical education and after the final review, the final questionnaire with a validity of 0.78 was approved. As the participants were completed all the questions, the returned electronic forms were saved.

2. METHOD

2.1. Statistical analysis

All statistical methods were performed by using SPSS software version 23. Quantitative data were presented as mean \pm SD and qualitative ones reported as frequency and percent. To analyze, due to normal distribution, Chi-square and Pearson correlation tests were used to identify the relationships in qualitative and quantitative variables, respectively. The statistical significance level was set at $p < .05$.

The present study was approved by the Ethics Committee of Birjand University of Medical Sciences (Ethical codes: IR.BUMS.REC.REC.1399.256).

3. RESULT

Three hundred and twenty students (28.9% male and 71% female) with a mean age of 21.17 ± 1.37 years (ranging from 19 to 29 years) participated in the study.

The percentages of subjects at the different grades were as follows: Bachelor students= 30.8%, Master students = 10.4%, Medical students = 39.1%, and PhD students = 17.2%. The majority of participants had basic (53.2%) and intermediate (47.5%) computer skills and whilst only 7.6% of them lacked any experience (Table 1).

Table 1. Distribution of Demographic Character of participants.

Variable	Mean+SD	N (%)
Age (y)	20.95 ± 1.65	
sex	male	93(28.9)
	female	227 (71)
Computer skills	Basic	170(53.2)
	Intermediate	152(47.5)
	Advanced	24 (7.6)
Grade	Bachelor	99(30.8)
	Master	33(10.4)

[†]-United Nations Educational, Scientific and Cultural Organization

[‡].Office of the United Nations High Commissioner for Human Rights

[§]-International Federation of the Red Cross and Red Crescent Societies-

^{**}-World Health Organization

Table 1. *Continues.*

	Medicine	125(39.1)
	PhD	24(7.6)
Having Practical lesson	Yes	231(72.4)
	No	(88(27.6)
Faculty	Health	55(17.2)
	Paramedical	48(15.1)
	medical	88(27.5)
	Dentistry	74(23.2)
	Nursing and Midwifery	67(21)
Semester	2 th	55(17.2)
	4 th	74(23)
	6 th	68(21.3)
	8 th	53(16.5)
	10 th	43(13.3)
	12 th	20(6.2)
GPA	18-20	151(47.2)
	16-18	124(38.6)
	<16	45(14)

The percentages of subjects at the different grades were as follows: Bachelor students= 30.8%, Master students = 10.4%, Medical students = 39.1%, and PhD students = 17.2%. The majority of participants had basic (53.2%) and intermediate (47.5%) computer skills and whilst only 7.6% of them lacked any experience ([Table 1](#)).

Table 2. *Distribution of students' total satisfaction with virtual education system.*

Total satisfaction level	Frequency (%)
Not at all	41(13.1)
low	61(19.3)
medium	135(42.2)
high	59(18.4)
Very high	22(6.8)

This results show that majority of students (42.2) % had a medium satisfaction level and only 6.8% of them were very satisfied with this management of virtual educational system ([Table 2](#)).

Table 3. *Correlation between satisfaction score and questionnaire dimensions.*

Satisfaction factors	Total satisfaction score	
	<i>r</i>	<i>p</i>
platform availability of system	0.32	0.032
Designed content	0.19	0.09
Interactive learning activities	0.61	<0.001
quality of service	0.29	0.047
Teacher evaluation	0.06	0.28

From the above analysis, we draw the conclusion that among the five major factors, designed content had no direct influence on user total satisfaction. Also, the above results show that, the influence index of user satisfaction mainly involved service quality. Users mainly hoped that the platform could meet their learning needs and provide necessary functions for learning; however, they did not have high expectations for the interface design of the platform (Table 3).

Table 4. Relationship between satisfaction levels with demographic variables.

Variable		low	medium	high	χ^2 (p.value)
Satisfaction level	low				
	medium				
sex	Male	7.8%	12.2%	9%	13.79
	Female	18.9%	38.6%	14%	(0.04)
Computer skills	Basic	24.3	21.6	7.8	10.54
	Intermediate	9.2	27.9	8.5	(0.038)
	Advanced	-	3.4	4.2	
Grade	Bachelor	7.2	20.4	4	4.198
	Master	3.2	4.6	2.7	(0.241)
	Medicine	8.4	19.6	11.3	
	PhD		5.2	2.4	
Practical lesson	Yes	34.7	26	12.4	8.21
	No	8.3	13.5	6.4	(0.32)
Faculty	Health	3.6	10.4	3.8	5.946
	Paramedical	-	11.3	4.7	(0.114)
	medical	5.9	15.8	6.5	
	Dentistry	4.6	12.7	5.1	
Semester	Nursing and Midwifery	1.3	17.1	2.6	
	2 th	3.2	9.5	4.1	3.092
	4 th	2.8	17	3.2	(0.378)
	6 th	1.9	14.8	4.3	
	8 th	3.5	8.6	3.9	
	10 th	0.96	10.1	2.3	
	12 th	0.7	3.9	1.2	
GPA	18-20	9.3	22	16	
	16-18	12.4	18	8.5	9.27
	<16	4.8	7.9	1.3	(0.051)

According to the above analysis, we draw the conclusion that among all considered demographic variables, only characteristic including sex, Computer skills, Semester and GPA had a significantly relationship with satisfaction level ($p < 0.05$) (Table 4)

4. DISCUSSION and CONCLUSION

In the coronavirus disease pandemic, educational system is no exception to undergo frequent occurred. Changing education is especially important. So, the education system needs comprehensive management and monitoring to maintain the best quality. The issuance of health guidelines for the observance of social distances necessitates a change in the educational systems. The crisis of COVID-19 pandemic has challenged the learning system too. In the face of the coronavirus disease crisis, creating a platform for virtual educational will create a new capacity for student education. In most countries, due to social distancing and closure of universities has prompted educators to teach in a virtual way and take online exams (Conroy et

al., 2008; González-Gómez et al., 2012). Education in the coronavirus pandemic entails a change in students' educational needs and educational systems. In order to achieve educational goals in medical sciences universities, measures must be taken to allocate limited resources to educational goals in the best way (Franz et al., 2015; Mahoney et al., 2016).

Thus, this study was designed to evaluate virtual educational performance by Student satisfaction evaluation amongst students in one of the medical sciences universities in Iran.

The results showed a significant difference in satisfaction with virtual educational in different teaching styles, so that blended method had more satisfaction than others, but there was no significant difference in satisfaction with online and offline-content among students.

Previous research has shown that the using online content in teaching as a non-synchronous e-learning tool has an effective role in satisfaction of students and helps them focus on content (Oliveira et al., 2017; Osmani et al., 2019). It is important to note that in a virtual learning environment, many factors including lecturers, courses, technology, system design and learning environments affect user satisfaction (Gholipour Mofrad Dashtaki et al., 2020; Moazami et al., 2014). As an example, the result of a study showed that while content is appropriate, factors such as problematic use, technical problems and lack of access to electronic equipment can be reasons for dissatisfaction with virtual education (Buşan, 2014). Several studies showed a moderate relationship between the strategy of using online content learning techniques and its satisfaction (Gholipour Mofrad Dashtaki et al., 2020). Also, another one has showed that the LMS method is better suited to support efficient learning (Franz et al., 2015). Another study has demonstrated that most students were very satisfied with the effect of using blended teaching methods (Bennett & Lockyer, 2004). In practical and skill-based discussions, it should be noted that training should be both virtual and in-person in order to achieve student satisfaction and optimal performance (Cohen & Davidovitch, 2020). On the other hand, various teaching styles are as notable educational concepts and the number of students and different kinds of educational content should be specified based on the teaching style (Donoghue, 2006; Kim & Bonk, 2006). Another result of this study showed that, designed content had no direct influence on user total satisfaction, indicating that users had a fair attitude. Instead, platform availability had a significance influence on user satisfaction. In terms of availability, the most important problems were: function design and operation of the online teaching platform. In terms of Interactive learning activities quality, the feedback for the homework assigned by teachers was the main effective factor (Viner et al., 2020). The influence of teacher evaluation quality on satisfaction of students was caused by matters such as timely response to problems and learning extension. The correlation between the overall Interactive activity's quality, Teacher evaluation quality, and designed content was not high, indicating that the influence on user satisfaction was not high (Anarinejad & Mohammadi, 2020).

In the current study, distribution of satisfaction level was almost similar in different faculty, so that, there was no significant difference between them. But in total, satisfaction in students of health faculty was more than others. It can be due to majority of students from this faculty are in bachelor grade without practical and clinical courses. A study indicated that participating in a virtual education course can improve attending students' attitudes towards virtual education in students with different learning styles (Ebadi & Heidaranlu, 2020; Setiawan, 2020).

In the present study, there was no difference in satisfaction score in the using of virtual education between different grades. Similarly, more participants in the current study were not satisfied with their VR experiences. Satisfaction and positive attitude towards VR education seem to be positively associated with gender, computer skills and previous experiences on VR. These results were consistent with a previous study (Zaharah et al., 2020).

In the present study, the majority were female and both genders were in medium level of satisfaction. In terms of gender, females showed higher satisfaction and a more positive attitude

than males towards virtual education, which is similar to other studies that females are more willing to virtual education. However, other studies have rejected the role of gender in relation to satisfaction with VR education. These differences may be due to difference in the assessment methods and even sample size (Akram et al., 2020; Cohen & Davidovitch, 2020).

Further, the satisfaction of 6th-semester students towards VR education and technical experience were significantly more than amongst their lower semester.

Lack of technical skills adversely affect the virtual educational learning process. Another study has shown that students' perception of virtual learning was related to the degree of essential computer skills and stable access to the Internet (Gholipour Mofrad Dashtaki et al., 2020). These results were in line with the result of our study. Computer skills potentially strengthen the connection between students and learning from VR education. Additionally, it enhances their proficiency in using various platforms and applications, most of the students considered social media to be unhelpful in the VR educational learning process, although it can be an effective tool in this era. Hence, the curriculum should be well-structured to ensure the effectiveness of these tools (Al-Taweel et al., 2020).

Previous studies have been reported conflicting results regarding satisfaction with VR educational learning, meanwhile some studies indicated a higher attrition rate for online courses than conventional one. This can be ascribed to overlapping in the timing of online lectures with personal daily activities (Chen et al., 2020). Generally, students' satisfaction with virtual programmers are influenced by multiple factors, like, quality of the course. Support of this concept was seen by students' need to improve the quality of the online lectures. Moreover, majority of them agreed with combining VR educational learning with classic classroom as the best method to attain the targets of the educational process. This result agrees with a previous published systematic review (Tanveer et al., 2020).

Evaluation of mean attitude of females towards VR education was significantly higher than male, which is consistent with results from a previous study which suggested that females are more teacher-oriented than males. Also, design and aesthetic presentation of the online material, and live interaction with the students are factors that positively affect the success of the online courses (Kim & Bonk, 2006; Osmani et al., 2018). Indeed, attitude towards virtual lecturers and it's quality are main factors in success of virtual education learning (Anarinejad & Mohammadi, 2020; Osmani & Hajizadeh, 2019).

This study, like other observational studies has limitations such as that only relationship with certain variables could be specified but not the cause-effect associations. Also, this study focused on evaluating user satisfaction for university students only. Therefore, the evaluation of other aspects of distant learning, such as interactive tutorials, webinars and online courses should be considered in the future studies.

Furthermore, the results of this study were compared to previous studies published in normal situations rather than a crisis period, which could have some bias, especially due to depression and anxiety related with the social restrictions.

This study collected students' experience data on virtual education platforms during the COVID-19 pandemic. Generally, the obtained results showed that majority of the participants feels that they are being adversely affected by VR education learning system in university. Also, a moderate level of satisfaction and positive attitude towards VR educational learning was observed. Although, COVID-19-associated events have caused to improve IT and computer skills amongst all users' members to prepare better for similar crises in the future.

We concluded that, if the situation prevails, drastically affect the teaching during fall, and even on the recurring semester 2021. It will be highly challenging to continue in-person class

sessions for both students and teachers. However, online teaching will also cost a lot in the shape of the internet and related facilities costs.

The next semester activities halted. No face to face interaction and universities closed for the time being. Shift to online classes with the help of virtual educational system. Assignments and open book exams may be used as an option. It would be highly challenging for teachers and students to shape back in the position of face to face learning. However, cost on internet use and related facilities shows an upward trajectory.

At last, we can use a strategic planning tool to meet the current challenges and to cope with any uncertain and risk situation in the future. Also, it is necessary to establish a dedicated center at the university that will work to cooperate internally and externally to decrease the line and virtual barriers.

Acknowledgments

We would like to thank all students participating in the study.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Freshteh Osmani: Investigation, Methodology, Visualization, Software, Formal Analysis, and Writing the original draft.

ORCID

Freshteh Osmani  <https://orcid.org/0000-0002-6112-7131>

5. REFERENCES

- Akram, W., Adeel, S., Tabassum, M., Jiang, Y., Chandio, A., & Yasmin, I. (2020). Scenario Analysis and Proposed Plan for Pakistan Universities–COVID–19: Application of Design Thinking Model.
- Al-Taweel, F. B., Abdulkareem, A. A., Gul, S. S., & Alshami, M. L. (2020). Evaluation of technology-based learning by dental students during the pandemic outbreak of coronavirus disease 2019. *European Journal of Dental Education*, 24(3), 81-89.
- Anarinejad, A., & Mohammadi, M. (2020). The practical indicators for evaluation of e-learning in higher education in Iran. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 5(1), 11-25.
- Bennett, S., & Lockyer, L. (2004). Becoming an online teacher: Adapting to a changed environment for teaching and learning in higher education. *Educational Media International*, 41(3), 231-248.
- Buşan, A.-M. (2014). Learning styles of medical students-implications in education. *Current health sciences journal*, 40(2), 104.
- Chen, T., Peng, L., Yin, X., Rong, J., Yang, J., & Cong, G. (2020). *Analysis of user satisfaction with online education platforms in China during the COVID-19 pandemic*. Paper presented at the Healthcare.
- Chiao, H.-M., Chen, Y.-L., & Huang, W.-H. (2018). Examining the usability of an online virtual tour-guiding platform for cultural tourism education. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 23, 29-38.
- Cohen, E., & Davidovitch, N. (2020). The Development of Online Learning in Israeli Higher Education. *Journal of Education and Learning*, 9(5), 15.

- Conroy, M., Durrheim, D. N., & Dalton, C. (2008). Likely impact of school and childcare closures on public health workforce during an influenza pandemic: a survey. *Communicable Diseases Intelligence Quarterly Report*, 32(2), 261.
- Donoghue, S. L. (2006). Institutional potential for online learning: A Hong Kong case study. *Journal of Educational Technology & Society*, 9(4), 78-94.
- Ebadi, A., & Heidarlanlu, E. (2020). Virtual Learning: A New Experience in the Shadow of Coronavirus Disease. *Shiraz E-Medical Journal*, 21(12), e106712.
- Franz, S., Behrends, M., Haack, C., & Marschollek, M. (2015). *Benefits and Barriers of E-Learning for Staff Training in a Medical University*. Paper presented at the ICIMTH.
- Gholipour Mofrad Dashtaki, D., Mohammadi, A., Zolfaghari, M., Imani, S., & Tahmasebian, S. (2020). The Relationship of Satisfaction and Usage of Virtual Learning Facilities with Learning Style in Medical, Health, and Operating Room Students. *Strides in Development of Medical Education*, 17(1), 1-6.
- González-Gómez, F., Guardiola, J., Rodríguez, Ó. M., & Alonso, M. Á. M. (2012). Gender differences in e-learning satisfaction. *Computers & Education*, 58(1), 283-290.
- Kim, K.-J., & Bonk, C. J. (2006). The future of online teaching and learning in higher education. *Educause quarterly*, 29(4), 22-30.
- Mahoney, N. R., Boland, M. V., Ramulu, P. Y., & Srikumaran, D. (2016). Implementing an electronic learning management system for an Ophthalmology residency program. *BMC medical education*, 16(1), 307.
- Moazami, F., Bahrapour, E., Azar, M. R., Jahedi, F., & Moattari, M. (2014). Comparing two methods of education (virtual versus traditional) on learning of Iranian dental students: a post-test only design study. *BMC medical education*, 14(1), 1-5.
- Oliveira, A. C., Mattos, S., & Coimbra, M. (2017). Development and assessment of an e-learning course on pediatric cardiology basics. *JMIR medical education*, 3(1), e10.
- Osmani, F., & Hajizadeh, E. (2019). Combining Multiple Imputation and Inverse-Probability Weighting for Analyzing Response with Missing in the Presence of Covariates. *Journal of Biostatistics and Epidemiology*.
- Osmani, F., Hajizadeh, E., Rasekhi, A., & Akbari, M. E. (2018). Analyzing Relationship Between Local and Metastasis Relapses with Survival of Patients with Breast Cancer: A Study Using Joint Frailty Model. *International Journal of Cancer Management*, 11(12), e81783.
- Osmani, F., Hajizadeh, E., Rasekhi, A., & Akbari, M. E. (2019). Prognostic factors associated with locoronal relapses, metastatic relapses, and death among women with breast cancer. Population-based cohort study. *The Breast*, 48, 82-88.
- Setiawan, A. R. (2020). Scientific Literacy Worksheets for Distance Learning in the Topic of Coronavirus 2019 (COVID-19).
- Tanveer, M., Bhaumik, A., & Hassan, S. (2020). Covid-19 Pandemic, Outbreak Educational Sector and Students Online Learning in Saudi Arabia. *Journal of Entrepreneurship Education*, 23(3).
- Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., . . . Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The Lancet Child & Adolescent Health*.
- Wilbur, K. (2016). Evaluating the online platform of a blended-learning pharmacist continuing education degree program. *Medical education online*, 21(1), 31832.
- Zaharah, Z., Kirilova, G. I., & Windarti, A. (2020). Impact of Corona Virus Outbreak Towards Teaching and Learning Activities in Indonesia. *SALAM: Jurnal Sosial dan Budaya Syar-i*, 7(3), 269-282.

An Evaluation of Pass/Fail Decisions through Norm- and Criterion-Referenced Assessments

Ismail Cuhadar^{1,*}, Selahattin Gelbal²

¹Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara, Turkey

²Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: June 02, 2020

Revised: Dec. 06, 2020

Accepted: Jan. 04, 2021

KEYWORDS

Norm-referenced
Assessment,

Criterion-referenced
Assessment,

Standard-setting,

Angoff Method,

Nedelsky Method.

Abstract: The institutions in education use various assessment methods to decide on the proficiency levels of students in a particular construct. This study investigated whether the decisions differed based on the type of assessment: norm- and criterion-referenced assessment. An achievement test with 20 multiple-choice items was administered to 107 students in guidance and psychological counseling department to assess their mastery in the course of measurement and evaluation. First, the raw scores were transformed into T-scores for the decisions from norm-referenced assessments. Two decisions were made to classify students as passed/failed comparing each student's T-score with two common cutoffs in education: 50 and 60. Second, two standard-setting methods (i.e., Angoff and Nedelsky) were conducted to get two cut scores for the criterion-referenced assessment with the help of experts in measurement and evaluation. Two more decisions were made on the classification of students into pass/fail group by comparing the raw scores and the cut scores from two standard-setting methods. The proportions of students in pass/fail categories were found to be statistically different across each pair of four decisions from norm- and criterion-referenced assessments. Cohen's Kappa showed that the decisions based on Nedelsky method indicated a moderate agreement with the pass/fail decisions from the students' semester scores in measurement and evaluation while the other three decisions showed a lower agreement. Therefore, the criterion-referenced assessment with Nedelsky method might be considered in making pass/fail decisions in education due to its criterion validity from the agreement with the semester scores.

1. INTRODUCTION

Educational institutions make high-stakes decisions about students to determine who has mastered the objectives of a course, who will be promoted to the upper grades or who will be selected to a particular school. Because a false decision can cause some problems in reaching the next level objectives in education, educational institutions should be careful in making these decisions. For example, students' learning should be assessed carefully in secondary school due to the fact that it may have impact on the dropout rates in higher education (Paura & Arhipova, 2014; Vossensteyn et al., 2015). Therefore, it is important to build appropriate decision

CONTACT: Ismail CUHADAR ✉ ismail.cuhadar@gmail.com 📍 Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

mechanism to minimize these false decisions. A valid assessment method is one required tool for the appropriateness of the decision-making mechanism in education (Aiken, 2000; Crocker & Algina, 2008).

Assessment is a decision-making process that involves the comparison of the measurement results with a criterion (Baykul, 2010; Turgut & Baykul, 2010). In other words, assessment is composed of three items: measurement results, criterion, and decision. Accordingly, a decision is made on the measured construct by comparing the measurement results with a criterion. For this reason, the criterion is required to be defined in order to evaluate the measurement results. Based on the type of the criterion in the assessment process, assessment is grouped in two categories: norm-referenced assessment and criterion-referenced assessment (Murphy & Davidshofer, 2005).

Norm-referenced assessment is a formal evaluation process where the performance of a student is compared with the performance of a scientifically selected group of test takers (Crocker & Algina, 2008; Kline, 2000). In the norm-referenced assessment, the performance of the selected group of test takers is the criterion for making the decision on the measurement results of students. On the other hand, a student's score is compared with a predetermined level of performance in the criterion-referenced assessment, and this particular level of performance can be determined using a specific ability set or knowledge area in an educational assessment (Cizek & Bunch, 2007; Murphy & Davidshofer, 2005). In the criterion-referenced assessment, one approach for establishing the criterion is to determine a cut score through the standard setting methods based on the subject matter experts' judgements (Cizek & Bunch, 2007; Crocker & Algina, 2008; Murphy & Davidshofer, 2005; Urbina, 2004).

Standard setting is a process to identify a number that separates different performance levels (Cizek, 1993). There are several standard setting methods in the literature. Jaeger (1989) categorized the standard setting methods in two groups: test-centered and examinee-centered methods. Nedelsky method (Nedelsky, 1954), Angoff method (Angoff, 1971), Ebel method (Ebel, 1972), and Yes/No method (Impara & Plake, 1997) are some examples of the test-centered standard setting methods while borderline group method (Zieky & Livingston, 1977) and contrasting groups method (Berk, 1976) are some examples of examinee-centered standard setting methods. Due to the convenience for the multiple-choice test items, the easiness to administer to the subject matter experts, and the popularity in literature and practice, Angoff and Nedelsky methods were used in the current study to identify the cut scores for classifying student into the performance levels (Cizek, 1993; Murphy & Davidshofer, 2005). These two methods were the basis of the criterion-referenced assessment, and briefly introduced in the next sections.

1.1. Angoff Method

Because Angoff method (1971) is a convenient procedure for the tests with the multiple-choice items, it is commonly used in the practice, including license and certificate programs (Cizek & Bunch, 2007). The first step in the application of Angoff method is to define the minimum qualification level for being categorized in the particular performance level with respect to the test purpose (Livingston & Zieky, 1989). Then, the subject matter experts determine the probability that each item can be answered correctly by examinees with this minimum qualification level. An average probability is obtained across all subject matter experts for each item in the test. The sum of these probabilities from each item corresponds to the cut score for the test. This process can be expressed as a formula using Equation 1 to obtain the cut score of a test via Angoff method.

$$Cut\ Score_{Angoff} = \frac{\sum_{j=1}^R \sum_{i=1}^K (p_{ij})}{R} \quad (1)$$

In Equation 1; R indicates the number of subject matter experts, K indicates the number of items in a test, and p_{ij} is the probability determined by expert j to item i .

1.2. Nedelsky Method

Because it is easy to apply the method proposed by Nedelsky in 1954, Nedelsky method is still in practice, and it was one of the methods that accelerated the transition from the norm-based performance level decisions to the assessment type showing examinees' true performance levels (Cizek & Bunch, 2007). The number of distractors in each item is important in determining the cut score using this method (Arrasmith, 1986). The subject matter experts determine how many distractors the minimum qualified examinees can eliminate in each test item taking the measurement construct into consideration. Accordingly, the number of options that the minimum qualified examinees cannot eliminate is determined for each test item. A probability of correct response is obtained considering the number of remaining options. The probabilities across all items are summed to get a cut score for each subject matter expert. The average of these cut scores across all experts indicates the cut score of the test by Nedelsky method. This process can be expressed as a formula using Equation 2 to obtain the cut score of a test via Nedelsky method.

$$Cut\ Score_{Nedelsky} = \frac{\sum_{j=1}^R \sum_{i=1}^K (d_i - e_{ij})^{-1}}{R} \quad (2)$$

In Equation 2; R indicates the number of subject matter experts, K indicates the number of items in a test, d_i is the number of options in item i , and e_{ij} is the number of distractors eliminated in item i by expert j .

1.3. Which Assessment Type?

The type of assessment depends on how measurement results are intended to be used. When the measurement results are used for the selection and placement purpose, the norm-referenced assessment is advantageous over the criterion-referenced assessment (McCauley & Swisher, 1984). However, the decisions from the norm-referenced assessment do not correspond to the true ability level in the target construct of the measurement tool (Johnson & Martin, 1980). For this reason, the norm-referenced assessment is open to misuse in evaluating examinees' performance levels and the effectiveness of a program (McCauley & Swisher, 1984). On the other hand, the criterion-referenced assessment is very useful in determining the examinees' performance levels and replanning curriculum based on the identified needs of the examinees from the criterion-referenced assessment (Freeman & Miller, 2001). Accordingly, the type of assessment that needs to be used in making decisions depends on the purpose of a measurement.

The goal of education is to provide intentional and sustainable changes in students' behavior through a curriculum and based on the objectives of that educational institution (Ertürk, 1998; Tyler, 2013). When the assessment types are reviewed in education and practice, it is seen that different approaches are taken for the similar educational goals. For example, the criterion-referenced assessment is used in the primary education and secondary education while the assessment type differs across the universities in the higher education, although the aforementioned goal of the education is the similar across all levels in the education system (e.g., the minimum scores for being evaluated as successful are 45 and 50 out of 100 in the primary and secondary education in Turkey, respectively; Milli Eğitim Bakanlığı, 2014, 2016). The assessment type is not consistent within the university among the departments, and either the norm- or criterion-referenced assessment can be chosen for evaluating student achievement in some universities (e.g., Akdeniz University, 2017; Ankara University, 2018; Erciyes University, 2015; Sakarya University, 2019). Furthermore, the passing grade is not consistent across the universities (e.g., 50 in Sakarya University, 60 in Ankara University). Thus, a score of 55 is considered insufficient to pass a course in some universities, but the same score means

a sufficient score in the others. In other words, the same score can result in pass or fail decision based on the assessment procedure in the educational institutions. Accordingly, it is important to determine which assessment procedure provides more valid decisions for which situations. Otherwise, the pass/fail decisions can be incorrect or inappropriate, and the incorrect decisions can cause problems in reaching the next level objectives of the curriculum (e.g., Paura & Arhipova, 2014; Vossensteyn et al., 2015).

There are studies in the literature for comparing the norm- and criterion-referenced assessments over different tests (e.g., Mohr, 2006; Oescher, Kirby, & Paradise, 1992; Pester, 2003; Visintainer, 2002). In addition, a few studies investigated the differences and the similarities in the decisions from these two assessment types (e.g., Jacobson, 2008; Nartgün, 2007; Toprakçı, Baydemir, Koçak, & Akkuş, 2007). However, the standard setting methods used in the criterion-referenced assessments were not compared with the norm-referenced assessments in these studies. Furthermore, two assessment types have not been investigated using the same test for decision making. Accordingly, this study purports to compare the decisions on the same group of examinees from the same test with two different assessment procedures: the norm-referenced assessment and the standard setting-based criterion-referenced assessment.

It is not only important to test the differences in the pass/fail decisions from the norm- and criterion-referenced assessments, but also to investigate which assessment type produces more valid decisions under which conditions. The criterion validity might be used to investigate the validity of decisions from the norm- and criterion-referenced assessments (see Aiken, 2000; Baykul, 2010; Kline, 2000; Montgomery & Connolly, 1987; Murphy & Davidshofer, 2005; Turgut & Baykul, 2010 for more information about validity). Despite several studies comparing the two assessment types (e.g., Jacobson, 2008; Nartgün, 2007), the criterion validity of the decisions based on two assessment types has not yet been investigated. Therefore, another purpose of this study is to investigate the criterion validity of the decisions from the norm- and criterion-referenced assessments. Based on two purposes of the study, two research questions were tested: a) “Is there a significant difference between the student-passing rates from the norm- and criterion-referenced assessments?”, and b) “How is the criterion validity of the decisions from the norm- and criterion-referenced assessments?”.

2. METHOD

2.1. Participants

Because the purpose was to compare the assessment types, and the findings were not generalized to a population, there was no sampling procedure in the current study. Accordingly, a purposive study group was chosen that fits the goal of the study. The study group was composed of the second-grade students studying in the guidance and psychological counseling department of Kayseri Erciyes University in Turkey. The fact that these students took a measurement and evaluation course, and the achievement test was designed to measure this content area were the reasons for the selection of them in the current study. In addition, some experts participated in the study for the application of the standard setting methods. These experts had at least a master’s degree in the measurement and evaluation field. In total, there were 107 students from the guidance and psychological counseling department, and there were 11 and 10 experts for the application of Angoff and Nedelsky methods, respectively.

2.2. Procedure and Instrument

In the study, the data were collected in three steps. First, a test was administered to the guidance and psychological counseling students to measure their achievements in the measurement and evaluation course. For this reason, an achievement test with the multiple-choice items was constructed considering the content of the measurement and evaluation course. After the items

were reviewed by two experts in the measurement and evaluation field, a test form with 36 items was obtained. A pilot study was conducted to investigate the statistical characteristics of the items. Then, a final test form composed of 20 items with five options in each was obtained considering the test content, item difficulties, and item discriminations from the pilot study. The data from the final test administration showed that the item discrimination indices ranged between 0.33 and 0.80, and the item difficulty indices ranged between 0.11 and 0.85. In addition, the internal consistency reliability based on Kuder-Richardson formula 20 (KR-20) was equal to 0.71. The second step of the data collection process involved using the experts' opinions to calculate the cut scores based on Angoff and Nedelsky methods. Angoff method was the first application for obtaining the cut scores, and Nedelsky method was administered one week after the application of Angoff method. At the last step of the data collection process, the pass/fail decisions for the students from the measurement and evaluation course at the end of the semester was gathered so that these decisions can be used as a criterion to examine the validity of the norm- and criterion-referenced assessment procedures in the current study.

2.3. Data Analysis

All data analyses were conducted using SPSS 18 (SPSS Inc., 2009) and Microsoft Excel (2013). First, Kendall's coefficient of concordance (i.e., Kendall's W; Kendall & Smith, 1939) was used to examine the agreement among the experts in the standard setting methods. Then, two decisions for each student on their achievements were made as "pass" or "fail" comparing their raw scores with the cut scores from Angoff and Nedelsky methods. In this way, two decisions based on the criterion-referenced assessment were obtained: one from Angoff method and one from Nedelsky method. For the norm-referenced assessment, the raw scores (i.e., the number of correct responses) were first transformed into T-scores (see Sönmez & Alacapınar, 2011). Two more decisions were made on the classification of students into pass and fail categories comparing the T-scores with two passing scores: 50 and 60. These two passing scores were chosen since they are commonly used in the assessment of the students' achievements (e.g., Ankara University, 2018; Sakarya University, 2019). At the end of whole process, there were four decisions for each student on their classifications into passing/failing groups: two decisions from the norm-referenced assessments (i.e., when 50 and 60 were the passing scores in T-score scale) and two decisions from the criterion-referenced assessments (i.e., when two cut scores from Angoff and Nedelsky methods were applied in the raw-score scale).

For the first research question, z-test was used to test whether the decisions based on the four methods in the study statistically differ. Z-test is used to analyze the statistical difference between two proportions from the same group of examinees (Calmorin & Calmorin, 2007). Z-statistic is calculated through dividing the observed proportion difference between the variables by its standard error, as seen in Equation 3 (Jekel, 2007).

Z-test Proportions

		Method II		
		Pass	Fail	
Method I	Pass	a	b	p_1
	Fail	c	d	q_1
		p_2	q_2	1.00

$$z = \frac{p_1 - p_2}{\sqrt{\frac{b+c}{N}}} \tag{3}$$

In Equation 3, N indicates the number of examinees; a indicates the proportion of examinees who pass from both methods; b indicates the proportion of examinees who pass from Method I, but fail from Method II; c indicates the proportion of examinees who pass from Method II, but fail from Method I; d indicates the proportion of examinees who fail from both methods;

$$p_1 = a + b; q_1 = 1 - p_1; p_2 = a + c; \text{ and } q_2 = 1 - p_2.$$

For the second research question, Cohen's Kappa (1960) was used to investigate the criterion validity of the decisions based on the norm- and criterion-referenced assessments by determining the agreement between the pass/fail decisions from the four methods and the pass/fail decisions from the students' semester scores. Cohen's Kappa statistic is used to determine the level of agreement between two categorical variables correcting the agreement rates by chance (Clark-Carter, 2005). The level of agreement based on Cohen's Kappa can be considered as poor for the values < 0.2 ; fair between .2 and 0.4; moderate between 0.4 and 0.6; good between 0.6 and 0.8; and perfect between 0.8 and 1 (Cohen, 1960; Landis & Koch, 1977; McHugh, 2012). Cohen's Kappa can be calculated using Equation 4 (Cohen, 1960).

$$\kappa = \frac{\text{Sum } f_o - \text{Sum } f_e}{N - \text{Sum } f_e} \quad (4)$$

In Equation 4; κ is Cohen's Kappa coefficient, $\text{Sum } f_o$ indicates the sum of observed frequencies in agreement between the methods, $\text{Sum } f_e$ indicates the sum of expected frequencies in agreement between the methods, and N is the number of examinees.

3. RESULT / FINDINGS

Before determining the cut scores from Angoff and Nedelsky methods, the agreement among the experts was examined. Kendall's W indicated a statistically significant agreement among 11 experts in the application of Angoff method ($W = 0.45, p < 0.01$). Similarly, a statistically significant agreement among 10 experts in the application of Nedelsky method was found ($W = 0.44, p < 0.01$). The cut scores across the experts ranged between 9.90 and 17.15 with an average of 13.20 in Angoff method, and between 5.28 and 15.58 with an average of 8.52 in Nedelsky method. Accordingly, the final cut scores of the achievement test was 13.20 and 8.52 based on Angoff and Nedelsky methods through the criterion-referenced assessments, respectively.

The results of the four methods for determining the passing and failing students from the achievement test in the current study was presented in Table 1. When the norm-referenced assessment was used with a cut score of 50 and 60 in T-score scale, 47% and 16% of the students passed the achievement test, respectively. For the criterion-referenced assessments, 23% of the students passed the test from Angoff method while it was 78% when the Nedelsky method was used to determine the cut score of the test. Accordingly, the minimum percent of passing students was from the norm-referenced assessment with a cut score of 60 in T-score scale, and the maximum percent of passing students was from the criterion-referenced assessment with Nedelsky method being the standard setting method. These two methods produced 62% gap with respect to the students classified as pass from the achievement test.

Table 1. *The Cut Scores, The Number of Passing Students, and The Proportion of Passing Students across The Four Methods (n = 107)*

Method	Assessment	Cut Score	Number of Passing	Proportion of Passing (%)
Angoff	Criterion-referenced	13.20	25	23
Nedelsky	Criterion-referenced	8.52	83	78
T-score	Norm-referenced	50.00	50	47
T-score	Norm-referenced	60.00	17	16

Z-test indicated that the proportion of passing students differed statistically among each pair of the four methods in the study at $\alpha = 0.01$ (i.e., $z = 2.83$ for the proportion difference between Angoff method and T-score of 60; $z = -5.00$ for the proportion difference between Angoff method and T-score of 50; $z = 8.12$ for the proportion difference between Nedelsky method and T-score of 60; $z = 5.74$ for the proportion difference between Nedelsky method and T-score of 50; $z = -5.74$ for the proportion difference between T-score of 50 and T-score of 60; $z = 7.62$ for the proportion difference between Angoff and Nedelsky methods). Accordingly, the passing rates depends on the chosen method, and a student can be classified into pass or fail category based on which method is applied in the assessment procedure. Therefore, it is important to determine which method produces more valid decisions among the four methods in the study.

The agreement between the four assessment procedures and the students' semester scores in classifying the students into pass/fail categories was presented in Table 2. When the norm-referenced assessment was used with a rule of 60 to pass the test, 36% of the pass/fail decisions was in agreement with the decision from the students' semester scores. When the passing score was 50 rather than 60 in the norm-referenced assessment, the level of agreement with the semester decisions went up to 61%. The former rule produced a poor agreement ($\kappa = 0.09$), and the agreement was fair from the later rule ($\kappa = 0.24$) in the norm-referenced assessments when the agreement was corrected by chance. For the criterion-referenced assessments, the percent agreement between the Angoff method and the external criterion was equal to 43% with a poor agreement based on Kappa value of 0.14. Among the four methods in the study, Nedelsky method produced the decisions with the highest agreement with the pass/fail categories from the students' semester scores. Nedelsky method and the semester scores resulted in classifying 81% of the students into the same category. In addition, Cohen's Kappa indicated a moderate agreement ($\kappa = 0.41$) between these two ways to categorize students into passing and failing groups. As a result, it was found that Nedelsky method, which is the procedure under the criterion-referenced assessment, provided the best decisions in classifying students into pass/fail categories with respect to the criterion validity.

Table 2. *The Agreement between the Pass/Fail Decisions from the Four Assessment Procedures and the Semester Scores (n=107)*

Method	Assessment	Cut Score	Frequency of Agreement	Percent of Agreement	Kappa
Angoff	Criterion-referenced	13.20	46	43	0.14
Nedelsky	Criterion-referenced	8.52	86	81	0.41
T-score	Norm-referenced	50.00	65	61	0.24
T-score	Norm-referenced	60.00	38	36	0.09

4. DISCUSSION and CONCLUSION

Norm- and criterion-referenced assessments are two major procedures in the assessment of student skills in education. Although which assessment type needs to be used depends on their advantages and disadvantages for different measurement situations, the two assessment types are sometimes used for the same measurement goals. Accordingly, the current study investigated the differences in the pass/fail decisions from the norm- and criterion-referenced assessments, and the criterion validity of the two assessment procedures. Under the norm-referenced assessments, two decisions were made on the classification of students into passing/failing groups using two cut scores: 50 and 60 in the T-score scale. Angoff and Nedelsky methods were used to determine two more cut scores in the raw score metric for categorizing students into passing/failing groups by the criterion-referenced assessment.

The findings indicated that all four methods produced statistically different rates of passing students from the achievement test. Accordingly, a different percent of students might pass a test depending on the type of assessment: norm- or criterion-referenced assessment. This difference between the norm- and criterion-referenced assessments is in line with the findings from Nartgün (2007), but inconsistent with the results in Oescher et al. (1992). The difference might have resulted from using a different subject area with two different tests (one test based on norm-referenced assessment and another test based on criterion-referenced assessment) in the study by Oescher et al. (1992). Nedelsky method provided the highest passing rate (i.e., 78%) among the four methods in the study, which is in line with the findings in Çetin and Gelbal (2010). The lowest percent of passing (16%) resulted from the norm-referenced assessment with a cut score of 60 in T-score scale. This result is understandable since the cut score is one standard deviation above the mean in T-score scale. Therefore, approximately 84% of students are expected to have a lower score than the cutoff in this norm-referenced procedure.

The analyses for investigating the criterion validity of the four methods in the norm- and criterion-referenced assessments indicated different agreement rates between the four methods and the external criterion (i.e., the pass/fail decisions from the students' semester scores). Nedelsky method provided the most valid decisions on the students' achievement groups with respect to the external criterion considering the agreement rates and Cohen's Kappa values. The reason for the consistency between Nedelsky method and the semester scores might be the high success of the students from the exams and projects in the measurement and evaluation course, and the relatively low cut score from Nedelsky method for the achievement test in the current study. However, unlike Nedelsky method, other three methods (i.e., Angoff method, T-score of 50, and T-score of 60) used a harder cut score to pass from the test, and so more students failed from these three assessment procedures causing poor to fair agreement rates with the decisions from the semester scores. This finding is not in line with the results in Jacobson (2008), where both norm- and criterion-referenced assessments were good at classifying examinees into two performance levels. Jacobson (2008) investigated the two assessment procedures in a different subject area and used one test per assessment. The difference in the findings might be attributed to the number of tests and the content of the tests in the studies.

Because the percentage of students classified in the passing performance level depended on the type of assessment in the current study, it is recommended that the educational institutions determine the assessment procedure based on their assessment purpose. When the purpose of the assessment is to determine the performance level of examinees or the proficiency in a construct, the criterion-referenced assessment is recommended. Accordingly, Nedelsky method can be used in determining how much is enough to pass from a course or curriculum considering the criterion validity of the method in the current study. However, the limitations of the current study need to be taken into consideration before generalizing the results into the other settings. For example, the course of measurement and evaluation was the subject area for the

achievement test in the current study. It is also possible to study the same research questions in other subject areas (e.g., comparing if the results differ across the subject areas requiring verbal skills or numerical skills). In addition, Angoff and Nedelsky methods were chosen for the criterion-referenced assessment in the current study, but some other standard-setting methods (e.g., borderline group method, contrasting groups method, etc.) can be considered in a future study. Furthermore, the number of performance levels was two in the current study. More than two performance levels can be studied in a future work (e.g., the number of letter grades in universities).

Acknowledgements

This study is composed of some parts of the master's thesis entitled as “*A Study upon Comparison of Norm- and Criterion-referenced Assessments*”.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Ismail Cuhadar: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Selahattin Gelbal:** Methodology, Supervision and Validation.

ORCID

Ismail CUHADAR  <https://orcid.org/0000-0002-5262-5892>

Selahattin GELBAL  <https://orcid.org/0000-0001-5181-7262>

5. REFERENCES

- Aiken, L. R. (2000). *Psychological testing and assessment*. Boston: Allyn and Bacon.
- Akdeniz University. (2017, September 17). *Akdeniz Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim ve Sınav Yönetmeliği [Akdeniz University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from <http://oidb.akdeniz.edu.tr/wp-content/uploads/2017/02/Akdeniz-Üniversitesi-Ön-Lisans-ve-Lisans-Eğitim-Öğretim-ve-Sınav-Yönetmeliği-17.09.2017.pdf>
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.
- Ankara University. (2018, September 4). *Ankara Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim Yönetmeliği [Ankara University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from <http://oidb.ankara.edu.tr/files/2018/04/ÖN-LİSANS-VE-LİSANS-EĞİTİM-ÖĞRETİM-YÖNETMELİĞİ.pdf>
- Arrasmith, D. G. (1986). *Investigation of judges' errors in Angoff and contrasting groups cut of score methods* [Doctoral dissertation, University of Massachusetts]. ProQuest Dissertations and Theses.
- Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması [Measurement in Education and Psychology: Classical test theory and applications]*. Ankara: Pegem Yayıncılık.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45(2), 4-9.
- Calmorin, L. P., & Calmorin, M. A. (2007). *Research methods and thesis writing*. Manila: Rex Book Store.

- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93-106.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousands Oaks, CA: Sage Publications.
- Clark-Carter, D. (2005). *Quantitative psychological research: a student handbook*. New York, NY: Psychology Press.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Erciyes University. (2015, December 27). *Erciyes Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim Yönetmeliği [Erciyes University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from <https://www.erciyes.edu.tr/kategori/ERU-E-BELGE/Yonetmelikler/131/136>
- Ertürk, S. (1998). *Eğitimde program geliştirme [Program development in education]*. Ankara: Meteksan.
- Freeman, L., & Miller, A. (2001). Norm-referenced, criterion-referenced, and dynamic assessment: What exactly is the point? *Educational Psychology in Practice*, 17(1), 3-16.
- Çetin, S., & Gelbal, S. (2010). Impact of standard setting methodologies over passing scores. *Ankara University, Journal of Faculty of Educational Sciences*, 43(1), 79–95.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34(4), 353-366.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed.; pp. 485-514). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Jacobson, R. Y. (2008). *Examination of the potential of selected norm-referenced tests and selected locally developed criterion-referenced tests to classify students into performance categories* [Doctoral dissertation, University of Nebraska]. ProQuest Dissertations and Theses.
- Jekel, J. F. (2007). *Epidemiology, biostatistics and preventive medicine*. Philadelphia: Saunders/Elsevier.
- Johnson, D. L., & Martin, S. (1980). Criterion-referenced testing: New wine in old bottles. *Academic Therapy*, 16(2), 167 - 173.
- Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3), 275-287.
- Kline, P. (2000). *Handbook of psychological testing*. London and New York: Routledge.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard - setting methods. *Applied Measurement in Education*, 2(2), 121–141.
- McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders*, 49(4), 338-348.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276-282.
- Microsoft Corporation. (2013). *Microsoft Excel*. Retrieved from <https://office.microsoft.com/excel>

- Milli Eğitim Bakanlığı. (2014, July 26). *Milli Eğitim Bakanlığı okul öncesi eğitim ve ilköğretim kurumları yönetmeliği [Ministry of National Education regulations for preschool and primary school education institutions]*. Retrieved May 27, 2020, from <http://mevzuat.meb.gov.tr/dosyalar/1703.pdf>
- Milli Eğitim Bakanlığı. (2016, October 28). *Milli Eğitim Bakanlığı ortaöğretim kurumları yönetmeliği [Ministry of National Education regulations for secondary school education institutions]*. Retrieved May 27, 2020, from https://ogm.meb.gov.tr/meb_iys_dosyalar/2016_11/03111224_ooky.pdf
- Mohr, A. K. (2006). *Criterion referenced and norm referenced predictors of student achievement: Teacher perceptions of, and correlations between, Iowa test of basic skills and the palmetto achievement challenge test* [Doctoral dissertation, University of South Carolina]. ProQuest Dissertations and Theses.
- Montgomery, P. C., & Connolly, B. H. (1987). Norm-referenced and criterion referenced tests: Use in pediatrics and application to task analysis of motor skill. *Physical Therapy*, 67(12), 1873-1876.
- Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications*. New Jersey: Pearson.
- Nartgün, Z. (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme [An investigation on whether the applications of the norm- and criterion-referenced assessments over the same scores make a difference in grading]. *Ege Eğitim Dergisi*, 8(1), 19-40.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14(1), 3-19.
- Oescher, J., Kirby, P. C., & Paradise, L. V. (1992). Validating state-mandating criterion-referenced achievement tests with norm-referenced test results for elementary and secondary students. *Journal of Experimental Education*, 60(2), 141-150.
- Paura, L., & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia-Social and Behavioral Sciences*, 109, 1282-1286.
- Pester, A. M. (2003). *Language intervention effects of norm-referenced and criterion referenced test scores* [Master's thesis, Miami University]. https://etd.ohiolink.edu/!etd.send_file?accession=miami1050351250&disposition=inline
- Sakarya University. (2019, April 18). *Sakarya Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim ve Sınav Yönetmeliği [Sakarya University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from <https://www.sakarya.edu.tr/yeni-lisans-ve-onlisans-egitim-ogretim-ve-sinav-yonetmeligi-d330.html>
- SPSS, Inc. (2009). PASW Statistics for Windows (Version 18.0) [Computer Program]. Chicago: SPSS Inc.
- Sönmez, V., & Alacapınar, F. G. (2011). *Örneklendirilmiş bilimsel araştırma yöntemleri [Scientific research methods with examples]*. Ankara: Anı Yayıncılık.
- Toprakçı, E., Baydemir, G., Koçak, A., & Akkuş, Ö. (2007, September). *Eğitim fakültelerinin eğitim-öğretim ve sınav yönetmeliklerinin karşılaştırılması [A comparison of regulations for education and examinations in faculty of education]*. Paper presented at the meeting of 16. Ulusal Eğitim Bilimleri Kongresi, Tokat, Turkey.
- Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Ankara: Pegem Yayıncılık.
- Tyler, R. W. (2013). *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.
- Urbina, S. (2004). *Essentials of psychological testing*. New York: Wiley

- Visintainer, C. (2002). *The relationship between two state-mandated, standardized tests using norm-referenced Terranova and the criteria-referenced, performance assessment developed for the Maryland school performance assessment program* [Doctoral dissertation, Wilmington College]. ProQuest Dissertations and Theses.
- Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., & Wollscheid, S. (2015). *Dropout and completion in higher education in Europe: Main report*.
- Yıldırım, C. (2011). *Bilim felsefesi [Philosophy of science]*. İstanbul: Remzi Kitabevi.
- Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, NJ: Educational Testing Service.

Comparison of confirmatory factor analysis estimation methods on mixed-format data

Abdullah Faruk Kilic ^{1,*}, Nuri Dogan ²

¹Adiyaman University, Faculty of Education, Department of Educational Sciences, Turkey

²Hacettepe University, Faculty of Education, Department of Educational Sciences, Turkey

ARTICLE HISTORY

Received: Aug. 27, 2020

Revised: Nov. 16, 2020

Accepted: Jan. 05, 2021

Keywords:

Bayesian estimation,
Monte-carlo simulations,
CFA.

Abstract: Weighted least squares (WLS), weighted least squares mean-and-variance-adjusted (WLSMV), unweighted least squares mean-and-variance-adjusted (ULSMV), maximum likelihood (ML), robust maximum likelihood (MLR) and Bayesian estimation methods were compared in mixed item response type data via Monte Carlo simulation. The percentage of polytomous items, distribution of polytomous items, categories of polytomous items, average factor loading, sample size and test length conditions were manipulated. ULSMV and WLSMV were found to be the more accurate methods under all simulation conditions. All methods except WLS had acceptable relative bias and relative standard error bias. No method gives accurate results with small sample sizes and low factor loading, however, the ULSMV method can be recommended to researchers because it gives more appropriate results in all conditions.

1. INTRODUCTION

Evidence for validity should be collected first in test development or adaptation studies. The process of collecting validity evidence for a test's structure mostly involves examining the relationships between the variables (Bollen, 1989). Factor analysis is one of the oldest and best known ways to investigate relationships between variables (Byrne, 2016; Osborne & Banjanovic, 2016). The use of confirmatory factor analysis (CFA) in the process of collecting evidence of construct validity is an accepted approach in the literature, and thus frequently used (AERA et al., 2014; DiStefano & Hess, 2005; Guilford, 1946; Nunnally & Bernstein, 1994; Thompson & Daniel, 1996). A search for the key term "confirmatory factor analysis" in the Scopus database resulted in 34.257 articles. When the search was limited to the field of psychology and social sciences, there were 19.546 articles. 461 of these articles were published in 2020. Confirmatory factor analysis is thus frequently used in the field of social sciences and psychology.

The use of CFA requires knowledge of which estimation method provides accurate results under which conditions, because estimation methods affect the results obtained when estimates

CONTACT: Abdullah Faruk KILIC ✉ abdullahfarukkilic@gmail.com 📍 Adiyaman University, Faculty of Education, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

were biased. There are thus numerous studies in the literature comparing CFA estimation methods. An examination of studies in which the observed variables were categorical found that some studies were performed with only five categories of observed data. The manipulated simulation conditions were the distribution of the observed or latent variables, the estimation methods used and the sample sizes in these studies (Babakus et al., 1987; B. O. Muthén & Kaplan, 1985, 1992; DiStefano, 2002; Ferguson & Rigdon, 1991; Lei, 2009; Morata-Ramirez & Holgado-Tello, 2013; Potthast, 1993). Examining other simulation studies with categorical data found that there were between two and seven categories of observed variables (Beauducel & Herzberg, 2006; Dolan, 1994; Flora & Curran, 2004; Green et al., 1997; Li, 2016; Liang & Yang, 2014; Moshagen & Musch, 2014; Rhemtulla et al., 2012; Yang-Wallentin et al., 2010). Studies comparing estimation methods on mixed item response type data, however, were few and limited (Depaoli & Scott, 2015; Oranje, 2003).

The study by Depaoli and Scott (2015) was retracted due to systematic error in the simulation codes. Item type (including different combinations of item types), factor loadings, factor correlations, sample sizes, and priors in the case of Bayesian conditions was examined, however, the percentage and distributions of polytomous items were not manipulated. In the simulation study conducted by Oranje (2003), sample size, number of factors, number of observed variables per factor, and item response-type were manipulated, and ML, WLS and WLS (estimated to Lisrel software), WLSM and WLSMV (estimated to Mplus software) estimation methods were compared. The study reported that as the number of categories increases, the sensitivity of the parameter estimates increases, because polychoric correlations are more appropriate in this condition. However, the distribution of polytomous items was not manipulated in this study, and the study was conducted in a single mixed format test (60% with 2 categories, 20% with 3 categories and 20% with 5 categories).

1.1. The Present Study

Despite the large number of studies comparing CFA estimation methods, there does not seem to be a study comparing both frequentist and Bayesian estimation methods in terms of mixed item response type data. Therefore, investigating this comparison will close this gap in the CFA literature. In addition, the current study studied in a large number of simulation conditions to close this gap. So, the current study can meet the needs of applied researchers who use CFA to collect validity evidence. This study will thus contribute to the literature on CFA estimation methods.

This study investigates which CFA estimation method gives unbiased and accurate results for simulation conditions with mixed item response type data. Research problems were therefore constructed as follows. According to the simulation conditions; which estimation methods have more accurate i) convergence rate and inadmissible solution rate, ii) percentage of accurate estimate (PAE), iii) relative bias (RB), iv) standard error bias (SEB) values? and v) how accurate is the performance of ML, MLR, ULSMV, WLS, WLSMV and Bayesian on four different empirical data sets in terms of convergence, inadmissible solution rate, RB and SEB values?

2. METHOD

A Monte Carlo simulation was used in the present study. Monte Carlo studies are statistical sampling investigations. In these studies, dataset suitable for empirical distribution is generated. The aim of these studies is to produce a data set suitable for empirical distribution. This situation separates Monte Carlo studies from simulation studies. Because in simulation studies, it is possible to generate dataset for population or to demonstrate a statistical analysis. However, sample data are generated in accordance with a certain distribution in Monte Carlo simulations (Bandalos & Leite, 2013). It compared CFA estimation methods in mixed item response type

data. Simulation and empirical data sets were both used in the study. The empirical data set included four tests of the Monitoring and Evaluation of Academic Skills (MEAS) research data sets which conducted by Turkish Ministry of National Education. The tests consist of 18 items, some items are scored as 1-0, some items are 0-1-2 and some 0-1-2-3. The tests included both binary and polytomous items.

2.1. Manipulated Factors

This study focused on achievement tests consisting of mixed item responses. Mixed item response type achievement tests are generally reported to be unidimensional (Bennett et al. 1990; Bennett, Rock, and Wang 1991; Lissitz, Hou, and Slater 2012; van den Bergh 1990). For this reason, the measurement model was defined as unidimensional. The percentage of polytomous items ((10%, 20%, 40%, 50%), skewness of polytomous items (left skewed, normal, right skewed), categories of polytomous items (3, 4 and 5), average factor loading (.40, .60 and .80), sample size (200, 500 and 1000) and test length (20, 30 and 40 items) were manipulated as simulation conditions. The simulation conditions were fully crossed, so, 972 (4x3x3x3x3x3) simulation conditions were manipulated, with 1000 replicates per cell.

The average factor loading was chosen as low (.40), medium (.60) and high (.80). Since the lowest factor loading in such tests is recommended as .40 (Howard, 2016), the low value of the average factor loading is .40, medium is .60 and high is .80. It is not common in practice that all items have the same factor loading, and so unlike other studies, the factor loadings of all the items in the test were not equal (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Forero et al., 2009; Li, 2016a; Liang & Yang, 2014).

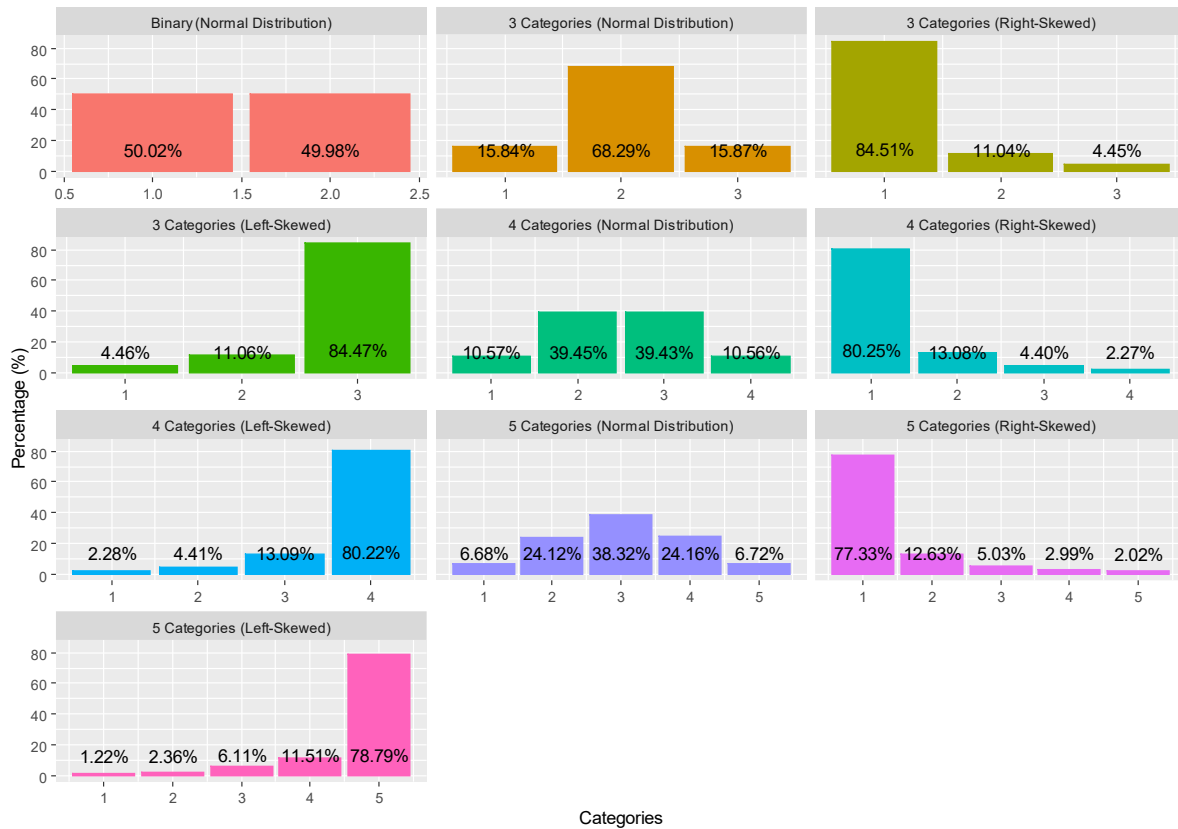
Sample sizes were determined as 200 (small), 500 (medium) and 1000 (large), as used in many other simulation studies (Beauducel & Herzberg, 2006; Li, 2016a; Oranje, 2003; West et al., 1995).

Considering the real test situations, the percentage of polytomous items and the categories of polytomous items were determined as 10%, 20%, 40%, 50%, and 3, 4, 5 respectively. Since it is thought that the distribution of polytomous items may have an impact on the estimates, the distribution of polytomous items was added to the simulation conditions as left-skewed, normal and right-skewed. The test length was manipulated to be short (20 items), medium (30 items) and long (40 items).

2.2. Data Generation

Continuous data sets (continuous latent variable) were first generated for each condition of the study, followed by multivariate normal distribution. Once the continuous data sets were generated, the data was categorized according to simulation conditions. This approach is commonly used in the literature (Beauducel & Herzberg, 2006; Lei, 2009; Morata-Ramirez & Holgado-Tello, 2013; Oranje, 2003; T. K. Lee et al., 2018). This approach also meets the assumption that the underlying variable is normally distributed in psychology (Crocker & Algina, 2008; Gulliksen, 1950). Continuous data sets were categorized as binary (normally distributed), 3 categories (left-skewed, normal and right-skewed), 4 categories (left-skewed, normal and right-skewed) and 5 categories (left-skewed, normal and right-skewed). The distribution of categorical variables used in the study is presented in in [Figure 1](#).

Figure 1. Distribution of variables.



2.3. Outcome Variables

Non-convergence or inadmissible solutions rate, relative bias for factor loadings (RB), percentage of accurate estimates for factor loadings (PAE) and standard errors bias (SEB) were used as outcome variables in the study.

Since 1000 there were replications in the study, estimation methods with 500 or more nonconvergence or inadmissible solutions were considered “NA” for that condition.

Relative bias was calculated via

$$RB = \frac{\hat{\varphi} - \varphi_{True}}{\varphi_{True}} \quad (1)$$

where $\hat{\varphi}$ is the mean of sample estimates over the 1000 replications of average factor loading and φ_{True} is the true average factor loading. In the literature, $|RB| < .05$ indicates trivial bias, $.05 \leq |RB| \leq .10$ indicates moderate bias and $|RB| > 0.10$ indicates substantial bias (Flora & Curran, 2004; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Moshagen & Musch, 2014; Rhemtulla et al., 2012). $|RB| \leq 0.10$ was thus considered “acceptable” in this study.

To determine the percentage of accurate estimate (PAE), the average factor loading value obtained from 1000 replications was examined as to whether it was within $\pm 5\%$ of the real factor loading determined in the simulation condition (Ferguson & Rigdon, 1991; Wolf, Harrington, Clark, & Miller, 2013). Methods with 95% or more of PAE were considered “acceptable” in this study.

Standard error bias (SEB) was calculated via

$$SEB = \frac{\frac{1}{n_{rep}} \sum_{t=1}^{n_{rep}} s\hat{e}(\hat{\theta}_{pt})}{sd(\hat{\theta}_{pt})} \quad (2)$$

where $\widehat{se}(\hat{\theta}_{pt})$ was the standard error of parameter p for replication t , $sd(\hat{\theta}_{pt})$, was the standard deviation of parameter estimates obtained from t replications (Forero et al., 2009; Holtmann, Koch, Lochner, & Eid, 2016; Rhemtulla et al., 2012). When the standard error estimates are equal to the standard deviation obtained empirically, the SEB value will be equal to 1. Accordingly, the SEB value was classified as follows (Holtmann et al., 2016): $5/6 < SEB < 6/5$ was negligible, $2/3 < SEB < 5/6$ and $6/5 < SEB < 3/2$ was medium and $SEB < 2/3$ or $SEB > 3/2$ was large.

The Psych Package (Revelle, 2019) in the R software (R Core Team, 2018) was used to generate the simulation data. Mplus software (L. K. Muthén & Muthén, 2012) was used for CFA. Since 1000 replications were used in the study, the data sets were analyzed in Mplus software using the MplusAutomation (Hallquist & Wiley, 2017) package.

2.4. Data Analysis in Real Data Sets

The empirical data sets were obtained from the Monitoring and Evaluation of Academic Skills (MEAS) research carried out in 2016 in Turkey. Different item types were used in the MEAS research. For this reason, “Open-Ended and Multiple-Choice Question Writing” training was given to item writers by academicians. It is emphasized that the prepared items were reviewed by measurement and evaluation experts and language experts, and after the necessary arrangements, a pilot application was carried out with approximately 5000 students in Ankara, Turkey. The actual application of MEAS research was conducted with the participation of about 38.000 students from 81 provinces of Turkey (MoNE, 2017). A rubric was developed for scoring open-ended items. Accordingly, firstly, correct and partially correct answers were formed. After the pilot application, the answers given by the students to the open-ended items were examined and the unpredictable answers were added to the rubric. Thus, the rubric was composed of four parts as true, partial true, false and empty. The research was conducted in the fields of Turkish, mathematics, science and social studies. The reliability coefficient of the tests as internal consistency ranged between .73 to .85. Test data from Turkish which was focused on reading comprehension (13 binary, 5 three categories), mathematics (12 binary, 6 three categories), science (14 binary, 4 three categories) and social sciences (15 binary, 2 three categories and 1 five categories) was used. Missing data was removed from the data sets via listwise deletion. After removal, the Turkish, mathematics, science and social studies test data consisted of 4745, 2247, 3143 and 3442 individuals, respectively.

Sampling was first undertaken for each data set. Since the sample size conditions were determined as 200, 500 and 1000, the same sample sizes were randomly taken from the Turkish focused on reading comprehension, mathematics, science and social studies test data sets. Sampling was repeated 100 times for each test, in order to avoid the sample bias.

The outcome variables in the analysis performed with real datasets were non-convergence or inadmissible solutions rate, relative bias for factor loadings (RB), and percentage of accurate estimates for factor loadings (PAE).

The true parameter value was needed to calculate the PAE and RB value. The true average factor loading value of the real data sets is unknown. The true value of the average factor loading was obtained using exploratory factor analysis (EFA). For this purpose, EFA was conducted with the whole sample in the Turkish, mathematics, science and social studies datasets. Unweighted least squares (unweighted least square [ULS]), which is claimed to be strong against the assumption that multivariate normality is severely violated, was used as a factor extraction method (Nunnally & Bernstein, 1994; Osborne & Banjanovic, 2016). EFA demonstrated that the data sets were unidimensional.

A repeated measures ANOVA was used to determine which simulation factor is more effective on PAE, RB and r-SEB values. Since the same data sets were analyzed using different

estimation methods, the estimation methods are defined as within-subject. The simulation conditions are defined as between-subject. Partial η^2 was used to examine the effect size. In partial η^2 .01 or less is interpreted as being a small, .06 or more a medium and .14 or more a large effect (Cohen, 1988).

In the real data set, analyses for EFA were performed using Factor 10.08 software (Lorenzo-Seva & Ferrando, 2020).

3. RESULT / FINDINGS

3.1. Convergence and Inadmissible Solution Rate

The convergence rates of maximum likelihood (ML), robust maximum likelihood (MLR) and Bayes are 100% and their inadmissible solutions rates are 0%. A detailed table for the convergence and inadmissible solution rates of other methods is given in Appendices A-E.

The convergence rate of the unweighted least squares mean-and-variance-adjusted (ULSMV) method is 100% and the inadmissible solution rate is 0.01%. The ULSMV method has an inadmissible solution under conditions where the sample size is 200, skewed 3 or 4 category polytomous items, and average factor loading is .80. The coverage rate of all methods except Bayesian is over 90% for all models.

The convergence rate of the weighted least squares mean-and-variance-adjusted (WLSMV) method was 99.99%, while its inadmissible solution rate was .02%. Data sets seem to have convergence problems under conditions where the sample size is 200, the average factor loading is .40, and the test length is 30 or 40 items for WLSMV method. There are inadmissible solutions in conditions similar to ULSMV where the sample size is 200, polytomous items were skewed, the average factor loading was .80, and there were polytomous items in 3 or 4 categories.

The convergence rate of the weighted least squares (WLS) method is 49.48%, and the inadmissible solution rate is 7.03%. The WLS method was not converged under any conditions with a sample size of 200. Additionally, when the sample size was 500, it was not converged under any conditions where the test length was 40 items. Accordingly, it can be said that the WLS method does not converge in small samples or long tests.

There was convergence problem for the WLS method when increasing the number of polytomous items under conditions where sample size was 500, test length was 30 items and average factor loadings were .40 and .60. Increasing the number of polytomous items categories to five resulted in convergence problems under conditions where percentage of polytomous items were 40% and %50. The convergence problems of the WLS method decreased as the sample size increased to 1000.

Examination of the inadmissible solutions in the WLS method suggests that this method has more inadmissible solutions under conditions where the sample size was 1000 and the average factor loading was .80. WLS has inadmissible solutions in about 40% of all data sets under conditions where sample size was 500, test length was 30 items, and the average factor loading was .60.

3.2. Percentage of Accurate Estimates

The PAE of WLS method was not examined due to its low convergence rate. The PAE values of other methods are presented in Appendix F in detail, for all conditions.

Under conditions where the sample size was 200 and average factor loadings were .40 and .60, the PAEs of the all estimation methods were less than 95%. When increasing the average factor loading to .80, the PAE of the methods were greater than 95%. Under 36 conditions where sample size was 200, the average factor loading was .80, and polytomous items had 3 categories

(3 conditions of distributions of polytomous items x 3 conditions of test length x 4 conditions of percentage of polytomous items = 36 conditions), the Bayesian method's PAE values were greater than 95% in more conditions (33 conditions). For the specified simulation conditions (3 conditions of distributions of polytomous items x 3 conditions of test length x 4 conditions of percentage of polytomous items = 36 conditions), the PAE values of the ULSMV method were close to those of the Bayesian method (26 conditions). Under conditions where sample size was 200 and 3 categories of polytomous items followed normal distribution, WLSMV, ULSMV and Bayesian methods had similar PAE values, but the distribution of polytomous items was skewed, and the WLSMV method's PAE values decreased. Under conditions where the sample size was 200, polytomous items had 4 or 5 categories, and polytomous items followed normal distribution, the PAE values were bigger in the Bayesian and ULSMV methods than the ML/MLR and WLSMV methods. When the ML/MLR, and WLSMV methods were compared, the WLSMV method had bigger PAE values.

The PAE value of the methods was below 95% in all conditions with a sample size of 500 and an average factor loading of .40. When the average factor loading increased to .60 and .80, the PAE value of the methods increased to 95%. When the sample size increased to 1000, the PAE values of ULSMV, WLSMV and Bayesian methods exceeded 95% under some conditions with an average factor loading of .40. Accordingly, it can be said that the PAE values increase in the estimation methods when sample size or average factor loading increase.

A repeated measures ANOVA was performed to determine which simulation condition was more effective as regards PAE values. In Mauchly's Test of Sphericity, sphericity was violated ($\chi^2(5) = .01, p < .001$). The Greenhouse-Geisser correction was thus used. There was a statistically significant main effect of the estimation method on PAE values overall $F(1.53, 1357.14) = 27797.19, p = .00, \text{partial } \eta^2 = .97$.

When the average PAE values of the methods were compared with the Bonferroni correction, ULSMV (mean = 89.56%, se = .55) was statistically significantly higher than other methods. The WLSMV (mean = 89.17%, se = .56) method's PAE was statistically significantly higher than both Bayesian (mean = 87.56%, se = .55) and ML/MLR (mean = 67.77%, se = 1.07) methods. The Bayesian method's PAE value, on the other hand, was statistically significantly higher than in the ML/MLR method.

When the test of within-subject effects was examined, the most important second order interaction was found to be method x average factor loading ($F(3.06, 1357.14) = 9053.40, p = .00, \text{partial } \eta^2 = .95$). The other second order interactions method x sample size ($F(3.06, 1357.14) = 1705.99, p = .00, \text{partial } \eta^2 = .79$), method x percentage of polytomous item ($F(3.06, 1357.14) = 877.43, p = .00, \text{partial } \eta^2 = .75$), method x distribution of polytomous item ($F(3.06, 1357.14) = 294.42, p = .00, \text{partial } \eta^2 = .40$) and method x categories of polytomous items ($F(3.06, 1357.14) = 112.99, p = .00, \text{partial } \eta^2 = .20$) had a large effect size, but the interaction of method x test length ($F(3.06, 1357.14) = 14.57, p = .00, \text{partial } \eta^2 = .03$) had a small effect size.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size ($F(6.11, 1357.14) = 1106.00, p = .00, \text{partial } \eta^2 = .83$). The other third order interactions method x sample size x percentage of polytomous items ($F(9.17, 1357.14) = 224.91, p = .00, \text{partial } \eta^2 = .60$), method x distribution of polytomous items x average factor loading ($F(6.11, 1357.14) = 61.04, p = .00, \text{partial } \eta^2 = .22$), method x percentage of polytomous items x sample size ($F(9.17, 1357.14) = 40.66, p = .00, \text{partial } \eta^2 = .22$), method x average factor loading x test length ($F(6.11, 1357.14) = 46.66, p = .00, \text{partial } \eta^2 = .17$) and method x categories of polytomous items x average factor loading ($F(6.11, 1357.14) = 36.14, p = .00, \text{partial } \eta^2 = .20$) had a large effect size. The other interactions had medium and small effect sizes ranging between .01-.13.

Between-subject effect was examined to investigate which simulation condition had a higher effect on PAE values. Average factor loading had the biggest effect on the PAE values ($F(2, 888) = 41569.78, p = .00, \text{partial } \eta^2 = .99$). Sample size ($F(2, 888) = 10670.12, p = .00, \text{partial } \eta^2 = .96$), percentage of polytomous items ($F(3, 888) = 376.89, p = .00, \text{partial } \eta^2 = .56$), test length ($F(2, 888) = 356.13, p = .00, \text{partial } \eta^2 = .45$) and categories of polytomous items ($F(2, 888) = 6.20, p = .00, \text{partial } \eta^2 = .01$) had an effect on the PAE values, however, PAE values do not differ significantly according to the distribution of polytomous items ($F(2, 888) = 0.49, p = .62, \text{partial } \eta^2 < .00$).

Because there were many between subject variables, only second and third order interactions were studied. When second order interactions were examined, the important interaction was found to be average factor loading x sample size ($F(4, 888) = 1693.53, p = .00, \text{partial } \eta^2 = .88$). When results were examined in terms of partial eta squared, the average factor loading x percentage of polytomous items ($F(6, 888) = 79.31, p = .00, \text{partial } \eta^2 = .35$), and average factor loading x test length ($F(4, 888) = 79.31, p = .00, \text{partial } \eta^2 = .26$) interactions had large effect size. Distribution of polytomous items x sample size ($F(4, 888) = 29.51, p = .00, \text{partial } \eta^2 = .12$), percentage of polytomous items x sample size ($F(6, 888) = 17.03, p = .00, \text{partial } \eta^2 = .10$) and test length x sample size ($F(4, 888) = 21.67, p = .00, \text{partial } \eta^2 = .09$) had a medium effect on PAE values. The other interaction effect sizes ranged between .01-.04, and some was not statistically significant.

Examination of the post-hoc tests found that average factor loading categories differed statistically significantly from each other. So, .80 had higher PAE values than .40 and .60. Similarly, .60 had higher PAE values than .40. At the same time, sample size categories were statistically significantly different from each other: 1000 had higher PAE values than 200 and 500. Similarly, 500 had higher PAE values than 200.

Polytomous items with 3 categories had a statistically significantly higher PAE value than those with 4 and 5 categories ($p = .01$). There were no statistically significant differences between polytomous items with 4 and 5 categories. No statistically significant difference was found between the distribution of polytomous items. Accordingly, it can be said that the distribution of polytomous items has no effect on the estimation method's PAE values.

Test length categories differed from each other statistically significantly ($p = .00$). So, 40 items had higher PAE values than 20 and 30. Similarly, 30 items had higher PAE values than 20. So, an increase in the number of items increases the PAE values of the methods. The percentage of polytomous items differed from each other statistically significantly ($p = .00$). So, 50% had higher PAE values than the others (10%, 20% and %40). Similarly, 40% had higher PAE values than 20% and 10%, and 20% had higher PAE values than 10%. As the percentage of polytomous items increases, therefore the PAE values of the estimation method increases.

A repeated measures ANOVA showed that the PAE values of the estimation methods differ from each other. The PAE value was obtained in the highest ULSMV method. This method was followed by WLSMV, Bayesian and ML/MLR. The most effective condition on the PAE of the methods is the average factor loading. This condition was followed by sample size (partial $\eta^2 = .96$), percentage of polytomous items (partial $\eta^2 = .56$), test length (partial $\eta^2 = .45$) and categories of polytomous items (partial $\eta^2 = .01$). When the interaction of conditions was examined, average factor loading x sample size (partial $\eta^2 = .88$) had the biggest effect on PAE values. This interaction was followed by the average factor loading x percentage of polytomous items (partial $\eta^2 = .35$), average factor loading x test length (partial $\eta^2 = .26$), distribution of polytomous items x sample size (partial $\eta^2 = .12$), percentage of polytomous items x sample size (partial $\eta^2 = .10$) and test length x sample size (partial $\eta^2 = .09$).

In summary, an increase in average factor loading, sample size, test length and percentage of polytomous items increases the PAE values of the estimation methods. Interestingly, the PAE values of the methods increased as the categories of polytomous items decreased.

3.3. Relative Bias

The RB value in the conditions converged by WLS generally decreased with an increasing number of items (substantial bias), and with a decreasing number of items, the value of RB increased (moderate bias). WLS has not been compared with other methods in which it has moderate or substantial bias under the conditions where WLS could converge. The RB values of all methods are presented in Appendix G in detail.

In the simulation conditions with a sample size of 200, the ULSMV and WLSMV had trivial RB. While the ML/MLR methods were moderately biased under conditions where average factor loading was .40, ML/MLR estimation methods have trivial RB when the average factor loading increased to .60 or .80. The Bayesian method has trivial bias in most conditions where the average factor loading was .40 and in all conditions with an average factor loading of .60 and .80.

Under conditions where the sample size was 500 and 1000, Bayesian, ULSMV and WLSMV methods had trivial bias. ML/MLR methods had trivial bias in most simulation conditions where average factor loading is .40, and in all simulation conditions where average factor loading is .60 and .80.

The RB values were acceptable ($|RB| \leq .10$) for all simulation conditions in all methods except WLS. A repeated measures ANOVA was performed to examine simulation conditions affecting RB values, and thus, to examine which conditions were more effective. Mauchly's Test of Sphericity showed that sphericity was violated ($\chi^2(5) = .00, p < .001$) the Greenhouse-Geisser correction was thus used. There was a statistically significant main effect from the estimation method on RB scores overall $F(1.00, 883.18) = 44.72, p = .00, \text{partial } \eta^2 = .05$.

When the average RB values of the methods were compared with the Bonferroni correction, it was observed that the ULSMV (mean = -.00, se = .00) method had a statistically significantly lower RB value than other methods. The Bayesian (mean = -.01, se = .00) method is lower than both the WLSMV (mean = .02, se = .01) and ML/MLR (mean = -.04, se = .00) methods. The WLSMV method, on the other hand, has a statistically significantly lower RB value than the ML/MLR method.

When tests of within-subject effects was examined, it was observed that the most important second order interaction was method x average factor loading ($F(2.00, 883.18) = 13.68, p = .00, \text{partial } \eta^2 = .03$) which has a small effect size. Method x sample size ($F(2.00, 883.18) = 5.93, p = .00, \text{partial } \eta^2 = .01$) also has a small effect size, but the other second order interactions were not statistically significant.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size ($F(4, 883.18) = 4.80, p = .00, \text{partial } \eta^2 = .02$) which has a small effect size. Method x sample size x percentage of polytomous items ($F(4, 883.18) = 2.21, p = .00, \text{partial } \eta^2 = .01$) also has a small effect size, but the other third order interactions were not statistically significant.

The between-subject effect was examined to investigate which simulation condition has a greater effect on RB values. Sample size had the largest effect on RB values ($F(2, 883) = 3.65, p = .03, \text{partial } \eta^2 = .01$). Average factor loading ($F(2, 883) = 3.58, p = .03, \text{partial } \eta^2 = .01$) had a smaller effect on RB values. Other simulation conditions had no effect on RB values.

When second order interactions were examined, the most important interaction was found to be test length x percentage of polytomous items ($F(6, 883) = 2.23, p = .04, \text{partial } \eta^2 = .01$) which has a small effect size. The other interactions were not statistically significant.

When post-hoc tests were examined, conditions where the average factor loading was .80 had statistically significantly smaller RB values than for .40, but there was no statistically significant difference between .80 and .60 conditions. The other simulation conditions did not affect RB values.

A repeated measures ANOVA demonstrated that the RB values of the methods differed from each other and ULSMV had the lowest RB value. This was followed by Bayesian, WLSMV and ML/MLR methods. The most effective condition regarding the RB values of the methods was sample size ($\text{partial } \eta^2 = .01$). This condition was followed by average factor loading ($\text{partial } \eta^2 = .01$). When the interaction of conditions was analyzed, method x average factor loading ($\text{partial } \eta^2 = .03$) had the largest effect on RB values.

In summary, the simulation conditions, generally, have no effect on RB values, but the condition where average factor loading was .80 had a smaller RB value.

3.4. Standard Error Bias

ML and MLR methods have negligible standard error bias in all conditions. Bayes, ULSMV and WLSMV methods were negligibly biased in most of the 200 sample size conditions. All estimation methods except WLS had negligible bias in conditions where sample size was 500 and 1000. WLS method, generally, have large bias in most conditions if converged. The SEB values obtained from the estimation methods according to the simulation conditions are presented in Appendix H for more information.

A repeated measures ANOVA was performed to examine simulation conditions affecting SEB values. Mauchly's Test of Sphericity showed that sphericity was violated ($\chi^2(9) = .00, p < .001$) so the Greenhouse-Geisser correction was used. Estimation method had a statistically significant main effect on SEB values $F(1.93, 1711.09) = 8991.97, p = .00, \text{partial } \eta^2 = .91$.

When the average SEB values of the methods were compared with the Bonferroni correction, the SEB values of the ULSMV (mean = .98 se = .00) and MLR (mean = .98 se = .00) methods differed statistically significantly from other methods and were observed to be closer to 1 (which means that there is no bias). ML (mean = .97, se = .00) differed statistically significantly from both WLSMV and Bayesian methods. The WLSMV method (mean = .96, se = .00) had a statistically significantly higher SEB value than the Bayesian method (mean = .92, se = .00).

When the test of within-subject effects was examined, the most important second order interaction was found to be method x sample size ($F(3.85, 1711.09) = 2703.63, p = .00, \text{partial } \eta^2 = .86$). The other second order interactions method x average factor loading ($F(3.85, 1711.09) = 572.71, p = .00, \text{partial } \eta^2 = .56$), method x categories of polytomous items ($F(3.85, 1711.09) = 175.55, p = .00, \text{partial } \eta^2 = .28$), method x percentage of polytomous items ($F(5.78, 1711.09) = 115.48, p = .00, \text{partial } \eta^2 = .28$), method x distribution of polytomous items ($F(3.85, 1711.09) = 153.21, p = .00, \text{partial } \eta^2 = .26$), and method x test length ($F(3.85, 1711.09) = 74.42, p = .00, \text{partial } \eta^2 = .14$) had a large effect size.

When the third order interactions were examined, the most important third order interaction was found to be method x average factor loading x sample size ($F(7.71, 1711.09) = 346.14, p = .00, \text{partial } \eta^2 = .61$). The other third order interactions method x sample size x percentage of polytomous items ($F(11.56, 1711.09) = 77.92, p = .00, \text{partial } \eta^2 = .34$), method x distribution of polytomous items x sample size ($F(7.71, 1711.09) = 68.06, p = .00, \text{partial } \eta^2 = .23$), method x categories of polytomous items x sample size ($F(7.71, 1711.09) = 33.01, p = .00, \text{partial } \eta^2 = .13$)

had a large effect size. The other interactions were medium and small effect size, which ranged between .01-.13.

The between-subject effect was examined to investigate which simulation condition had a greater effect on the SEB values of the methods. Percentage of polytomous items had the greatest effect on SEB values ($F(3, 888) = 303.42, p = .00, \text{partial } \eta^2 = .51$). The other simulation conditions, sample size ($F(2, 888) = 144.21, p = .00, \text{partial } \eta^2 = .25$) and distribution of polytomous items ($F(2, 888) = 104.24, p = .00, \text{partial } \eta^2 = .19$) had a large effect on the SEB value overall. Average factor loading ($F(2, 888) = 61.69, p = .00, \text{partial } \eta^2 = .12$) and categories of polytomous items ($F(2, 888) = 56.96, p = .00, \text{partial } \eta^2 = .11$) had a medium effect on SEB value overall. Test length ($F(2, 888) = 21.09, p = .00, \text{partial } \eta^2 = .05$) had a small effect on SEB value overall.

When second order interactions were examined, the most important interaction was found to be average factor loading x sample size ($F(4, 888) = 65.61, p = .00, \text{partial } \eta^2 = .23$) which had a large effect. The other interaction effect sizes ranged between .01-.03, and some was not statistically significant.

When post-hoc tests were examined, average factor loading categories were found to differ from each other statistically significantly: .80 had higher SEB values than .40 and .60. Similarly, .60 had higher SEB values than .40. At the same time, the condition where sample size was 1000 had statistically significantly higher SEB values than the sample size was 200. Polytomous items with 3 categories had statistically significantly smaller SEB values than those with 4 and 5 categories ($p = .00$). There was no statistically significant difference between polytomous items with 4 and 5 categories. Accordingly, the SEB values of the methods are more accurate in 4 and 5 categories polytomous items. Polytomous items which followed normal distribution had more accurate SEB values than right or left skewed ones ($p = .00$). No statistically significant difference was observed between the right or left skewed polytomous items.

The condition where test length was 20 items had more accurate SEB values than 30 and 40 item conditions ($p = .00$). No statistically significant difference was observed between the test length for 30 and 40 items. The condition where the percentage of polytomous items was 10% had more accurate SEB values than the others (10%, 20% and 40%). The increase in the percentage of polytomous items caused the SEB values to decrease. Accordingly, the decrease in the percentage of polytomous items caused more accurate SEB values.

Repeated measures ANOVA revealed that the SEB values of the methods differed from each other, and the most appropriate SEB value was in the ULSMV and MLR methods. These methods were followed by ML, WLSMV and Bayesian methods. The most effective condition of SEB values in the estimation methods was percentage of polytomous items ($\text{partial } \eta^2 = .51$). This condition was followed by sample size ($\text{partial } \eta^2 = .25$), distribution of polytomous items ($\text{partial } \eta^2 = .19$), average factor loading ($\text{partial } \eta^2 = .12$), categories of polytomous items ($\text{partial } \eta^2 = .11$) and test length ($\text{partial } \eta^2 = .01$). Interaction of average factor loading x sample size ($\text{partial } \eta^2 = .23$) had the largest effect on SEB values. The effect sizes of other interactions were small (range between .01-.03).

In summary, an increase in categories of polytomous items, average factor loading, and sample size resulted in more accurate SEB values. A decrease in the test length and percentage of polytomous items resulted in more accurate SEB values. Polytomous items followed a normal distribution which makes SEB values more accurate.

3.5. Analysis of The Empirical Data Set

For sample sizes of 200, 500 and 1000 in Turkish, mathematics, science and social science tests, the convergence and inadmissible solution rates of ML/MLR and Bayesian methods were 100%

and 0%, respectively. The ULSMV method converged on all datasets but produced 8% and 6% inadmissible solutions in Turkish and mathematics datasets of 200 sample sizes, respectively. WLSMV converged in all data sets, similar to ULSMV, with 22% and 4% inadmissible solutions in Turkish and mathematics datasets with a sample sizes of 200, respectively.

When the results were examined in terms of PAE, the ML, MLR and WLS methods did not exceed 95% in any sample size. The PAE values of the Bayesian method were bigger than 95% when sample size was 1000, while it is generally below 95% when sample sizes were 200 and 500. As the average factor loading increased, the PAE values of the Bayesian method increased. The PAE values of the ULSMV and WLSMV methods were greater than 95% when the sample size was 1000. The PAE values of the WLSMV and ULSMV methods tended to increase as the average factor loading increased.

When the RB values of the estimation methods were examined, ULSMV and WLSMV methods were found to have trivial bias. The Bayesian method, on the other hand, had moderate bias only in the 200 and 500 sample sizes of the mathematics data set, and trivial bias in the other data sets.

The ML and MLR methods generally have medium bias except in the Turkish data set with a sample size of 500 and social science data set with a sample size of 200. These methods have negligible bias for these data sets.

The WLS method generally estimated the factor loadings more highly than it would if it converged. It has a large bias in 200 and 500 sample sizes. WLS has a negligible bias in the Turkish, science and social science data sets with sample sizes of 1000, however, when the PAE values of the WLS method were analyzed for these data sets, PAE values were 20%, 27% and 13%, respectively.

4. DISCUSSION and CONCLUSION

The estimation methods used for CFA in the current study were compared with mixed item response types, and thus, the performance of CFA estimation methods in mixed format tests were examined. Adding the Bayesian method as well as frequentist estimation methods allowed their performance in mixed format tests to be compared in a large number of conditions. Previous studies comparing CFA estimation methods have reported WLSMV or ULSMV methods as giving better results than estimation methods in many respects (Forero et al., 2009; Li, 2014; Rhemtulla et al., 2012; Savalei & Rhemtulla, 2013; Shi, DiStefano, McDaniel, & Jiang, 2018), however, all the items in these studies have the same number of categories.

As a result of the study, the following findings were obtained. First, the convergence rates of ML/MLR and Bayesian methods were 100% and the inadmissible solutions were 0%, similar to other studies (Forero et al., 2009; Jin, Luo, & Yang-Wallentin, 2016; Lee & Song, 2004; Li, 2016; Liang & Yang, 2014, 2016; Moshagen & Musch, 2014; Zhao, 2015). While convergence rate and inadmissible solutions of ULSMV were 100% and 0.01% respectively, WLSMV was 99.99% and 0.02%. The WLS method did not converge in small samples, as found in other studies, and the convergence rate of WLS was 49.48% and the inadmissible solution rate of WLS was 7.03% (Bandalos, 2014; Finney & DiStefano, 2013; Olsson, Foss, Troye, & Howell, 2000; Oranje, 2003).

Second, similar to other studies in the literature (Forero et al., 2009; Li, 2014; Rhemtulla et al., 2012; Savalei & Rhemtulla, 2013; Shi et al., 2018), ULSMV estimated factor loadings more accurately than other methods. Mixed item response type data thus gives similar results to non-mixed data. The WLSMV method also had similar results to ULSMV. ULSMV was more accurate in parameter estimates in this study, however, when the sample size was small ($n =$

200) and the average factor loading was low (.40), no estimation method had sufficient PAE values (PAE > 95%).

Third, when evaluated in terms of relative bias, all methods except WLS were within the acceptable range ($|RB| < .10$). The simulation study conducted by Shi et al. (2018) compared WLSMV, ULSMV and WLSM methods, and found that ULSMV and WLSMV methods had acceptable bias for all sample sizes (200, 500 and 1000). They also emphasized that the ULSMV method performed slightly better than the WLSMV method. Similarly, it was observed in the current study that ULSMV was less biased than other methods at a statistically significant level. The same methods were suitable in mixed item response type data. Lei (2009) found that ML and WLSMV had unbiased parameter estimates. The estimation methods gave similar results in mixed item response type data to five point categorical data. Liang and Yang (2014) stated that the WLSMV method is slightly better than the Bayes method in terms of bias. Since non-informative priors were used in the current study, the Bayesian method may have had a larger bias than other methods, however, the RB value of the Bayesian method was also within the acceptable range ($|RB| < .10$).

Fourth, the standard error bias (SEB) values of all methods, except WLS, were negligible with increasing sample size. The SEB values of all methods are acceptable, except those for WLS, however, the SEB values differed statistically significantly according to the methods. The ULSMV and MLR methods had the least SEB value. Repeated measures ANOVA demonstrated that the SEB values of the methods differed from each other, and that the ULSMV and MLR methods had the most appropriate SEB value for all simulation conditions. Generally, the increase of categories of polytomous items, factor loading, and sample size make the SEB value more accurate, and the decrease in the test length, the percentage of polytomous items and polytomous items follow normal distribution and make the SEB values more accurate. Jin et al. (2016) also noted that the SEB values of WLSMV, ULS and ML methods were acceptable. Mixed item response type data does not cause a big change in the SEB values of the methods.

Similar results to those of the simulation study were obtained in the analyses performed with real data sets. ML/MLR and Bayesian methods converged in all datasets and had no inadmissible solution. ULSMV and WLSMV converged in all datasets but had a small number of inadmissible solutions. All methods, except WLS, had acceptable RB values.

In conclusion, the ULSMV estimation method is preferable when performing CFA with mixed item response type data, so that parameter estimates can be more accurate. Although the results of the methods are within the acceptable range in terms of RB and SEB values, when evaluated in terms of PAE, ULSMV is slightly better than WLSMV in parameter estimates. However, it should be remembered that the method's PAE values were not in the acceptable range for small sample sizes and low average factor loading. No estimation method is suitable for every condition in mixed item response type data, and the estimation method should be selected considering the sample size and the average factor loading. In future studies, researchers could perform simulation studies manipulating the number of factors, and correlations between factors, using informative priors for the Bayesian method. This study is limited to MEAS data sets collected in 2016. This study is also limited to the 491 simulation conditions at the time. In the current study, mixed item response type data was created to be binary and three categories, binary and four categories or binary and five categories polytomous data independently. This could be manipulated in future studies as binary and three, four and five categories, simultaneously.

Acknowledgments

This paper was produced from the first author's doctoral dissertation.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Abdullah Faruk KILIÇ: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Nuri DOĞAN:** Investigation, Methodology, Supervision, and Validation.

ORCID

Abdullah Faruk KILIÇ  <https://orcid.org/0000-0003-3129-1763>

Nuri DOĞAN  <https://orcid.org/0000-0001-6274-2016>

5. REFERENCES



- AERA, APA, NCME, American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement In Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 24(2), 222-228. <https://doi.org/10.2307/3151512>
- Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 102-116. <https://doi.org/10.1080/10705511.2014.859510>
- Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625-666). Information Age.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118619179>
- Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Depaoli, S., & Scott, S. (2015). Frequentist and bayesian estimation of CFA measurement models with mixed item response types: A monte carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, (September), 1-16. <https://doi.org/10.1080/10705511.2015.1044653> (Retraction published 2015, *Structural Equation Modeling: A Multidisciplinary Journal*, 318)
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(3), 327-346. https://doi.org/10.1207/S15328007SEM0903_2
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, 23(3), 225–241. <https://doi.org/10.1177/073428290502300303>

- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47(2), 309–326. <https://doi.org/10.1111/j.2044-8317.1994.tb01039.x>
- Ferguson, E., & Rigdon, E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, 28(4), 491–497. <https://doi.org/10.2307/3172790>
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Information Age.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4), 625–641. <https://doi.org/10.1080/10705510903203573>
- Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2), 108–120. <https://doi.org/10.1080/10705519709540064>
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427–438. <https://doi.org/10.1177/001316444600600401>
- Hallquist, M., & Wiley, J. (2017). *MplusAutomation: Automating Mplus model estimation and interpretation*. Retrieved from <https://cran.r-project.org/package=MplusAutomation>
- Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*, 51(5), 661–680. <https://doi.org/10.1080/00273171.2016.1208074>
- Jin, S., Luo, H., & Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 680–694. <https://doi.org/10.1080/10705511.2016.1189334>
- Lee, T. K., Wickrama, K., & O’Neal, C. W. (2018). Application of latent growth curve analysis with categorical responses in social behavioral research. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(2), 294–306. <https://doi.org/10.1080/10705511.2017.1375858>
- Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4), 653–686. https://doi.org/10.1207/s15327906mbr3904_4
- Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, 43(3), 495–507. <https://doi.org/10.1007/s11135-007-9133-z>
- Li, C.-H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* [Unpublished Doctoral dissertation]. Michigan State University.
- 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. <https://doi.org/10.3758/s13428-015-0619-7>
- Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of*

- Quantitative Research in Education*, 2(1), 17-38. <https://doi.org/10.1504/IJQRE.2014.060972>
- Liang, X., & Yang, Y. (2016). Confirmatory factor analysis under violations of distributional and structural assumptions: A comparison of robust maximum likelihood and bayesian estimation methods. *Journal of Psychological Science*, 39(5), 1256–1267. <https://doi.org/10.1504/IJQRE.2013.055642>
- Lorenzo-Seva, U., & Ferrando, P. J. (2020). *Factor* (Version 10.10.03) [Computer software]. Universitat Rovira i Virgili.
- MoNE. (2017). *Monitoring and evaluation of academic skills report for eight graders*. MONE. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf
- Morata-Ramirez, M. de los A., & Holgado-Tello, F. P. (2013). Construct validity of likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *International Journal of Social Science Studies*, 1(1), 54-61. <https://doi.org/10.11114/ijsss.v1i1.27>
- Moshagen, M., & Musch, J. (2014). Sample size requirements of the robust weighted least squares estimator. *Methodology*, 10(2), 60-70. <https://doi.org/10.1027/1614-2241/a000068>
- Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38(2), 171–189. <https://doi.org/10.1111/j.2044-8317.1985.tb00832.x>
- Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, 45(1), 19–30. <https://doi.org/10.1111/j.2044-8317.1992.tb00975.x>
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0* [Computer software]. Muthén & Muthén.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). McGraw-Hill.
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4), 557–595. https://doi.org/10.1207/S15328007SEM0704_3
- Oranje, A. (2003, April 21-25). *Comparison of estimation methods in factor analysis with categorized variables: Applications to NEAP data* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education. Chicago, IL, USA.
- Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS®*. SAS Intitute Inc.
- Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, 46(2), 273–286. <https://doi.org/10.1111/j.2044-8317.1993.tb01016.x>
- R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.r-project.org/>.
- Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research*. <https://cran.r-project.org/package=psych>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. <https://doi.org/10.1037/a0029315>
- Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, 66(2), 201–223. <https://doi.org/10.1111/j.2044-8317.2012.02049.x>

- Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(6), 924-945. <https://doi.org/10.1080/10705511.2018.1449653>
- Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, 56(2), 197–208. <https://doi.org/10.1177/0013164496056002001>
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *National Institutes of Health*, 76(6), 913–934. <https://doi.org/10.1177/0013164413495237>
- Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3), 392–423. <https://doi.org/10.1080/10705511.2010.489003>
- Zhao, Y. (2015). The performance of model fit measures by robust weighted least squares estimators in confirmatory factor analysis [Doctoral dissertation, The Pennsylvania State University]. <https://etda.libraries.psu.edu/catalog/24901>

Developing and validating a computerized oral proficiency test of English as a foreign language (Coptefl)

Cemre Isler ^{1,*}, Belgin Aydin ²

¹Department of Foreign Language Education, English Language Education, Faculty of Education, Fırat University, Elazığ, Turkey

²Department of Foreign Language Education, English Language Education, TED University, Ankara, Turkey

ARTICLE HISTORY

Received: May 19, 2020

Revised: Oct. 13, 2020

Accepted: Jan. 05, 2021

Keywords:

Language testing,
Oral proficiency testing,
Computerized oral proficiency testing,
English language learners,
EFL context.

Abstract: This study is about the development and validation process of the Computerized Oral Proficiency Test of English as a Foreign Language (COPTEFL). The test aims at assessing the speaking proficiency levels of students in Anadolu University School of Foreign Languages (AUSFL). For this purpose, three monologic tasks were developed based on the Global Scale of English (GSE, 2015) level descriptors. After the development of the tasks, it was aimed to develop the COPTEFL system and then compare the test scores and test-takers' perspectives on monologic tasks between the COPTEFL and the face-to-face speaking test. The findings from these quantitative and qualitative analyses provided substantial support for the validity and reliability of the COPTEFL and inform the further refinement of the test tasks.

1. INTRODUCTION

Testing students' overall language ability in an efficient manner is one of the primary challenges faced by large-scale preparatory school programs in the universities of Turkey (Aydın et al., 2016). The demands of efficiency often take precedence over in the proficiency tests of these programs and as a result, in most cases, the administrations of oral proficiency tests are not held for reasons of impracticality and difficulty of implementation (Aydın et al., 2016; 2017). That is, the administration of oral proficiency testing is a time consuming and labor-intensive process (Kenyon & Malabonga, 2001; Mousavi, 2007). For example, the employment of a trained interviewer, such as in the face-to-face oral proficiency interviews, brings about its logistical issues when large numbers of test-takers are to be tested. Other practices, such as paired or group testing procedures, also consume much time and attention in the process of the administrations and are most feasible for small-scale assessments (Malabonga, Kenyon & Carpenter, 2005). Thus, the demands for testing speaking make it impractical to systematically measure in foreign language programs. For this reason, many institutions don't even try to test

CONTACT: Cemre Isler ✉ cemreisler@anadolu.edu.tr 📍 Fırat University, Department of Foreign Language Education, English Language Education, Faculty of Education, Fırat University, Elazığ, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

speaking skills (Aydın et al., 2016).

Due to not given the same evaluative attention as the other skills, Turkish learners of English do not experiment with the oral language as much as they do with the written language. This situation, in turn, causes a lack of motivation for the achievement of communicative oral skills on the part of the students (Aydın et al., 2016). Harlow & Caminero (1990) articulated this point as: “If we pay lip service to the importance of oral performance, then we must evaluate that oral proficiency in some visible way” (p.489). Indeed, most English language instructors in Turkey are well aware of the importance and necessity to test directly the speaking skill in the proficiency tests. Teachers, however, are confronted with the fact that there does not exist an oral proficiency instrument or a model that is easy to implement for a large group of students in terms of time and logistics. One of the current studies on this topic was conducted by Aydın et al. (2016) in which they carried out a series of interviews with the administration of twelve schools of foreign languages in Turkey. There were two purposes that leading this study. First, it was aimed to explore the practices used by the universities to prepare reliable and valid language proficiency and to discuss the feasibility of these practices in their contexts. Second, it was aimed to collect opinions from these state universities in Turkey about the use of computer-assisted assessment techniques in the assessment of language proficiency, as well as to identify the existing practices if there any. The findings of the study revealed a detailed picture of the present practices of universities concerning language proficiency tests. The most prominent findings of the study showed that (1) all institutions believe the importance of including four skills in a proficiency test; namely reading, listening, writing, and speaking. Yet, most of them cannot test the speaking skill due to practical reasons; (2) most of the institutions refer to not having sufficient human resources and technical equipment for the preparation, administration, and assessment procedures of proficiency tests. These tests are mostly prepared and administered by the instructors assigned for this job or volunteers to do it. The number of staff in testing units who received education in assessment and evaluation is quite low; (3) they also state experiencing certain problems in the administration of proficiency tests. Accordingly, it is not possible to pilot the tests due to time limitations both for administration and assessment procedures. Due to the high number of students, tests are provided in multiple-choice format and the statistical analyses of test results are not done by experts in most institutions because of the reasons mentioned above. Within the purpose of this study, particularly about the speaking skill, the data gathered from the leading universities of Turkey clearly show that among all skills, testing oral proficiency is referred to as the most problematic one which results in not testing at all. The results, all in all, clearly depict the lack of agreed content and the administration and assessment of the framework for proficiency tests. However, establishing certain standards in foreign language education seems inevitable to catch up with the developed countries with regard to internationally recognized language tests in terms of validity, reliability, and usability. In this regard, all the universities that participated in the study emphasize the necessity of establishing certain standards in foreign language education. Also, all of them except one state that they support the idea of developing a nation-wide proficiency test by using technology.

When we have a look at the studies on educational technology, we see that with the recent advancements in computer technology, the use of computers in the delivery of oral proficiency tests has begun appealing due to its potential benefits such as increased reliability of the test as a consequence of the standardization of test delivery process, more efficient test administration and the flexibility in the delivery of tasks (Mousavi, 2007; Zhou, 2015). Although recent advances in computer technology have promoted the computer delivery of oral proficiency tests, the absence of an interviewer has resulted in concerns about the validity of using them as a replacement for face-to-face speaking tests (Zhou, 2015). Accordingly, the most ubiquitous concern was that test-takers’ performance on a computer-based speaking test may not reflect

their ability measured by face-to-face speaking tests in which test-takers are required to interact with an interviewer (Zhou, 2008; 2015). Examining this issue of importance, since it concerns fundamental questions of test validation, i.e. to ensure the score interpretations (Zhou, 2008). So, there has been a call for more research on comparing computer-based tests with conventional face-to-face speaking tests. Given that the score equivalence is significant and should be established prior to the interpretations of computer-based speaking tests, in the present study it was attempted firstly to develop the Computerized Oral Proficiency Test of English as a Foreign Language (COPTTEFL) and then investigate the equivalence of the semi-direct (COPTTEFL) and the direct (face-to-face) versions of a test of oral proficiency. The present study is, therefore, comparability research and it primarily relied on concurrent validation which focuses on the equivalence between test scores. However, this study argues that examining the relationship between test scores only through concurrent validation might provide insufficient evidence as to whether the COPTTEFL measures what it intended to measure. It suggests demonstrating the validity of the test from multiple perspectives. In this respect, it suggests that test-taker attitudes might represent an important source to obtain a deeper understanding with regard to the construct validity and face validity of the tests. If test-takers' attitudes towards the test seriously affect their scores, the scores may not reflect their real language ability, which the test is intended to assess and consequently, the test would lack construct and face validity. With these purposes, it was firstly aimed to develop a computer-based speaking test system, namely the COPTTEFL which would be established on a framework of test validation.

1.1. A framework for validating a speaking test

The most useful starting point for the test development is to have a framework of validation to support the claims made for the tests. If the study is to establish whether the test is valid as a testing instrument, it is essential to utilize a framework of validation in order to collect data systematically and objectively. The socio-cognitive framework (Weir, 2005) for validating the test was used in the present research. It was operationalized from the initial stages in the development of the test of speaking to the comparability of scores by each mode of testing. Several frameworks for language test validation have been proposed by earlier theorists, but as put forward by O'Sullivan (2011a), they have been unable to offer an operational specification for test validation. The approach taken by Weir (2005), however, defined each aspect of validity with sufficient detail as to make the model operationalizable for each of the four skills (O'Sullivan, 2011b).

The socio-cognitive framework for validating the speaking test (Weir, 2005) was used as the major reference by which the speaking tasks of the study were developed. The framework offers a guideline for validating the speaking tests by demonstrating the steps that need to be followed for validity and reliability concerns. The essential components to be investigated in the framework are as follows: (1) Test-taker characteristics, (2) Theory-based validity, (3) Context validity, (4) Scoring validity, (5) Criterion-based validity, (6) Consequential validity.

Firstly, test-taker concern has been raised by Weir (2005) and it was argued that it is directly related to the theory-based validity since test-taker characteristics have an impact on the way test-takers process the test task. He stated that physical, psychological, and experiential differences of the individuals should be considered during the test development process so that bias for or against a particular group can be avoided. Secondly, theory-based validity is related to considerations regarding how well a test task correlates with cognitive (internal mental) processes resembling those which language users employ when undertaking similar tasks in non-test conditions. Thirdly, context validity is related to the appropriacy of the contextual properties of the test tasks to assess specific language ability. Moreover, scoring validity is concerned with the extent to which test results are consistent with respect to the content

sampling and free from bias. Criterion-based validity is about the relationship between test scores and other external measurements that assess the same ability. Finally, consequential validity refers to the impact of tests and test scores interpretations on teaching, learning, individuals, and society.

The present study only focused on theory-based validity, context validity, and scoring validity. The other aspects of the framework; test-taker characteristics, criterion-based and consequential validity were not investigated. This was decided on for the reason that it was beyond the scope of the study to collect data on all components due to time constraints. Therefore, only those included in the study were discussed in the following part of the study.

1.2. Theory-based validity

Theory-based validity, construct validity, or later renamed as cognitive validity (Khalifa & Weir, 2009), is one of the components of Weir's (2005) socio-cognitive framework for validating language tests and concerned with the internal mental processes. In relation to the cognitive processes elicited from test-takers, Field (2013) argues that the main concern is not whether the tasks are close to an actual speaking or listening event, but whether these tasks require test-takers to employ the internal mental processes that a language user normally undertakes in similar tasks during non-test conditions. Reflecting on the representatives of the mental processes in test tasks is the main concern for cognitive validity. Therefore, the focus in studies of cognitive validity is not on the speech produced by the test-taker, but rather the mental processes that a test-taker undertakes in speech production during a speaking test. At this point, the relationship between theory-based validity and context validity is a symbiotic one. The context in which the test task is presented has an impact on the mental processes of the test-taker. For example, the mode of input, whether it is listening to the dialogue or looking at pictures will influence how the test-taker conceptualizes and processes these messages as pre-verbal messages (Zainal Abidin, 2006). The speaking skill descriptors provided by the Global Scale of English (GSE, 2015) were used in the present study to define the language construct and determine the target sub-skills of the construct.

1.2.1. GSE descriptors for the speaking skill

After Messick's (1989) challenge against the traditional view of validation, validity is not seen as a characteristic of a test, but a feature of the inferences made on the basis of test scores. The focus here is the test score or the results of the test since this is what is used to make interpretations about test-takers' ability (Chapelle, 2013). As stated in Chapelle (2013), in current approaches, scores are interpreted with regard to pre-determined standards of knowledge. For example, the increasingly used Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) represents an ordered set of statements through six common reference levels (A1, A2, B1, B2, C1, C2; *ranging from lowest to highest*) that describing language proficiency. It is claimed, for example, that a speaker assessed as meeting the standard for level B1:

“Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst traveling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes, and ambitions and briefly give reasons and explanations for opinions and plans” (Council of Europe, 2001, p.24),

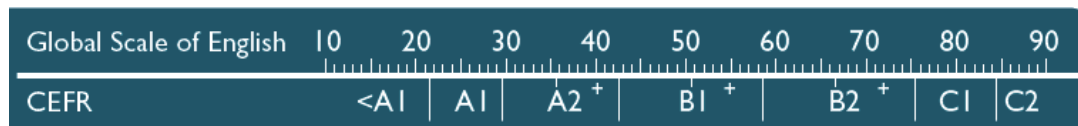
while a speaker at B2:

“Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint

on a topical issue giving the advantages and disadvantages of various options” (Council of Europe, 2001, p.24).

The development of a test instrument begins with such a set of standards (Chapelle, 2013). These may be rather general, as in the case of the CEFR, or more granular, as in the GSE. The GSE proficiency scale was created with reference to the CEFR, but the main difference between the GSE proficiency scale and the CEFR proficiency scale stems from its granular structure (see Figure 1).

Figure 1. Global Scale of English aligned with the CEFR.



As shown in Figure 1, the GSE presents a more granular measurement of proficiency within a single CEFR level (GSE, 2015). It is a proficiency scale from 10 to 90 and defines what a learner can do across four skills at a specific GSE range. For example, a language learner at GSE range 27 “can understand a phone number from a recorded message, but a learner at 74 “can follow an animated conversation between two fluent speakers” in listening skill. As for reading skills, a learner at 43 on the scale “can understand simple technical information (e.g. instructions for everyday equipment)” whereas a learner at 58 “can recognize the writer’s point of view in a structured text”. As for speaking skills, a learner at 42 “can give a short basic description of events and activities” while the ones at 61 “can engage in extended conversation in a clearly participatory fashion on most general topics”.

Most of the preparatory programs in Turkey use the CEFR as a proficiency scale where the learner proficiency is classified from A1 (low basic) to C2 (fully proficient) (Council of Europe, 2001). However, in the 2014-2015 academic years, Anadolu University School of Foreign Languages (AUSFL) moved away from CEFR towards the GSE which is psychometrically aligned to CEFR (GSE, 2015). The reason for this shift from CEFR to GSE was explained as:

“The wide proficiency ranges covered by each of the 6 CEFR levels (from A1 to C2) made it difficult for everybody to agree on the exact nature of each proficiency level. Considering the nature and difficulties of the language learning process, especially in a foreign language context, the inability to demonstrate how much progress has been achieved and how much more remains might be a demotivating factor. The time it takes for students to move up from one level to another varies greatly depending on their starting level, the amount of exposure to the language, their context, mother tongue, age, abilities and a range of other factors. For this reason, it is difficult to estimate how much time is needed to pass from one CEFR level to the next, especially in a context where input is mainly limited with the classroom boundaries. These limitations, in addition to the lack of clarity on how to interpret the CEFR levels, required searching for a different proficiency framework which resulted in the discovery of the Global Scale of English (GSE), a psychometric tool” (Aydin et al., 2017, p. 308-309).

The curriculum of the speaking course was designed based on the GSE (2015) Learning Objectives between 51-66 levels. 66 on the GSE proficiency scale, which corresponds to the initial stages of B2 in the CEFR was established as the optimum point to be reached by the end of the program. The reason why 66 was determined as an exit level was that “considering the entry-level and the length of time available for both in and out-of-class study, 66 was determined to be an achievable point on the GSE” (Aydin et al., 2017, p. 311).

Since the program aims to give general English from 51 to 66 on the GSE scale, it was therefore decided to take the range between 51-66 levels for the speaking skills as a basis for the test tasks developed in the present study. The fact that the GSE (2015) identifies language

proficiency in different levels and offers illustrative descriptors of “can-do” activities at each range of a proficiency level makes it a useful reference in task design especially when a specific range is targeted for a task. These descriptors are used in the study to guide the alignment of the tasks to the different proficiency levels.

1.3. Context Validity

Context validity, which is often named as content validity (i.e. Fulcher & Davidson, 2007) is related to the context coverage, relevance and representativeness. The contextual components of the test tasks in the study are examined based on the aspects of context validity for speaking proposed by Weir (2005). According to Weir (2005), it is “the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample” (p.19). This description implies that task characteristics and settings of tasks should reflect “performance conditions of the real-life context” as much as possible (Shaw & Weir, 2007, p.63).

Weir (2005) notes that in test development, various elements regarding the task and administration setting, as well as task demands in terms of linguistic characteristics and speakers should be taken into account to develop a theoretically sound basis for the choices made with respect to contextual features of the test tasks. Therefore, presenting as much evidence as possible for each of these elements will provide test developers with pieces of evidence to validate the choices they would like to make about test-takers based on their test performances. In this sense, the tasks developed for the purposes of the current study were ordered according to their assumed difficulty based on the GSE scale descriptors and targeted specific range in the scale (between 51-66 levels, see Section 2.3.2 for further information). The GSE scale descriptors provide an opportunity for producing a wide range of speech functions as describing, comparing, elaborating and expressing preferences, explaining, and justifying opinions. The current test taps into these various functions since different functions require different kinds of cognitive processing and may increase/decrease task difficulty (Galaczi & ffrench, 2011).

1.4. Scoring validity

Scoring validity is concerned with all test aspects that can influence scores’ reliability. Zainal Abidin (2006) highlights that scoring validity is an inevitable aspect of test validation procedure since the scores obtained from the tests may not be totally due to their performances, but influenced by other factors i.e. sources of error. Such problems of inconsistency can threaten the validity of the test and lead to the involvement of construct-irrelevant variance in the testing process. Therefore, it is identification and minimization of such errors of measurement that test developers should concern for the reliability of the scores produced by a test.

In testing speaking, rating is an important factor affecting the reliability of the test. It includes criteria/rating scale, rating procedure, raters, and grading and awarding. The chief concern in the testing of speaking, in this sense, is rater reliability and how scores are awarded based on a rating scale (Zainal Abidin, 2006). As for the investigation of the scoring validity in the study, it was examined how well the COPTEFL scores and the face-to-face speaking test scores are compared in terms of inter-rater reliability, order-effect and test scores.

1.5. The present study

When we review the research on the use of computers in oral proficiency testing, it is seen that the studies have focused largely on correlations or analyses of test outcomes/products including test scores (Jeong, 2003; Kiddle & Kormos, 2011; Öztekin, 2011; Thompson, Cox & Knapp, 2016) underlying constructs of test-taker language output in different modes of tests (Zhou, 2015), and test-taker reactions (i.e. attitudes) (Joo, 2008; Kenyon & Malabonga, 2001; Qian,

2009). As O’Loughlin (1997) states while these approaches have offered valuable insights into the comparability issue, there is a need to complement them with other perspectives as well. In particular, apart from investigating test outcomes or products, limited attention has been paid to “the examination of test design, test taking and rating processes and how an understanding of these components of assessment procedures may provide the basis for a more complex comparison between the two kinds of tests when combined with the analysis of test products” (p. 72). The perspective that O’Loughlin (1997) addressed above is the methodological approach taken in the current study. Particularly, apart from investigating comparability of test scores and test-taker attitudes obtained from open-ended questions, the study placed emphasis on the examination of the test development process including test design, development and administration stages. To date, such an approach has been seldom adopted in comparability research. Notable examples of studies combining a focus on process and product in testing speaking are O’Loughlin (1997) and Mousavi (2007). This study differs from O’Loughlin (1997) because its aim was not to compare a tape-based speaking test with a face-to-face speaking test. It also differs from Mousavi (2007) because its aim was not to compare a computer-based oral proficiency test with a face-to-face speaking test (International English Language Testing System, IELTS) that already was in use. Instead in the present study, first, a computer-based oral proficiency test format was developed and then a face-to-face version of the test was created to provide contrast and the test-method effect. By attending to both process and product, it was aimed to offer greater insight into construct validity and thus establish a stronger basis from which to compare test scores and attitudes towards both test delivery modes. The research questions that guided the present study are as follows:

1. How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability; (b) order-effect; (c) test scores?
2. What are the attitudes of test-takers in relation to the test delivery modes?

2. METHOD

2.1. Research context and participants

The study was conducted at Anadolu University School of Foreign Languages (AUSFL), Turkey. The school is a preparatory program that aims to equip the students with the necessary language skills in order to follow the academic education in their departments. The curriculum of the program is designed to help students to be able to reach that exit proficiency level required to be accepted as successful and constructed based on the GSE Learning Objectives (2015) between 51-66 levels. This test was designed for newly arrived students to AUSFL who have finished their high school or the students who are studying in preparatory schools of the university and have to take an exit exam to demonstrate that they have gained sufficient proficiency in English for academic study in their departments. The participants of the study were forty-five non-native speakers of English whose first language is Turkish. Participation in the study was on a voluntary basis.

2.2. Test development team

2.2.1. Item writers and raters

There were eight-item writers in the study. Item writers were also the raters. They are professional instructors who work full-time for the testing institution in AUSFL. They are experienced teachers of similar students and they have relevant experience for teaching speaking, writing speaking tasks for proficiency exams and assessing students’ speaking proficiency.

2.2.2. Editors

In the editing committee, there were four experts who were asked for opinions about the written

items. One of the experts was a professional item writer and rater who did not participate in producing items and scoring. Another one was experienced in teaching speaking and language testing in Anadolu University Foreign Languages Department. The others were subject experts, the researcher was one of them.

2.2.2. Software developers

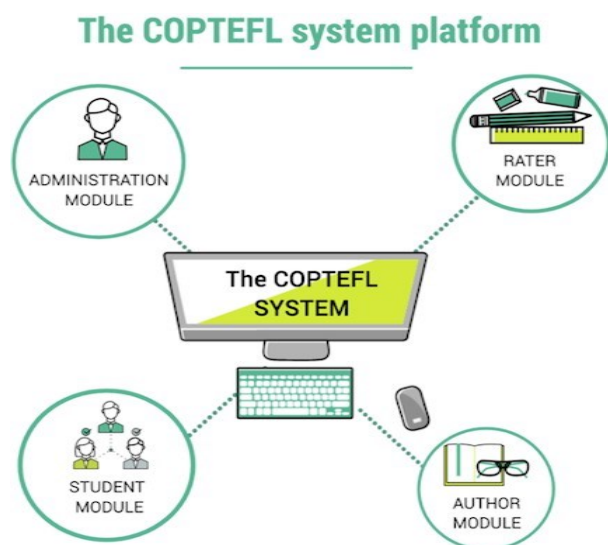
Two software developers, who have the necessary formal professional qualifications, developed the COPTEFL system. One of them worked for the programming of the system and the other worked for the web-page design.

2.3. Instruments

2.3.1. The COPTEFL system platform

The COPTEFL is a computer-based speaking test of general proficiency designed for adult learners of English as a Foreign Language (EFL). It uses the Web as its delivery medium. It was designed to offer users an alternative to face-to-face oral proficiency tests. The COPTEFL system platform comprises four types of users as (1) Administrator, (2) Author, (3) Rater, and (4) Student. Each user has its own module and is only allowed to access the information and functions they are given permission to access. In order to keep the system reliable, only the administrator(s) have access to other modules (see [Figure 2](#)).

Figure 2. *The COPTEFL system platform.*



In order to keep the system reliable, only the administrator(s) have access to other modules. Each module and its functions were explained below:

1. Administration Module:

This module is used by the administrator(s) in order to manage the system platform. Accessing this module allows direct access to all components of the COPTEFL system. The functionalities built into this module include:

- (a) Managing users (authors, raters and students) i.e. allowing or restricting user access, accessing the data of a user in particular, editing users' personal data, setting user deadlines (see [Figure 3](#)).

Figure 3. Administration module: Managing users.

COPT EFL
New Generation Oral Proficiency Test

Home

Students

Instructors

Test

Questions

Settings

Please fill in the textbox

Photo
Choose File No file chosen

Mail:

Phone:

Type:
Instructor

Save

Instructor List
Add Instructor
Send Message
All Messages

(b) Setting the time for the test and its announcement to the student module for the test-taker registration (see [Figure 4](#)).

Figure 4. Administration module: Setting time for the test.

COPT EFL
New Generation Oral Proficiency Test

Home

Students

Instructors

Test

Questions

Settings

Create Test

Name:

Date:

Time:
18:00

Student Count:



Last Registration Date:

Location:

Create Test

- (c) Allowing test to start on the scheduled time,
- (d) Monitoring the processes of test item creation and rating, and sending messages to the authors/raters,
- (e) Control over the written test items i.e. editing or deleting before they are included in the item pool by the system automatically (see [Figure 5](#)).

Figure 5. Administration module: Control panel for the written test items.

Part 1 Part 2 Part 3						
Number	Image	Label	Question	Writer	Proofreader	Proofreader Review
1		Daily life	Describe this picture and explain why it is important to have a good night's sleep.	Abdulkadir Durmuş	Proof Reader	Approved
2		Education - school life	Describe this picture and explain why it is important to have a university education.	Abdulkadir Durmuş	Proof Reader	Approved

- (f) Monitoring the test questions for each student before and after the administration of the test,
 (g) Changing the test questions before the test starts,
 (h) Accessing the students' answers to the questions during the test administration,
 (i) Accessing the test results (see [Figure 6](#)).

Figure 6. Administration module: Accessing to the test results.

17 Mayıs 5/17/2018		Time: 14:00 Location: c 107 Last Registration Date: 5/17/2018				
All						
E-mail	Part Number	Rater Grade	Inter Rater Grade	Proofreader Grade	Result Grade	Result
hasanalibuyuksahin@hotmail.com	Part 1	65	70	0	62	Passed
	Part 2	70	55	0		
	Part 3	55	55	0		
tekdemir1558@gmail.com	Part 1	55	85	75	60	Failed
	Part 2	55	85	70		
	Part 3	55	70	50		

2. Author Module:

This is the module used by the item writers to develop and edit tasks for the test item pool (see [Figure 7](#)).

Figure 7. Author module: Adding questions.

The screenshot shows the 'New question' form in the COPT EFL author module. The form is titled 'New question' and is located on the right side of the page. On the left side, there is a dark blue sidebar with navigation options: Home, Test Bank, Rating, and Review. The form itself has a teal header. Below the header, there are several input fields: 'Select Part' and 'Select Label' are dropdown menus with 'Please select...' as the placeholder text. The 'Photo' field has a 'Choose File' button, a '(500x500)' size indicator, and an 'Upload' button. The 'Question' field is a large text area. The 'Note' field is a smaller text area with '(If any)' written below it. At the bottom of the form, there is a 'Save' button.

The tasks written by the item writers are shown in the administrator’s module so that any necessary final changes can be made before the tasks become ready for the test. Item writers can only manage their own tasks and cannot access the tasks developed by other writers. However, they can see the total number of tasks for each part in the item pool.

3. Rater Module:

This is the module through which raters can monitor the students’ tests to score. These tests are assigned to raters by the computer automatically. Each rater is also an inter-rater. Raters do not know for which test they are assigned as rater or inter-rater. They just give the scores for each test that showed up on their module. Only administrator(s) can access the data of a person about for which test s/he was a rater or inter-rater. Scoring is anonymous on the system. The identities of the students remain confidential. Each task is delivered successively to score. Depending upon the extent of the discrepancy between scores, two or three rater scores were compared to get more accurate results. In AUSFL’s rating system if the scores by two raters are discrepant by more than ten points, a third rater independently scores. The score of one of the two raters whose score is close to the third rater is accepted as valid. This procedure is adapted in the ratings of the study. A proofreader who is the administrator gives scores as a third rater in order to reach a consensus in the ratings among two raters. The scores of the one whose scores are close to the proofreader’s are accepted as valid by the computer and therefore, the final score comes from the average score of these two raters.

In the development of the rating scale, an existing scale -which was developed by a formal testing body in AUSFL was adopted. But, it was modified since it involved “interaction” as a rating element. Because the tasks were monologic ones, “interaction” would be useless. The process in specifying the procedures for scoring started with expert judgments and evaluations. Appropriate changes were made based on those decisions by the experts and therefore, the tasks are decided to be scored according to the following assessment criteria: (1) Pronunciation, (2) Fluency, (3) Grammar range and accuracy, (4) Adequacy of vocabulary for purpose and (5) Task fulfillment. In the rater module, each criterion is shown with its explanation on the page

and raters see the criteria section while they are listening to the answers. They can give scores at the same time they are listening to (see Figure 8).

Figure 8. Rater module: Giving scores.

The screenshot shows the 'Please Listen to Answer' and 'Please Rate' sections of the rater module. The 'Please Listen to Answer' section includes an image of four people socializing, a question asking to describe the picture and explain why it is important to make friends, and a 'Current point = 65' indicator. The 'Please Rate' section shows two criteria: 'Task fulfillment' and 'Fluency', each with four rating options (1-4) and their descriptions. The '3 - GOOD' option is selected for both criteria.

When they complete marking, the computer shows the total grade that a student gets after the scoring process and then the rater can submit the score (see Figure 9).

Figure 9. Rater module: Completing marking.

The screenshot shows a table of completed marking tasks. The table has columns for Image, Part Number, Question, Status, Point, and Rating. The 'Status' column shows 'Rated' for all three tasks. The 'Point' column shows 60, 65, and 70 for the three tasks respectively. The 'Rating' column shows a green arrow icon for each task. Below the table, there is a 'Current point = 65' indicator and a 'SUBMIT' button.

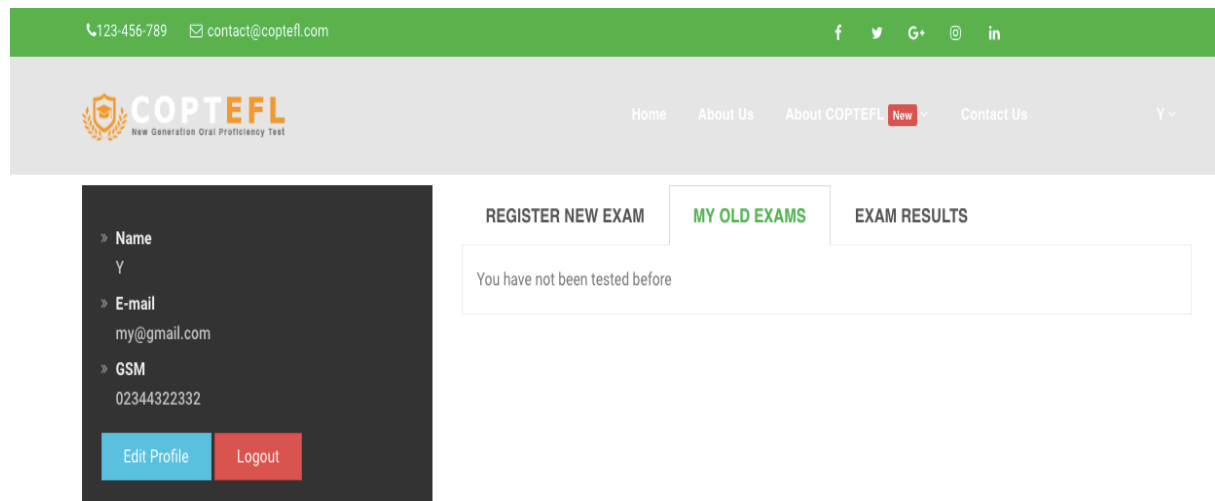
Image	Part Number	Question	Status	Point	Rating
	1	Describe this picture and explain why it is important to have a good night's sleep.	Rated	60	↩
	2	Look at these two pictures and explain whether the fathers or the mothers should spend more time with their children.	Rated	65	↩
	3	1. Talk about an unforgettable moment in your life. 2. How did it affect your life? 3. Do unforgettable moments shape your future life?	Rated	70	↩

Current point = 65 SUBMIT

4. Student Module:

This module has two functions: (1) to deliver the test and (2) to publish the test scores achieved by the students. The students who registered to the system log into the delivery platform in order to start their test. When raters and inter-raters submit their scores, the computer averages the grades and publishes them on students' own page. The students can log into their accounts and see the grades they get for each criterion and their final grade on this module (see [Figure 10](#)).

Figure 10. Student module.



2.3.2. Speaking tasks

The speaking tasks used in the present study were created by the item writers. Three speaking tasks were developed in order to evaluate students' general proficiency with regard to oral competency. Since the test was intended to be used as a general proficiency test, the tasks were prepared according to the Global Scale of English (GSE) 'can do' statements from 51 to 66 levels (see some examples from the range 51-66 below):

- 51 Expressing and responding to feelings (e.g. surprise, happiness, interest, indifference).
- 53 Comparing and contrast alternatives about what to do, where to go, etc.
- 60 Justifying a viewpoint on a topical issue by discussing pros and cons of various options.
- 62 Constructing a chain of reasoned argument.
- 66 Developing a clear argument with supporting subsidiary points and relevant examples.

The COPTEFL included monologic tasks that can elicit individual discourses without the test-takers' interacting with an interlocutor. These tasks were discourse type tasks. The first task was a description and giving an opinion task. In this task, students were required to describe the picture and then give/express/justify an opinion related to the picture. The second task was a comparison task. In this task, students were asked to look at two pictures and choose one and provide a reason for their choice. The final task was a discussion task in which students were required to justify a viewpoint.

The final version of the task types was based on the pilot test. In this phase, the prototype tasks were piloted in a face-to-face test with small groups of learners in order to find out which tasks do not work as planned, and which should be included or excluded after revision. Announcements explaining that students had the chance of testing their speaking skills were made at the school. Six volunteer AUSFL students participated in the study. The researcher conducted the test and rated for the scores. The analysis showed that some of the questions might elicit a small range of language and repeated answers from the students. The presence of

general questions such as “Why is it important to celebrate national holidays?” provided plenty of scope for answers. This, in turn, caused the detailed question related to the general one such as “Talk about a national holiday celebrated in your country” to be covered in advance. Therefore, the necessary changes and modifications were made after discussing problematic items with the team of test writers. According to this revision, test tasks were redesigned. Item writers and editors were asked for their opinions with regard to the final version of the task types. After getting their approval, test tasks were written. The editors edited the written tasks and the final version of the tasks was entered into the COPTEFL system. After this process, the COPTEFL system was ready to test.

2.3.3. Face-to-face speaking test

The tasks and the instructions used in the face-to-face speaking test were the same as the COPTEFL (for an example task for the face-to-face speaking test see [Figure 11](#)).

Figure 11. An example for Task 2.

PART 2: I will ask you to compare two pictures, choose one and provide reasons for your choice. You will have 15 seconds to think about your answer and 90 seconds to reply.

Look at these two pictures and talk about which of these travelling ways you would prefer to choose.



2.3.4. Open-ended questions

Open-ended questions were designed according to the opinions of the two subject experts. Since the aim of the study was to find out the usability of the COPTEFL in comparison with a face to face equivalent, the questions were targeted to depict attitudes of students towards the COPTEFL system and its perceived advantages and disadvantages with regard to a face to face speaking test. Therefore, open-ended questions were used for investigating test-taker attitudes towards testing speaking in the COPTEFL and the face-to-face mode. Test-takers were first asked to evaluate the COPTEFL system and then, state their preferences by making comparisons between the two modes. The questions were:

1. How do you evaluate the COPTEFL system as a speaking test delivery medium?
2. What are the advantages and disadvantages of the COPTEFL when compared to the face-to-face speaking test?
3. What are the advantages and disadvantages of the face-to-face speaking test when compared to the COPTEFL?

2.4. Data collection procedure

2.4.1. A priori construct validation: Processes followed in the development of the tests

In an effort to adapt the best practices in language test development in the present study, Bachman & Palmer’s (1996) stages in test development were followed with slight modifications in order to more suitably fit the purposes of the research. In their book, test

development was organized into three stages as design, operationalization and administration.

1. Test design and operationalization

(a) Developing test task specifications: The purpose for which the test would be used and the target population for the test were our starting point in designing test specifications. Having determined this, relevant literature was reviewed in order to find out what language would be needed by test candidates in the case of oral proficiency tests. From this consultation, similar tasks and texts were sampled to arrive at a manageable test design. Then, with the help of the eminent experts in the field, draft specifications and sample tasks within which the test might be constructed were designed. Since the principal user is probably the test writers, the team of test writers was then asked for their opinions about whether the draft specifications and sample tasks were appropriate for the purposes of the test and the target population. They revised the tasks and specifications, and discussed whether the tasks would work, that is whether each task which was intended to assess a particular aim actually would do so. Many of the responses to whether tasks would work or not gave the impression that a trial on a small group of learners who are similar in background and language level to the target population would provide helpful insights in understanding the kind of language being elicited for each task.

Trialing for test tasks: The development of the speaking tasks process consisted of various stages such as the selection of task types, writing of task items, consulting with experts, and pilot tests. The final version of the task types was based on the pilot test. In this phase, the prototype tasks were piloted in a face-to-face test with small groups of learners in order to find out which tasks do not work as planned, and which should be included or excluded after revision. Announcements explaining that students had the chance of testing their speaking skills were made at the school. Six volunteer preparatory program students participated in the study. The researcher conducted the test and rated for the scores. The analysis showed that some of the questions might elicit a small range of language and repeated answers from the students. The presence of general questions such as “Why is it important to celebrate national holidays?” provided plenty of scope for answers. This, in turn, caused the detailed question related to the general one such as “Talk about a national holiday celebrated in your country” to be covered in advance. The result was that, however thoughtfully designed to avoid pitfalls, some of the questions failed to elicit the targeted responses. Therefore, the necessary changes and modifications were made after discussing problematic items with the team of test writers. According to this revision, test task specifications were redesigned to generate test tasks.

Once a coherent system was created for specifications and tasks whose parts fitted together, item writers and editors were asked for their opinions with regard to the final version of the task types. After getting their approval, test tasks were written. The editors edited the written tasks and the final version of the tasks was entered into the COPTEFL system. After this process, the COPTEFL system was ready to test.

(b) Writing test tasks and instructions: After making explicit any constraints in test design, test writers began writing tasks and instructions with the test’s specifications. The writers needed to find suitable communication activities for the tasks such as expressing an opinion on an issue, a view by contrasting it with other possible views, or discussing ideas. The writers also needed to find pictures that serve the purpose of the task. After completing the test writing process, each writer made responsible for editing another writer’s set of tasks. Once their editing process concluded, tasks became subject to a number of reviews before they reached their final draft stage. Two editors revised the items and assembled them into a draft test paper for the consideration of other editors. These editors examined each item for the degree of match with the test task specifications, ambiguities in the wording of the items, and match between the questions and pictures. The changes made after editing processes were reported and shared with the team of test writers.

(c) Specifying the procedures for scoring: In this process, considerable effort was put into developing a practical analytic scale for decision making in which there is less to read and remember than in a complicated descriptor with many criteria and unfamiliar technical terminology. Once the scale was modified, it was then refined by raters who use it so that they understand the meaning of the levels with regard to each particular feature in the scale.

(d) Software development: In developing testing software, there are some key requirements of the standard steps taken by the researchers (Mousavi, 2007; Shneiderman, 2004; Zak, 2001) as (1) Analysing (defining the problem), (2) Choosing the interface, (3) Coding, (4) Test and debug, and (5) Completing the documentation. In this study, the same standard steps were followed. These steps were explained briefly below.

STEP 1: Analysing (defining the problem): This pertains to the definition of a problem in any research project. In this step, a statement of the problem was presented to provide guidance to the rest of the programming steps. That is to say, what exactly the programmer wants to achieve with this programming was stated in this step. The core problem that drew this study was the low degree of the practicality of administering oral proficiency tests to a large group of preparatory program students through the use of a live face-to-face interview. In this step, meetings with the administrator and instructors were held in order to determine the requirements of the COPTEFL system i.e. who would be the rater, whether raters would write items, what to include item writing and rating pages in the system. These were general questions that were answered during the analyzing phase. In order to translate requirements into design, meetings between the researcher and the software developer were held twice a week. They analyzed the requirements of the system for the possibility of incorporating to the COPTEFL system program.

STEP 2: Choosing the interface: The interface of any computerized test involves the actual objects the test-takers see and deal with during a testing session. These objects may consist of videos, text boxes, command buttons, animations, progress bars, date/time indicators, and so on. Here, the key to developing a good user interface is to have a complete understanding of the target user (Luther, 1992). As suggested in Mousavi (2007), this understanding may be achieved by a process referred to as user task analysis where the developer assumes himself/herself as the target user and identifies a series of possible scenarios to come up with the most convenient one. The relationship between the application interface and test-takers is important because, as Fulcher (2003) states interface design can be the threat of interface-related construct irrelevant variance in test scores, and therefore should be avoided. For this purpose, he identifies a principled approach in the development of a good interface design. This approach includes three phases as (1) planning and initial design, (2) usability testing, and (3) field testing and fine-tuning. The present study followed these phases in choosing the interface. Each was explained below.

Phase 1: Planning and initial design. This involved hardware and software considerations, navigation options, page layout, terminology, text, color, toolbars and controls, icons, and the rest of the visible objects on a typical computerized test.

Phase 2: Usability testing. This included activities such as searching for problems and solutions, selecting test-takers for usability studies, item writing and banking, pre-testing, try-out for scoring rubrics.

Phase 3: Field testing and fine-tuning. This consisted of try-out for the interface with a group of samples drawn from the target test-taking population and also making sure that the logistics of data collection, submission, scoring, distribution and retrieval, and feedback would work as planned. This phase provided an opportunity to trial and test for (possible) variation in the appearance of the interface across sites, machines, platforms, and operating systems.

STEP 3: Coding: Coding is the translation of the algorithm into a programming language. It is generally the most complex and time-consuming step in the development of the computerized tests. Once the planning of the application and the building of the user interface were complete, programming instructions were written to direct the objects in the interface on how to respond to events. After deciding on the system design requirements, the COPTEFL system was divided into four modules as (a) administration module, (b) author module, (c) rater module, and (d) student module. Coding was developed according to the modules. The code was developed based on the needs of the program from scratch, and this stage was the most challenging part and took the longest time.

STEP 4: Test and debug: Debugging is the process of tracking down and removing any errors in the computer program. Errors in a computer program could be the result of typing mistakes, flaws in the algorithms, or incorrect use of the computer language rules. Testing and debugging step was an inevitable part of the operation because of the complexity in the coding phase of the programming as well as the possible persistence of syntactic anomalies in the programming language. Caution must be exercised at this stage for possible problems. After the code was developed, the system went through a pilot study to see if it was functioning properly. The researcher and the developer assessed the software for errors and document bugs if there were any. The developer did the necessary changes to the system due to the results.

STEP 5: Completing the documentation (or distributing the application): This step included developing an installable setup file along with all its components for new users and new platforms. That is, all the materials that described the program were compiled to allow other people, involving test users, raters, administrators, and item writers to understand the scope of the program and what it does. Distributing the application made it possible to run the application on different platforms and with different operating systems and to secure the compiled files, projects, and codes. So, this stage was the try-out stage for the COPTEFL system. It was passed over to the users to get feedback. Any bugs and glitches experienced during this stage were fixed.

2. Test administration:

This stage consists of two phases:

(a) **Try-out phase:** After the development of the software, the next step was to test it with users. 15 students who consented to participate were included in the trial. These were the students having their regular laboratory classes as part of their language program. In this phase, we gathered information on the usefulness of the test itself and for the improvement of the test and testing procedures.

(b) **Operational testing:** In this phase, the aim was to gather information on the usefulness of the test, but this time administering the test involved the goal to accomplish the specified use/purpose of the test. A total of 45 volunteer preparatory program students from various proficiency levels participated. One week before the administration, test-takers were divided into groups, each of which takes portions of the test at different times. One group was tested by the COPTEFL first, and then interviewed in the face-to-face speaking test, and a second group was provided the face-to-face speaking test first and then being tested by the COPTEFL. The test-takers who took the COPTEFL first were asked to take the face-to-face speaking test after 3 weeks and the test-takers who took the face-to-face speaking test first were asked to take the COPTEFL after 3 weeks. With a counterbalanced design like this, it was aimed to find out whether the test in one mode followed by the other could affect the score for the second mode. The following section described the step-by-step procedure of the program, the COPTEFL, and its administration:

STEP 1: The COPTEFL web page was loaded: www.coptefl.com

STEP 2: Test-takers registered and logged in to the testing system.

STEP 3: Microphones were set up and tested.

STEP 4: Once the testing program started, the test-takers were presented with a short introductory page. In this introductory screen, the speaker welcomed the test-takers and introduced the test, providing information about its steps, format, function, procedure and length.

STEP 5: As soon as the test-takers listened to the instructions and clicked on the “next page” button, testing started.

STEP 6: Once the test-takers responded to all tasks, a final page was shown to express appreciation for taking the test. At this point, the program terminated and the test-takers exited the program.

After the administration of the tests, open-ended questions were given immediately in order to explore test-takers’ attitudes towards the test modes. After they completed writing, the researcher collected the answer sheets.

2.5. Data analysis

2.5.1. Score comparability

In order to evaluate the consistency between judges’ ratings, inter-rater reliabilities were calculated for each test delivery mode. To investigate the inter-rater reliability, a two-way random absolute agreement intra-class correlation coefficient (ICC) was performed. ICC assesses the consistency between judges’ ratings of a group of test-takers. Before proceeding to compare the magnitude of raw scores, the order effect on test scores was examined. To assess the effect of delivery mode and the mode-by-order interaction statistically, an Analysis of Covariance (ANCOVA) with within-subject effect was run on total scores.

For the investigation of the equivalence of test scores across modes, the magnitude of raw scores was compared by means of the two-way random absolute agreement intra-class correlation coefficient (ICC).

For better interpretations of the findings, mean score differences were also taken into account and therefore, mean scores by each test delivery mode were compared for total scores and task types. A paired samples t-test was run to compare the mean scores between the COPTEFL and the face-to-face speaking test.

2.5.2. Test-taker attitudes

The answers of the test-takers were classified into themes each of which represented an idea related to their attitudes towards test conditions. After that, certain themes based on the ideas were coded and then placed into the categories each of which represented with. Finally, emerging themes were expressed in frequencies. In order to reduce research bias and establish the reliability of research findings, an expert from Anadolu University, who studies in the English Language Teaching Department as a research assistant and has experience in qualitative data analysis analyzed the data. To ensure the credibility of the findings, the consistency between the emerging findings from two researchers was investigated. Similarities and differences across the categories identified were sought out. After member checking sessions and rigorous discussions between the researchers, a final consensus on the categories was achieved. How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability, (b) order-effect, and (c) test scores?

3. RESULT / FINDINGS

3.1. Research Question 1

How well are the COPTEFL scores and the face-to-face speaking test scores compared in terms of (a) inter-rater reliability, (b) order-effect, and (c) test scores?

3.1.1. Inter-rater reliability

ICC results for overall test scores and each task type test scores across delivery modes were provided in [Table 1](#) below.

Table 1. Intra-class correlation coefficient (ICC) results for inter-rater reliabilities across test delivery modes

Test score	<u>COPTEFL</u> ICC	<u>Face-to-Face</u> ICC
Overall score	.806*	.889*

* significant at the .01 level

[Table 1](#) presented the results of the ICC on overall test scores for both delivery modes. As shown in the table, the ICC score for the face-to-face speaking test (ICC= .889) was slightly higher than the score in COPTEFL (ICC= .806). The average measure ICC for the COPTEFL was .806 with a 95% confidence interval from .64 to .89 ($F(44,44)=5.075, p<.001$) and the average measure ICC for the face-to-face speaking test was .889 with a 95% confidence interval from .80 to .93 ($F(44,44)=9.033, p<.001$). These ICC values between .75 and .90 indicated good reliability for both tests (Larsen-Hall, 2010). [Table 2](#) revealed the ICC results for each task type (see [Table 2](#)).

Table 2. Intra-class correlation coefficient (ICC) results of inter-rater reliabilities for each task type across test delivery modes.

Task type	<u>COPTEFL</u> ICC	<u>Face-to-face</u> ICC
Opinion	.686*	.796*
Comparison	.702*	.842*
Discussion	.772*	.890*

* significant at the .01 level

For the COPTEFL, the findings showed that the average measure ICC for the opinion task was .688 with a 95% confidence interval from .42 to .82 ($F(44,44)=3.151, p<.001$), for the comparison task, it was .702 with a 95% confidence interval from .45 to .83 ($F(44,44)=3.317, p<.001$), and finally, for the discussion task, it was .772 with a 95% confidence interval from .58 to .87 ($F(44,44)=4.320, p<.001$). As for the face-to-face speaking test, the findings showed that the average measure ICC for the opinion task was .796 with a 95% confidence interval from .63 to .88 ($F(44,44)=4.943, p<.001$), for the comparison task, it was .842 with a 95% confidence interval from .71 to .91 ($F(44,44)=6.260, p<.001$), and finally, for the discussion task, it was .890 with a 95% confidence interval from .79 to .94 ($F(44,44)=8.913, p<.001$).

These results indicate a significant direct relationship between inter-rater reliability scores for each task type across test delivery modes that those who get higher scores from a task in the COPTEFL by the rater also get higher scores from the same task in the face-to-face speaking test by the inter-rater or vice versa.

For the COPTEFL, the findings revealed that the ICC values for opinion and comparison tasks were between .50 and .75, meaning that the level of reliability was moderate. For the discussion

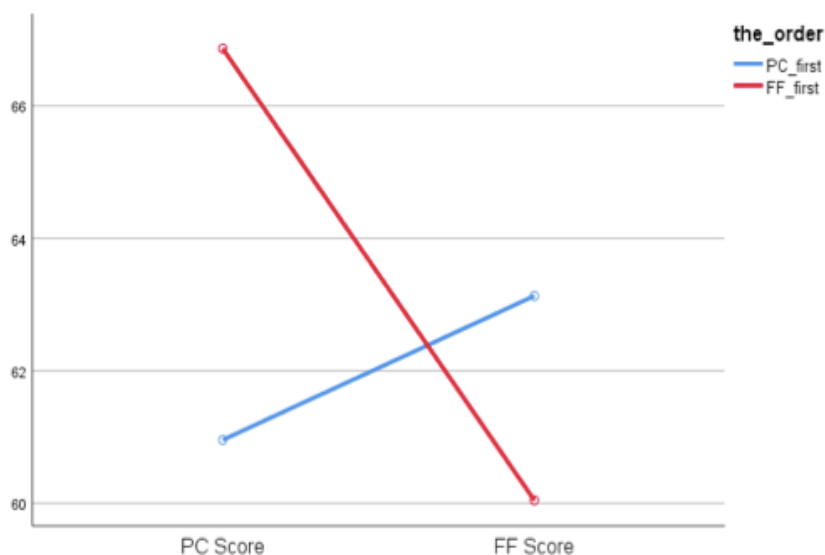
task, it was .77, which indicated good reliability. As for the face-to-face speaking test, each ICC value was between .75 and .90, revealing that the level of reliability was good.

3.1.2. Order-effect

Prior to comparing raw scores awarded to the two delivery modes, an Analysis of Covariance (ANCOVA) was computed to assess whether the existence of order has an effect on speaking test scores. The findings revealed that there is a significant interaction between test order and test scores ($F(1,43)=6.89, p=.012$).

As Figure 12 shows, the groups which took the COPTEFL first did better on the face-to-face speaking test ($M=63.13, SD=11.83$) than on the COPTEFL ($M=60.96, SD=11.82$); and the groups which took the face-to-face speaking test first did better on the COPTEFL ($M=66.86, SD=11.77$) than on the face-to-face speaking test ($M=60.05, SD=11.39$) independent of their level.

Figure 12. The interaction between group and test mode.



The results suggest that both groups did better in their second test than in their first test no matter which type of test they took first, which shows that there was a practice effect in general.

3.1.3. Comparability of raw scores

The analyses in this part focused on the differences in test scores between the COPTEFL and the face-to-face speaking test. In order to determine differences, ICC and paired samples t-tests were computed across delivery modes.

a. The relationship between scores by test delivery modes:

A two-way random absolute agreement ICC was conducted in order to find out how well the COPTEFL scores and the face-to-face speaking test scores were correlated. The analyses were run for total scores and task type scores across test delivery modes (see Table 3).

A significant moderate degree of ICC was found between the COPTEFL total scores and face-to-face speaking test total scores. The average measure ICC was .632 with a 95% confidence interval from .33 to .79 ($F(44,44)=2.735, p<.001$). This result indicates a direct relationship between the scores across test delivery modes that those who get higher scores from COPTEFL also get higher scores from the face-to-face speaking test, or vice versa.

Table 3. ICC results for total scores and task type scores across test delivery modes.

	Total score	COPTEFL		
		Opinion task	Comparison task	Discussion task
<u>Face-to-face</u>				
Total score	.632*			
Opinion task	-	.382		
Comparison task	-	-	.576**	
Discussion task	-	-	-	.704*

* significant at the .01 level

** significant at the .05 level

As for the scores from each task type, a low degree of ICC was found between the COPTEFL opinion task scores and the face-to-face speaking test opinion task scores. The average measure ICC was .382 with a 95% confidence interval from -.09 to .65 ($F(44,44)=1.648, p>.05$). For the comparison task scores, there was a significant moderate degree of absolute agreement ICC between the COPTEFL and the face-to-face speaking test. The average measure ICC was .576 with a 95% confidence interval from .22 to .76 ($F(44,44)=2.347, p<.05$), which indicates that those who get higher scores from the comparison task in the COPTEFL also get higher scores from the comparison task in the face-to-face speaking test, or vice versa. Finally, as for the discussion task scores, the results revealed a significant moderate degree of ICC between the COPTEFL and the face-to-face speaking test. The average measure ICC was .704 with a 95% confidence interval from .45 to .83 ($F(44,44)=3.333, p<.001$), which shows that those who get higher scores from the discussion task in the COPTEFL also get higher scores from the discussion task in the face-to-face speaking test, or vice versa.

b. Comparing the mean scores by test delivery modes:

For better interpretations of the findings, mean score differences were also taken into account and therefore, mean scores by each test delivery mode were compared for total scores and task types. A paired samples t-test was run to compare the mean scores between the COPTEFL and the face-to-face speaking test (see Table 4).

Table 4. Paired samples t-test results for each task type across test delivery modes.

	COPTEFL		Face-to-face		df	t	p
	Mean	SD	Mean	SD			
Total score	63.84	12.04	61.62	11.59	44	1.219	.229
Opinion task	65.36	13.57	61.07	12.28	44	1.808	.077
Comparison task	63.80	12.66	62.27	12.68	44	0.743	.462
Discussion task	62.82	13.85	62.33	12.57	44	0.258	.798

As Table 4 presented, the findings showed that there is not a statistically significant difference between the COPTEFL and the face-to-face speaking total test scores ($t(44)= 1.219; p>.05$). When the mean scores of each task type were investigated, the findings showed that there is not a statistically significant difference between the COPTEFL scores and the face-to-face speaking test scores for task types, which are the opinion task ($t(44)= 1.808; p>.05$); the comparison task ($t(44)= 0.743; p>.05$), and the discussion task ($t(44)= 0.258; p>.05$). Therefore, it can be concluded that test delivery mode was found not to have a significant effect on test-takers' speaking test scores.

3.2. Research Question 2

What are the attitudes of test-takers in relation to the test delivery modes?

3.2.1. General attitudes towards the COPTEFL

To explore the face validity of the COPTEFL from the test-takers' perspective, an open-ended question assessing participants' attitudes towards the COPTEFL was posed. The findings are presented in order of frequency below (see Table 5).

Table 5. Test-takers' general attitudes towards the COPTEFL.

Positive comments on COPTEFL		Negative comments on COPTEFL	
Categories	Num.**	Categories	Num.**
1. Test system	42	1. Test system	11
User friendly		Weak microphones	
Well designed		Abrupt transition between	
Easy to start and follow		instructions and tasks	
Easier to understand the pronunciation		The effect of the countdown	
Practical		timer on the screen	
Good sound quality			
2. Tasks	10	2. Tasks	2
At the right level of difficulty		Tough questions	
3. Time limit	11	3. Time limit	13
Enough time to give answers		Little answering time	
Enough time to think about answers		Little thinking time	

**The number of the comments by test-takers

Nearly all of the participants stated that the test system was well designed, quite easy to operate, and works well:

"The COPTEFL system was quite easy to access and operate." P3

"I think the COPTEFL is much more practical than the face-to-face speaking test. It does not require teachers to interview and this saves time for them." P12

According to them, the system was user friendly and provided practical experience for test-takers and teachers. Some of them reported that the sound quality was satisfactory and they had no difficulty in hearing or understanding the instructions or the questions:

"In face-to-face speaking tests, it is sometimes difficult to understand the interviewer's pronunciation. In the computerized test, on the other hand, the correctness of pronunciation was controlled beforehand, and the questions were shown on the screen. This made the tasks clear to understand." P8

Although most of the comments were positive in relation to the test system, there were a few constructive comments for improvement of the system or the tasks:

"It would be better if the time limit for speaking was longer." P16

"Tasks were tough, so the number of tasks could be reduced or the time limit could be multiplied." P24

In some of the comments, test-takers stated that microphones could be better in order to achieve the best possible results for sound recording. Also, some of them referred to the bad effects of seeing countdown timer on the screen:

"Countdown-timer made me feel nervous." P7

Apart from those, one of the participants also made a comment on the transition from instructions to tasks. According to her, that transition was abrupt and due to this, she got anxious.

3.2.2. The direct comparison of two modes

To better understand test-takers' test method preferences, the responses to the second open-ended question were analyzed. Of those analyses, three categories were developed based on the comments that favored the face-to-face mode. These were presented in order of frequency below (Table 6).

Table 6. The direct comparison of two modes.

Attitudes towards the face-to-face speaking test		Attitudes towards the COPTEFL	
Categories	Num.**	Categories	Num.**
1. Interaction	15	1. Less anxiety	29
2. Naturalness	4	2. Better control	4
3. More time	1	3. Test fairness	2

** The number of the comments by test-takers

The main reason test-takers had more favorable attitudes to the face-to-face speaking test was the interaction with the interviewer. The participants remarked that they performed better on this test since the reactions from the interviewer such as smiling and nodding helped them feel comfortable and relaxed. Although the interviewers did not assist, test-takers were still trying to figure out whether or not they were being understood thanks to the facial expressions of the interviewers. According to them, non-verbal communication should be involved in an examination atmosphere, because no reactions could make them unable to gauge how far they came to the correct answer:

“During the exam, I would prefer to have feedback from the teacher to get certain about the correctness of my responses. So, I would prefer face to face speaking exams.” P17

“Interaction with the teacher helped me speak more.” P3

Some of the participants stated that it felt more natural to talk in the presence of the interviewer since it was similar to a real-life conversation where the communication is between two or more people. Two of them perceived the face-to-face speaking test as a better measure of their spoken English because of this sense of naturalness. Although some of the test-takers preferred having a conversation with an interviewer who could accommodate their responses and the use of time, the opposite was also true for some others who preferred the COPTEFL due to lack of influence of the interviewer and the use of time:

“I got nervous in face-to-face exams. But, the COPTEFL made me feel relaxed since I was testing myself alone.” P3

“There would be no influence of the interviewer who was faced with a problem just before the exam and reflected it on us.” P8

“Sometimes the way interviewers behave in the test makes me nervous. But, in the COPTEFL, there is not such a problem.” P15

In conclusion, the quantitative data showed that the test-takers performed better on the COPTEFL compared to the face-to-face speaking test ($M= 63.84$, $M= 61.62$, respectively). The qualitative data provided insights into the attitudes of test-takers in relation to both test modes and revealed that if given choice, many of them preferred the face-to-face speaking test due to the opportunity of interaction with the interviewer while some of them have a strong preference for the COPTEFL due to its provoking less anxiety. These results showed that different types of learners have different testing experiences and thus preferred either the COPTEFL or the face-to-face speaking test.

4. DISCUSSION and CONCLUSION

The subjectivity in the rater judgments is one of the major sources of measurement error and a threat to the reliability and validity of test scores (Bachman, Lynch & Mason, 1995). Inter-rater reliability estimates in the present study were .80 (ICC) for the computer mode and .88 (ICC) for the face-to-face mode, indicating that the level of reliability for each mode was good. One possible interpretation of this result is that a potential problem of inconsistency in different raters' scores was effectively controlled. The issue of experience at this point is considered as "the most important reason for rating scales appearing to be meaningful and providing reliable results" (p.97). In the present study, experts in testing speaking rated to an existing scale developed by a formal testing body in AUSFL. In order to achieve a common standard –which no one would wish to disagree with, rater training and socialization into the use of the scale were valued in the study. Such training was perceived as the way to ensure greater reliability and validity of scores produced in language performance tests (Fulcher, 2014). With rater training sessions, it was intended to "socialize raters into a common understanding of the scale descriptors, and train them to apply these consistently in operational speaking tests" (Fulcher, 2014, p.145). These efforts could possibly lead to the achievement of good inter-rater reliability scores for both tests. Having two raters instead of one might also be one of the reasons for achieving good reliability. As argued in Fulcher (2014), the use of a double rating can avoid the potential effect that an individual rater may have on the test score. That is, multiple ratings of each performance help minimize the subjectivity of ratings and therefore, improve reliability (Carr, 2011).

Both groups did better on their second test than on their first test no matter which test mode they took first. This finding revealed that there was a practice effect in general. The question then arises as to why there was a practice effect. One possible explanation lies in that the group who took the COPTEFL first performed poorly in the COPTEFL mode because they had never taken a speaking test in a CBT mode and therefore, might not achieve the best performance due to their unfamiliarity with the test format. When provided the opportunity to take a second test, they might demonstrate better performance. Similar to those who took the COPTEFL first, test-takers who took the face-to-face speaking test first might achieve a higher score on the second test since they became familiar with the test content. In line with this finding, Öztekin (2011) also reported a test order effect in her study results in which both groups did better on their second test than on their first test no matter which type of test they took first. Similarly, Zhou (2009) revealed that the test order effect was present in the findings of the study. But, this time, the group who took the computer-based speaking test first performed better on the later face-to-face speaking test, whereas the group who took the face-to-face speaking test first did not perform better on the computer-based test. In relation to this finding, Zhou (2009) states that the reason behind this finding may lie perhaps in the reactions from the interviewer. Accordingly, during the face-to-face speaking test, the reactions from the interviewer might have motivated the test-takers to give better verbal responses to the tasks and therefore they may have felt encouraged to do their best by the presence of the interviewer.

The lack of statistically significant differences between the mean scores indicated that test-takers who did well on the COPTEFL did almost equally well on the face-to-face speaking test and there was no major change in test-takers' performance on monologic speaking tasks when the response was elicited through non-human elicitation techniques. This finding suggested that test delivery mode did not account for the variance in test scores in the present study. With regard to the comparisons between the computer mode and the face-to-face mode, some studies have also shown a considerable overlap between delivery modes for speaking tests, at least in the correlational coefficient sense that test-takers who score high in one mode also score high in the other or vice versa (Mousavi, 2007; Thompson, Cox & Knapp, 2016; Zhou, 2015). This

research was correlational in nature and the correlation coefficient between the test modes was found to be .63, which is considered a moderate index of reliability. In conformity with the traditional requirements for concurrent validation (Alderson, Clapham & Wall, 1995), a correlation coefficient of .9 or higher is indicated to be the appropriate level of standards at which test users could consider “the semi-direct testing results closely indicative of probable examinee performance on the more direct measures” (Clark, 1979, p.40). Even though the correlation coefficient score in the present study was lower than the figure of .9, as Mousavi (2007) put forward for a figure of .63, the finding highlighted the usability of the newly developed prototype test of oral proficiency as a reasonable alternative mode of test delivery.

One possible explanation of the findings in the study might be that test-takers performed similarly in both test delivery modes and small differences between individual scores that are statistically non-significant might not be detected by using the paired-sample t-test. As Kiddle and Kormos (2011) report, the correlational analysis measures the strength of the relationship between the two delivery modes and high correlations might be accomplished even if the test-takers score differently in the two modes. If, for example, test-takers were consistently awarded higher scores in the face-to-face mode than in the computer mode, correlations can still remain high. At this point, Kiddle and Kormos (2011) suggest further empirical analysis such as Rasch analysis that

“Unlike analyses such as t-tests and correlations, the Rasch analysis does not rely on raw scores but uses logit scores instead, and consequently can yield reliable information on whether the fact that the test was administered under different conditions has an effect on test performance” (p.353).

As for the interpretation of the lack of significant differences in the mean scores across modes, one of the reasons might be that test-takers performed differently between two modes, but raters tended to award similar scores to the test-takers across tasks based on their overall impression about them on a particular task or the overall test. Gülle (2015), for example, pointed out that raters might show a tendency to assign similar scores to the test-takers across tasks due to their holistic judgments. In the current study, in order to minimize possible halo effects, the raters were assigned the scoring criteria for each task separately. Instead of rating one test-taker on all three tasks, they were asked to award scores for the test-takers’ performances on the first task and then continue with the second task and the third task. However, as Gülle (2015) states, it is still possible for the raters to assign similar scores across different tasks based on their overall judgments of the test-taker performances. The present results may also be attributed to the possibility that test-takers performed differently between two tests, but not to extent that the raters were able to discern due to the little difference between bands on a given subscale.

Qualitative analysis of the data revealed that test-takers had favorable attitudes towards the COPTFL in many aspects and the majority of them did not show a particular preference in terms of the testing modes. But, it appears from the responses, if given the choice, most of the test-takers were found to prefer the face-to-face speaking test. However, this finding did not necessarily imply that their reactions to the COPTFL were negative. This finding corroborated with the finding of Qian (2009) who also found that participants did not have a particular preference with respect to the testing modes, and only partially corroborated with the findings of most researchers including McNamara (1987); Shohamy, Donitsa-Schmidt, and Waizer (1993) and Joo (2008), who all found that an overwhelming of participants showed a particular preference to the direct testing mode. The finding of the current study was at odds with Brown’s (1993) study that test-takers preferred semi-direct testing mode to the direct testing one. At this point, as Qian (2009) suggested that we should

“be cautious about drawing a conclusion as to which testing mode is more amenable to test-takers as their preferences might be test and context dependent: Test-takers’ attitudes may be

influenced by various factors, such as test quality, the stakes of the test to the test-taker, test-takers' cultural traditions and personalities, and so forth" (p.123).

4.1. Limitations of the study

The findings of the study must be considered in the context of several potential limitations and therefore, some caution is warranted when interpreting and generalizing the study results. It must be noted that the COPTEFL, as a newly developed testing format, was conducted on a small scale with a relatively small number of test-takers. The sample size which was drawn only from AUSFL might not have been representative or provided sufficient data for the study. Also, it should be stated that the sample only included AUSFL students who were mostly pre-intermediate or intermediate level students having a similar educational background and high computer familiarity due to their classes in the laboratories as a part of their regular language program. Thus, in more diverse and larger sample size, a more convincing and distinct or even different set of results might have been arrived at. The conclusions drawn from the analyses were, therefore, tentative.

4.2. Implications of the study

The findings of the study suggested that before serving the COPTEFL as a substitute for the face-to-face speaking tests, making students familiar with it through several practices and therefore, helping them get used to it is important. So, during these practice sessions, test administrators should explain test-takers the differences in testing formats and take their preferences into consideration, and thus support them when selecting the testing format that best meets their needs and interests. When test-takers get used to the COPTEFL, they can benefit from it as a new learning experience. On successful administration of it, the COPTEFL can be administered in language laboratory classes where students practice their English. This can help students assess their language on their own and see the progress in their level of English in time since the sounds are recorded. In time, practicing with the COPTEFL may help students to reduce the levels of nervousness mostly associated with face-to-face speaking tests. By giving award scores to their students constantly, the teachers can be informed about the profile of their students during the education period. The results of the experiment also showed that the use of the COPTEFL as a testing format helped reduce the amount of time, human resources, number of proctors, space and hard copy material required for a face-to-face speaking test. In addition to its advantages to test administrators and language testers, raters can also benefit from the convenience and user-friendliness of the rating platform in the COPTEFL system, which is attractive for its availability at any time and any place, low cost (i.e. no need for the use of cameras and CDs and a technical team to set the cameras) and potentially increased effectiveness compared with traditional face-to-face delivery. In foreign language programs at universities where there are a large number of incoming international students or exchange students (i.e. Erasmus students), it is generally a problem to group newcomers into appropriate instructional levels due to restricted time for organizing and delivering an entry or a placement test. For such cases, the COPTEFL can serve as a standardized oral proficiency test.

The issues in the development of the COPTEFL have also important implications for future computer-based speaking test developers. The step-by-step processes in the test design, construction and administration can be used as a roadmap for the test developers. This study can only be considered a first approach for a computer-based speaking test. Future researchers can improve on this study by changing the nature of task types, the number of tasks or the scoring system. Finally, the limited number of research of this type in Turkey, also, may be a reason to further study in this field.

Acknowledgements

This study is a part of the Doctoral Thesis of Cemre İşler in the Program of English Language Teaching in Anadolu University.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Cemre Isler: Investigation, Resources, Software Development, Data Collection, Data Analysis, and Writing the original draft. **Belgin Aydın:** Methodology, Supervision and Validation.

ORCID

Cemre İşler  <https://orcid.org/0000-0002-3622-0756>

Belgin Aydın  <https://orcid.org/0000-0002-4719-7440>

5. REFERENCES

- Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
- Aydın, B., Akay, E., Polat, M., & Geridönmez, S. (2016). Türkiye’deki hazırlık okullarının yeterlik sınavı uygulamaları ve bilgisayarlı dil ölçme fikrine yaklaşımları. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 16(2), 1-20.
- Aydın, B., Geridönmez, S., Polat, M. & Akay, (2017). Feasibility of computer assisted English proficiency tests in Turkey: A field study. *Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi*, 7(1), 107-122.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-57.
- Carr, N. T. (2011). *Designing and analyzing language tests: Oxford handbooks for language teachers*. Oxford University Press.
- Chapelle, C. A. (2013). Conceptions of validity. In *The Routledge handbook of language testing* (pp. 35-47). Routledge.
- Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies*, 35-49. TESOL.
- Council of Europe, (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative*. Pearson Education.
- East, M. (2016). Mediating Assessment Innovation: Why Stakeholder Perspectives Matter. In *Assessing Foreign Language Students’ Spoken Proficiency* (pp. 1-24). Springer.
- Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*, *Studies in language testing*, 35 (pp.77-151). Cambridge University Press.
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384-408.
- Fulcher, G. (2014). *Testing second language speaking*. Routledge.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

- Galaczi, E. D., & French, A. (2011). Context validity of Cambridge ESOL speaking tests. In L. Taylor (Ed.), *Examining speaking*, 30. Cambridge University Press.
- GSE. (2015). *New Global Scale of English Learning Objectives*. Pearson English. Retrieved May 18, 2017, from <http://www.english.com/gse#>
- Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish* [Unpublished master's thesis]. Boğaziçi University.
- Jeong, T. (2003). *Assessing and interpreting students' English oral proficiency using d-VOCI in an EFL context* [Unpublished doctoral dissertation]. Ohio State University, Columbus.
- Joo, M. (2008). *Korean university students' attitudes to and performance on a Face-To-Face Interview (FTFI) and a Computer Administered Oral Test (CAOT)* [Doctoral dissertation]. University of London.
- Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology*, 5(2), 60-83.
- Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, 8(4), 342-360.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading* (Vol. 29). Ernst Klett Sprachen
- Larsen-Hall, J. (2010). *A guide to doing statistics in second language research*. Routledge.
- Luther, A. C. (1992). *Designing interactive multimedia*. Multi-science Press Inc.
- Malabonga, V., Kenyon, D. M. & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59-92.
- Messick, S. (1989) Validity. In R. L., Linn (Eds.), *Educational measurement* (pp. 13-103). Macmillan/American Council on Education.
- McNamara, T. F. (1987). *Assessing the language proficiency of health professionals. Recommendations for the reform of the Occupational English Test* (Report submitted to the Council of Overseas Professional Qualifications). Department of Russian and language Studies, University of Melbourne, Melbourne, Australia.
- Mousavi, S. A. (2007). *Development and validation of a multimedia computer package for the assessment of oral proficiency of adult ESL learners: implications for score comparability* [Doctoral dissertation]. Griffith University.
- Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93(1), 836-847.
- O'Loughlin, K. (1997). *The comparability of direct and semi-direct speaking tests: A case study* [Unpublished doctoral dissertation]. University of Melbourne.
- O'Sullivan, D. B. (Ed.). (2011a). *Language testing: Theories and practices*. Palgrave Macmillan.
- O'Sullivan, B. (2011b). Language testing. In *The Routledge handbook of applied linguistics* (pp. 279-293). Routledge.
- Öztekin, E. (2011). *A comparison of computer assisted and face-to-face speaking assessment: Performance, perceptions, anxiety, and computer attitudes* [Master's thesis]. Bilkent University.
- Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113–125.
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.
- Shneiderman, B. (2004). *Designing the user interface: Strategies for effective human-computer interaction* (4th edition). Addison-Wesley.
- Shohamy, E., Donitsa-Schmidt, S., & Waizer, R. (1993). *The effect of the elicitation mode on the language samples obtained in oral tests* [Paper presentation]. 15th Language Testing Research Colloquium, Cambridge, UK.

- Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, 49(1), 75-92.
- Weir, C. J. (2005). *Language testing and validation*. Palgrave MacMillan.
- Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, 18(2), 187-206.
- Zainal Abidin, S. A. (2006). *Testing spoken language using computer technology: A comparative validation study of 'Live' and computer delivered test versions using Weir's framework* [Doctoral dissertation]. Universiti Teknologi Mara.
- Zak, D. (2001). *Programming with Microsoft Visual Basic 6.0: Enhanced edition*. Course Technology Thomson Learning.
- Zhou, Y. (2008). A comparison of speech samples of monologic tasks in speaking tests between computer-delivered and face-to-face modes. *JLTA Journal*, 11, 189-208.
- Zhou, Y. (2009). *Effects of computer delivery mode on testing second language speaking: The case of monologic tasks* [Doctoral dissertation]. Tokyo University of Foreign Studies.
- Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia*, 5(1), 2.

Examining the Measurement Invariance of TIMSS 2015 Mathematics Liking Scale through Different Methods

Zafer Erturk ^{1,*}, Esra Oyar ¹

¹Gazi University, Department of Educational Science, Ankara, Turkey

ARTICLE HISTORY

Received: Mar. 17, 2020

Revised: Dec. 01, 2020

Accepted: Jan. 05, 2021

KEYWORDS

Measurement Invariance,
TIMSS,
Latent Class,
Mixed Rasch Model,
Factor Analysis

Abstract: Studies aiming to make cross-cultural comparisons first should establish measurement invariance in the groups to be compared because results obtained from such comparisons may be artificial in the event that measurement invariance cannot be established. The purpose of this study is to investigate the measurement invariance of the data obtained from the "Mathematics Liking Scale" in TIMSS 2015 through Multiple Group CFA, Multiple Group LCA and Mixed Rasch Model, which are based on different theoretical foundations and to compare the obtained results. To this end, TIMSS 2015 data for students in the USA and Canada, who speak the same language and data for students in the USA and Turkey, who speak different languages, are used. The study is conducted through a descriptive study approach. The study revealed that all measurement invariance levels were established in Multiple Group CFA for the USA-Canada comparison. In Multiple Group LCA, on the other hand, measurement invariance was established up to partial homogeneity. However, it was not established in the Mixed Rasch Model. As for the USA-Turkey comparison, metric invariance was established in Multiple Group CFA whereas in Multiple Group LCA it stopped at the heterogeneity level. Measurement invariance for data failed to be established for the relevant sample in the Mixed Rasch Model. The foregoing findings suggest that methods with different theoretical foundations yield different measurement invariance results. In this regard, when deciding on the method to be used in measurement invariance studies, it is recommended to examine the necessary assumptions and consider the variable structure.

1. INTRODUCTION

In a world of rapid development and globalization, the information in the social, geographical, political, healthcare and educational fields of countries are easily accessed through a variety of organizations. An international database is thus possible because information regarding all countries is accessible. TIMSS -Trends in International Mathematics and Science Study and PISA - Programme for International Student Assessment are among the international educational databases. By way of these large-scale assessments, students from different educational systems can be compared for their both cognitive (e.g., mathematics, science

CONTACT: Zafer ERTÜRK ✉ zerturk35@gmail.com 📍 Gazi University, Department of Educational Science, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

achievement) and affective (attitude, perception, self-confidence, motivation, etc.) latent traits (Buchholz & Hartig, 2017).

There are a number of studies in the literature conducting cross-cultural comparisons thanks to the accessibility to international data (e.g. Alathl, Ayan, Demir & Uzun, 2016; Asil & Gelbal, 2012; Rutkowski & Svetina, 2014). In international assessments such as TIMSS, data is collected by administering a single measurement instrument to all participants from different countries. However, people from different socio-cultural backgrounds are likely to have different social, ethical and value judgments and interpret the scale items differently from each other. Thus, when collecting data from individuals from different cultures, researchers need to ensure that the items in measurement instruments mean the same in every culture.

The measurements need to be valid to obtain accurate results from the group comparisons made using the same measurement instrument (scale, questionnaire, test, etc.). In TIMSS, individuals from different cultures are administered the same measurement instruments. Therefore, the original versions of these instruments are translated into the languages spoken in all countries. The fact that measurement instruments can be translated flawlessly into other languages does not guarantee that each culture interprets the questions in the same way (Kankaras, 2010). Thus, there is an increasing need for addressing the methodological problems arising from the comparison of the data obtained from different countries and different cultures. One of these problems in intercountry comparisons is the invariance of measurements. In this regard, one of the basic concerns in any cross-cultural studies is whether or not the measurement invariance is established in testing the differences among groups (Hui & Triandis, 1985). In their study, Arım and Ercikan (2014) examined to what extent TIMSS 1999 U.S. and Turkey mathematics test results are comparable. In the comparison of the two countries, measurement invariance was taken into account and changing item function analyzes were performed for this. Accordingly, in the analysis made by comparing the test characteristic curves, it was determined that approximately 23% of the mathematics items operate differently between these two countries.

1.1. Measurement Invariance

Bryne and Watkins (2003) defined measurement invariance as the perception and interpretation of the items in the measurement instrument in the same way by individuals who in different sub-groups with respect to a certain variable. Invariance of measurements and methods that are adopted in cross-cultural studies across groups is referred to as the methodological invariance. Scale invariance and item invariance indicate the methodological invariance and concentrate on the degree of similarity between measurement methods across cultures (Kankaras, 2010).

Measurement invariance is a proof of validity employed to show that the same measurement instrument which is administered to different cultures in a study measures the same construct. In addition, since measurement instruments are created to measure a specific construct, the participants' responses should reveal their position about that specific construct. If their responses are influenced by additional factors which are different across cultures aside from the aimed construct, the invariance of measurements will fail to be established. In this case, the results to be obtained about the individuals by means of the measurement instrument will not reflect the real scores.

In order to test whether or not measurement invariance, a prerequisite in international comparison studies, is established, Multiple Group Confirmatory Factor Analysis (MG-CFA), an extension of Structural Equation Modeling (SEM), and the methods under the Item Response Theory (IRT) are adopted (Eid, Langenheine & Diener, 2003). In addition to these methods, mixed distribution models in which measurement invariance is examined by way of identifying the heterogeneous sub-groups are also implemented. Mixed distribution models have been

developed for the Item Response Theory (Mislevy & Verhelst, 1990; Rost & von Davier, 1995; von Davier & Rost, 1995;) and Structural Equation Models (Yung, 1997). These methods are combinations of a latent trait or the latent class analysis and a structural equation model (Eid & Rauber, 2000). Multiple Group Latent Class Analysis (MG-LCA), which is a method among mixed distribution models and is dependent on latent class analyses, may also be employed in measurement invariance studies (Magidson & Vermunt, 2001; Moors, 2004; Moors & Wennekers, 2003). MG-CFA is the method which is used when the observed and latent variables are continuous but cannot be used when both are categorical (Somers, Korkmaz, Dural & Can, 2009). MG-LCA, on the other hand, which is covered by the latent class models, can be used in measurement invariance studies if the two data structures mentioned are categorical. In addition, in their study in which MG-CFA and Differential Item Functioning (DIF) are compared based on IRT and MG-LCA, Kankaras, Vermunt and Moors (2011) stated that MG-LCA was an excellent alternative to the other two methods.

Another mixed distribution model is the Mixed Rasch Model (MRM). In mixed distribution Rasch models (Rost & von Davier, 1995), latent classes may be formed under a Rasch model for all individuals in a population and item difficulty parameters may differ across the unknown sub-groups (latent classes). Using this methodology, the number of groups required to account for the differences in item parameters can be identified.” In addition, the probability that an individual may belong to different classes can be calculated and individuals may be assigned to a latent class where their membership probability is maximum (Eid & Rauber, 2000). For ordered response categories (e.g., Likert-type scales), polytomous mixed Rasch model can be applied (Rost, 1991).

1.1.1. Multiple Group Confirmatory Factor Analysis

MG-CFA, a commonly preferred method in measurement invariance studies in various disciplines (Meredith, 1993; Mullen, 1995; Steenkamp & Baumgartner, 1998), is a parametric and linear approach investigating the similarity between measurement model parameters named as factor loadings, intercepts and error variances for the same factor models across groups.

Measurement invariance within the scope of MG-CFA is defined and tested through four hierarchical models (Byrne & Stewart, 2006; Meredith, 1993; Wu, Li & Zumbo, 2007). The measurement invariance levels that are tested in MG-CFA can be listed respectively as follows:

- i. Configural Invariance:** The configural model is the first level where the measurement invariance is tested in MG-CFA. This step allows freely estimating the factor loadings, regression constants and error variances concerning the groups.
- ii. Metric (Weak) Invariance:** Metric invariance, the second level, is the step where measurement units of groups regarding the latent variable are tested to find out whether they are similar or not. To this end, factor loadings are also restricted in addition to the factor number and factor pattern in groups.
- iii. Scalar (Strong) Invariance:** This model involves the restriction of the regression constants as well as the factor pattern and factor loadings (Tucker, Ozer, Lyubomirsk, & Boehm, 2006, p. 344).
- iv. Strict Invariance:** It is the last step of measurement invariance. The hypothesis that error terms concerning the items in the measurement invariance are equivalent across comparison groups is tested on this level (Önen, 2009).

There is a myriad of measurement invariance studies in Turkey conducted through MG-CFA. Based on TIMSS 1999 data for Turkey, Uzun and Öğretmen (2010) identified the affective factors that are influential in students' science achievement and tested these factors' measurement invariance by gender. In another study, Bahadır (2012) modeled the variables affecting students' reading skills by means of PISA 2009 data for Turkey. Then she tested the

measurement invariance of the obtained model across regions using MG-CFA. There are also studies which investigate the measurement invariance by gender and regions (Gülleroğlu, 2017; Ölçüoğlu 2015; Uzun, 2008) as well as those that compare the countries (Asil & Gelbal, 2012; Güzeller, 2011), by means of MG-CFA and based on the data on Turkey obtained from international assessments such as TIMSS and PISA. In his study, Güzeller (2011) examined whether the factor structure of the Computer Attitude Scale in PISA 2009 is similar across 10 different countries, in other words, its cross-cultural measurement invariance is made through MG-CFA. He obtained a similar factor structure as a result of the confirmatory factor analysis performed for all countries and showed that computer attitude has a cross-cultural invariance. Asil and Gelbal (2012) analyzed the cross-cultural and interlingual invariance of the student questionnaire administered within the scope of the Programme for International Student Assessment (PISA) 2016 comparatively based on the samples of Australia, New Zealand, USA and Turkey. In the conclusion part of their study, they stated that the measurement invariance failed to be established because of translation-related problems and cultural differences. Wu, Li, and Zumbo (2007) investigated the cross-country measurement invariance using TIMSS 1999 data in their study. Accordingly, by using the mathematics achievement scores of 21 countries participating in TIMSS 1999, it was checked whether the measurement invariance was achieved with MG-CFA. These countries include the U.S.A and Canada. According to the results obtained from the study, strict invariance was provided between the U.S.A and Canada. In the study conducted by Bowden, Saklofske, and Weiss (2011), the invariance of the measurement models of the Weschler Adult Intelligence Scale in U.S.A and Canada samples were examined. The model met the subtest scores that reflect similar structure measurement in both country samples and the assumption of invariance between samples. The results showed that structural validity was ensured in the measurement of cognitive abilities in U.S.A and Canadian samples and emphasized the importance of local norms.

1.1.2. Multiple Group Latent Class Analysis

MG-LCA as a concept is similar to MG-CFA in that it examines the relationship between categorical variables and latent constructs. MG-LCA analyzes the categorical latent constructs under the categorically observed variables whereas MG-CFA and IRT assume that latent variables are continuous. MG-LCA models the latent constructs as ordered categorical or nominal. Thus, instead of using the correlation/covariance matrix of data as done by MG-CFA, MG-LCA analyzes the cross-classification of the responses concerning the relevant variable (Kankaras, 2010). Measurement invariance within the framework of the latent class model is defined as the situation where the individuals who belong to different groups but are in the same latent class have the same observed response pattern and conditional probabilities (Millsap & Kwok, 2004).

Whether observable behaviors of individuals, such as attitudes, self-confidence, interest, willingness to study, and expressing that they find the lesson fun, arise from a latent structure is examined with latent variable models. There are three basic variables in these models: latent, observed and error. Observed variables are predicted by error and latent variables which explain the relationship between the observed variables, but the observed variables are not the cause of the latent variable (Collins & Lanza, 2010). In other words, if there is a latent variable that can be defined, the relationship between the observed variables disappears and this relationship is explained by the latent variable or variables (Goodman, 2002). Various models are available according to the fact that the variables are continuous and discontinuous. In latent class analysis, latent and observed variables are discontinuous. Latent variables observed in a traditional Latent Class Analysis consist of data at categorical or nominal scale level.

The latent class has at least two classes, if a model that can be defined with a single class is obtained, the observed variables are statistically independent of each other, so no latent

variables can be defined. The size of latent classes gives researchers information about subgroups in the universe. Another parameter used in the latent class analysis is conditional probabilities. Conditional probabilities can be likened to factor loadings in factor analysis. These parameters indicate the probability that an individual / observation in the t class of the X latent variable is at a certain level of the observed variable. Like the latent class probabilities, the sum of the conditional probabilities equals 1 (McCutcheon, 1987).

The most prominent reason why MG-LCA is preferred in measurement invariance studies is that almost all of the questions covered by the studies contain discrete (categorical or ordinal) response categories and can be used to identify the latent constructs from within the set of discrete observed variables (Kankaras, Moors & Vermunt, 2010). In addition, unlike MG-CFA and multiple group IRT which have strong assumptions about the distribution of data, MG-LCA is a rather flexible method feasible for all types of data. Second, while MG-CFA necessitates the invariance of at least two items under each factor to establish at least partial validity, there is no such requirement in MG-LCA. MG-LCA allows comparisons between groups even though each response in the model cannot establish the measurement invariance of the variable (Kankaras, 2010). In MG-LCA, the measurement invariance is gradually compared based on three basic models:

- i. **Heterogeneous Model:** In this model, which is tested in the latent class analysis on the first level of measurement invariance, parameters to be estimated (conditional or latent class probabilities) are not restricted. In other words, each parameter is allowed to be estimated separately in comparison groups (McCutcheon & Hagenaars, 1997).
- ii. **Partial Homogeneous Model:** Partial homogeneous model is the model in which slope parameters are tested by restriction. In this model, whether or not latent class probabilities differ across groups can be examined by way of removing the group-latent variable interaction effect from the model (Kankaras, Moors & Vermunt, 2010).
- iii. **Homogeneous Model:** This is the next step after the partial homogeneous model is tested. The homogeneous model step in the Latent Class Analysis is equivalent to the scalar (strong) invariance model in the structural equation modelings and fixed parameters are also restricted in addition to the slope parameters.

There are also measurement invariance studies carried out through MG-LCA in Turkey (Güngör, Korkmaz & Somer, 2013; Yandı, Köse & Uysal, 2017). Güngör, Korkmaz and Somer (2013) carried out a study which examined the measurement invariance by gender of the love capacity dimension of Values in Action Inventory through MG-LCA. They obtained two latent classes for both men and women and established the homogeneous model among the measurement invariance steps. In their study, Yandı, Köse and Uysal (2017) compared measurement invariance results acquired from the models having different statistical assumptions. In the data obtained from the Openness for Problem Solving Scale in PISA 2012, when the measurement invariance is examined through the invariance of mean covariance structures analysis having the assumption of normality, the steps up to strict invariance were accepted whereas, in MG-LCA, which does not require the assumption of normality, the partial homogeneous model was accepted.

1.1.3. Mixed Rasch Model

MRM is the combination of the Rasch model and the latent class analysis (Rost, 1991). In MRM, the probability of answering correctly is a function of both the individual's skill, which is a continuous variable and the individual's group, which is a categorical variable. The standard unidimensional Rasch model assumes that the responses or answers to the items of individuals who are at the same skill level have the same response technique (Fischer & Molenaar, 2012). Thus, the estimation of item difficulty to be obtained from the analyses remains constant across

different latent groups at the same skill level (Baghaei & Carstensen, 2013). If the measurement invariance in a dataset having two or more latent classes is examined through the standard Rasch model, the results may be misleading for the researcher since they will be interpreted based on a single class (Frick, Strobl & Zeileis, 2015).

In the mixed Rasch model, first, the number of the latent classes is identified in the examination of the measurement invariance. The formation of a single latent class is interpreted as the establishment of measurement invariance. If more than one class is formed, the establishment of measurement invariance is said to fail and effort is made to find out whether an item-based Differential Item Functioning (DIF) is present or not. (Yüksel, 2015). DIF is the case where individuals from different groups but at the same θ level are not likely to give the same answer to an item. A DIF investigation involves the comparison of the differences between item difficulties in different latent classes. Researchers argued that interpreting the response patterns of the individuals in each latent class would be more efficient than attempting to define the latent classes formed through MRM by the observed groups at hand (Bilir, 2009; Cho, 2007; Cohen & Bolt, 2005). In addition, Kelderman and Macready (1990) stated that approaching the DIF problem through MRM is more advantageous. The Mixed Rasch Model can be used in the analysis of the tests measuring the affective traits as well as in the achievement tests (Rost, Carstensen & von Davier, 1997).

Many studies tested the measurement invariance by means of MRM (Aryadoust, 2015; Aryadoust & Zhang, 2016; Cohen & Bolt, 2005; Eid & Rauber, 2000; Pishghadam, Baghaei & Seyednozadi, 2017; Şen, 2016; Yalçın, 2019; Yüksel, 2015; Yüksel, Elhan, Gökmen, Küçükdeveci & Kutlay, 2018). Tee and Subramaniam (2018) analyzed the measurement invariance of the attitudes towards eighth grade science in the UK, Singapore and USA countries that entered TIMSS 2011 with Rasch analysis. According to the results obtained from the research, there are some differences between students in Asia and students in the West. More specifically, Singaporean students acknowledge the instrumental value of science more than students in the UK and the US. Although Singaporean students are more successful than students from the USA and the UK, they are less confident in science. When it comes to their feelings for science, again, Singaporean students love science more than U.S.A and U.K students.

Ölmez and Cohen (2018) in their study, Partial Credit Model of Mixed Rasch Models of the sixth and seventh grade students in Turkey are used to identify differences in mathematics anxiety. Two latent classes were identified in the analysis. While students in the first latent class have less anxiety about understanding mathematics lessons and the use of mathematics in daily life, students in the second class have more self-efficacy for mathematics. Students in both classes are similar in terms of exam and assessment anxiety. In addition, it was observed that students in the first latent class were more successful in mathematics, mostly liked mathematics and mathematics teachers, and had better-educated mothers than students in the second latent class. In addition, observed variables such as gender, private or public school attendance, and education levels of fathers did not differ significantly between latent classes.

1.2. Purpose

Measurement instruments are created based on the assumption that "an instrument measures the same construct in each group" (Başusta & Gelbal, 2015). The results of the studies in which the measurement invariance of the measurement instruments administered to different groups and different cultures remains untested may raise a lot of question marks in minds. Thus, the invariance of the measurement instruments needs to be tested before the initiation of intergroup, intercountry or cross-cultural comparisons. Since testing the measurement invariance makes a significant contribution to the validity of the results in comparison studies, the selection of the method to be utilized in compliance with the data structure when testing the measurement

invariance and fulfillment of the assumptions are of such importance. Thus, the validity of measurements would be further proved as the researchers adopt various methods to test the measurement invariance (Kankaras, Vermunt & Moors, 2011).

The purpose of this study is to investigate the measurement invariance of the data obtained from the "Mathematics Liking Scale" in TIMSS 2015 through MG-CFA, MG-LCA and MRM, which are based on different theoretical foundations and compare the obtained results. To this end, the country level was taken into consideration when forming the sub-groups. Mathematics achievement rankings were taken into account when determining the 3 countries included in the study. Comparisons were made between America, which is in the middle in the success ranking, and Canada, which is more successful. The analysis was also made between Amerika and Turkey which is less successful. In addition, the measurement invariance between the countries where the same language is spoken (USA and Canada) and the countries where different languages are spoken (USA and Turkey) was tested.

In this study, the Mixed Rasch Model, which is one of the methodologically prominent Mixed Item Response Theory models in test development and measurement invariance studies, and MG-LCA model and MG-CFA methods are focused on. The comparison of KRM and MG-LCA methods, whose mathematical methodologies are similar, will provide guiding results for researchers who will use these methods. In addition to the KRM and MG-LCA methods, the MG-CFA method, which has been used in measurement invariance studies for many years, was included in the study, and the validity of the study results was increased. In this study, the theoretical foundations of analysis methods used in the field of measurement invariance are explained in detail. In addition, testing the linguistic measurement invariance will also provide us with more valid information about the significance of the comparisons made according to cultural differences in the TIMSS 2015 student survey.

2. METHOD

2.1. Research Design

The purpose of this study is to investigate the invariance of the "Mathematics Liking Scale" in TIMSS 2015 in American, Canadian and Turkish cultures through MG-CFA, MG-LCA and MRM. The current research is a descriptive study as aims to identify the cross-cultural validity level of the "Mathematics Liking Scale" in TIMSS 2015 study (Karasar, 2013).

2.2. Population and Sample

6079, 8068 and 9509 eighth-grade students from Turkey, Canada and the USA, respectively, participated in the TIMSS 2015 developed by the International Association for the Evaluation of Educational Achievement (IEA). A two-step path is pursued in the sample selection for TIMSS 2015. In this process, the schools are first selected from both public and private schools in each country through random sampling. Afterward, a class is chosen from each school (Olson, Martin & Mullis, 2008). The reason why eighth grade students were chosen in the study is that students' interests and attitudes towards mathematics are more pronounced in this age range. Since eighth grade students are in the last grade of primary education, they know themselves better than fourth grade students and their interests and attitudes towards lessons do not change much.

2.3. Data Collection Tool

The Mathematics Liking Scale in TIMSS 2015, which aims to identify whether or not students like math class, consists of a total of 9 items (TIMSS, 2015). Items were translated into Turkish by the researchers. The reason for using the "Mathematics Liking" scale within the scope of the study is the high number of items. In addition, the "Mathematics Liking" scale reflects general affective expressions towards mathematics. Thus, the perception of the statements in

the items is similar for the students of each country participating in TIMSS. The items are presented in the [Appendix Table A1](#) both in English and in Turkish with their codes.

2.4. Data Analysis Procedures

The study employed three different methods, namely Multiple Group CFA, Multiple Group LCA and Mixed Rasch Model, in testing the measurement invariance. The steps followed in the analysis of data are as follows:

- i. Calculation of the required statistics for the missing data, extreme value, normality, homogeneity of variance and multi-variant normality (testing of assumptions).
- ii. Performance of CFA
- iii. Performance of MG-CFA and testing of the levels of measurement invariance
- iv. Performance of Latent Class Analysis and testing of the levels of measurement invariance
- v. Implementation of the MRM and examination of the results
- vi. Comparison of the methods based on the obtained results

2.4.1. Assessment Criteria

The MG-CFA method involves calculating the differences between the CFI and TLI values in comparing the two models in order to find out whether the measurement invariance is established. Measurement invariance is not established when Δ CFI and Δ TLI values are below -0.01 or above 0.01 (Byrne, Shavelson & Muthen, 1989; Li, Wang, Shou, Zhong, Ren, Zhang & Yang, 2018; Liang & Lee, 2019; Schnabel, Kelava, Van de Vijver & Seifert, 2015; Wu, Li & Zumbo, 2007).

In the LCA model selection process, the simplest (parsimony) model, in other words, the model having the least number of latent classes and in which less parameter is predicted is sought. Statistical criteria, parsimony and interpretability should be considered in the model selection process (Collins & Lanza, 2010; Silvia, Kaufman & Pretz, 2009). There are several criteria in MG-LCA that are frequently used in the assessment of model-data fit. The likelihood ratio chi-square (L2) statistics are used as a standard criterion for the inconsistency between the observed and expected frequencies in the model. In addition to L2 statistics, various information criteria including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), modified AIC (AIC3) and consistent AIC (CAIC) are used in testing the measurement invariance in MG-LCA. When the sample size is large, BIC and CAIC are used for the model-data fit. When the sample size is small or medium, however, usually AIC statistics is used (Kankaras, Moors & Vermunt, 2011).

In order to identify the appropriate model-data fit in Mixed Rasch Model, aside from the criteria such as AIC and BIC as in MG-LCA, different statistics may be used, for example, the significance levels of Cressie Read and Pearson Chi-square values. Accordingly, the model obtained when p-value of Cressie Read or Pearson Chi-square is equal to or above 0.05 is said to be the appropriate model (von Davier, 2000). In addition, a common problem concerning chi-square parameters for the scale data observed in item-response models is that the number of cells significantly greater than the number of response models. The bootstrap method is recommended as a solution to this problem (Langeheine, Pannekoek & van der Pol, 1996). Thus, bootstrapped p-values of Cressie Read and Pearson Chi-square values are employed in this study to decide the appropriate number of latent classes.

In the event that a 1-class model is selected as the most appropriate model in model-data fit in MRM, it can be said that measurement invariance has been established, in other words, Differential Item Functioning (DIF) is not present in any of the items. However, if model-data fit cannot be ensured for a 1-class model, some items will be understood to have DIF. In testing DIF in items, item difficulties are calculated for the items in each class starting from the 1-class model to the latent class where the most appropriate model is identified. Identification of the

items displaying DIF involves the comparison of the differences between the item difficulty indices calculated for each latent class (Yüksel, Elhan, Gökmen, Küçükdeveci & Kutlay, 2018). Finally, contingency table analysis is performed to investigate whether the latent classes and observed variables (age, gender, status, country, etc.) are interrelated to find out the source of DIF occurring in some items.

2.4.2. Testing of the Assumptions

Items were reverse-coded as required before the pre-analysis. The missing data were removed from the dataset and excluded from the analyses. Deletion is preferred for the missing data, as it is not more than 5% in data and has a sufficient sample size. The testing of the assumptions was continued with 9509, 8068 and 5741 student data from the USA, Canada and Turkey, respectively.

In examining the extreme values, z score concerning the total scale score was calculated separately for each country and the values obtained were observed to be in the range of -1.54 and +1.95. In this regard, the data contained no extreme value. Skewness and kurtosis values were examined in testing the normality. Values for skewness and kurtosis were found to be in the range of ± 1 for the entire group and for each country. Thus, the data were proved to fulfill the coefficient of normality (Büyüköztürk, 2017). In the analysis, LISREL 8.80 for MG-CFA; LATENT GOLD 5.0 for MG-LCA and WINMIRA 2001 package programs for MRM were used.

2.4.3. Confirmatory Factor Analysis Results

Firstly, in order to identify whether or not the measurement model developed in each step of the measurement invariance test established model-data fit, was performed and the obtained fit indices were reported and interpreted. CFA results for each country are shown in Table 1.

Table 1. Model Fit Indices of Each Country Obtained from Measurement Models

Fit Index	Measurement Model Results			Perfect Fit	Acceptable Fit
	US	Canada	Turkey		
RMSEA	0.09	0.08	0.10	$0.00 \leq \text{RMSEA} \leq 0.05$	$0.05 \leq \text{RMSEA} \leq 0.10$
CFI	0.98	0.99	0.98	$0.97 \leq \text{CFI} \leq 1.00$	$0.95 \leq \text{CFI} \leq 0.97$
TLI	0.98	0.98	0.97	$0.95 \leq \text{TLI} \leq 1.00$	$0.90 \leq \text{TLI} \leq 0.95$
NFI	0.98	0.99	0.98	$0.95 \leq \text{NFI} \leq 1.00$	$0.90 \leq \text{NFI} \leq 0.95$
AGFI	0.88	0.91	0.88	$0.90 \leq \text{AGFI} \leq 1.00$	$0.85 \leq \text{AGFI} \leq 0.90$
GFI	0.93	0.95	0.93	$0.95 \leq \text{GFI} \leq 1.00$	$0.90 \leq \text{GFI} \leq 0.95$

Table 1 shows that, based on the results of the measurement models developed separately for each country, the RMSEA values are in the acceptable range (Hooper, Coughlan & Mullen, 2008; Kelloway, 1989; Steiger, 1990) while CFI, TLI and NFI values are in the perfect fit range (Sümer, 2000). AGFI and GFI values display perfect fit in the measurement model developed for Canada (Anderson & Gerbing, 1984; Cole, 1987) and are in the acceptable range for the USA and Turkey.

3. RESULT / FINDINGS

In this section, findings concerning MG-CFA, MG-LCA and MRM, which were employed to test the measurement invariances of the models obtained from the countries matched with respect to language (the same language or different languages) are presented.

3.1. Findings Obtained from MG-CFA

The results of MG-CFA that was performed to test the measurement invariance of data for "Mathematics Liking Scale" are presented in Table 2.

Table 2. MG-CFA Results for USA-Canada and USA-Turkey Data

	Steps	χ^2	sd	CFI	GFI	TLI	RMSEA	Δ CFI	Δ TLI
US-Can.	Configural Invariance ¹	4003.86	51	0.99	0.98	0.98	0.094	--	--
	Metric (Weak) Invariance ²	4136.61	60	0.98	0.98	0.98	0.088	-0.01	0.00
	Scalar (Strong) Invariance ³	4647.68	68	0.98	0.99	0.98	0.088	-0.01	0.00
	Strict Invariance ⁴	5070.21	77	0.98	0.98	0.98	0.086	-0.01	0.00
USA-Tur.	Configural Invariance ¹	3714.90	50	0.98	0.98	0.97	0.098	--	--
	Metric (Weak) Invariance ²	4064.85	60	0.98	0.97	0.98	0.094	0.00	0.01
	Scalar (Strong) Invariance ³	6429.77	69	0.97	0.97	0.97	0.110	-0.01	0.00
	Strict Invariance ⁴	7918.75	78	0.96	0.94	0.96	0.115	-0.02	-0.01

¹ Factor loadings, factor correlations and error variances are free

² Factor loadings are fixed (factor correlations and error variances are free)

³ Factor loadings and factor correlations are fixed (error variances are free)

⁴ Factor loadings, factor correlations and error variances are fixed

It is seen in Table 2 that model-data fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model developed in the configural invariance step given under USA-Canada comparison show a perfect fit. Therefore, it can be argued that the measurement model is the same for both countries. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the model developed in the metric invariance step display that the model-data fit is perfect. Examination of the difference between CFI and TLI values suggests that the difference is in the range of ± 0.01 (Δ CFI = -0.01, Δ TLI = 0.00) and metric invariance is established. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model created to test the scalar invariance show that model-data fit is established. Examination of Δ CFI and Δ TLI reveals that the values are in the range of ± 1 (Δ CFI = -0.01, Δ TLI = 0.00) and scalar invariance is established. Finally, model-data fit is seen to be established when the fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) in the developed strict variance model are examined. Examination of Δ CFI and Δ TLI reveals that the values are in the range of ± 1 (Δ CFI = -0.01, Δ TLI = 0.00) and strict invariance is established. In conclusion, as a result of the analyses performed based on data on the USA and Canada, all steps of measurement invariance have been observed to be established.

Comparison of USA-Turkey samples shows that the model-data fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model which was developed to test the configural invariance reflect a perfect fit. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the model which was developed in the metric invariance step suggest perfect model-data fit. The difference between Δ CFI and Δ TLI values is shown to be in the range of ± 0.01 (Δ CFI = 0.00, Δ TLI = 0.01). Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model created to test the scalar invariance show that model-data fit is established. The Δ CFI and Δ TLI values are observed to be in the range of ± 1 and the scalar invariance is established (Δ CFI = -0.01, Δ TLI = 0.00). Finally, model-data fit is seen to be established when the fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) in the developed strict variance model are examined. Examination of Δ CFI and Δ TLI values reveals that Δ TLI value is in the range of ± 0.01 whereas Δ CFI is out of this range (Δ CFI = -0.02, Δ TLI = -0.01). In this case, strict invariance cannot be established. In brief, the results of the analyses performed based on the data on the USA and Turkey indicate that among the measurement invariance steps, configural, metric and scalar invariances are established but strict invariance cannot be established.

3.2. Findings Obtained from MG-LCA

In order to test the measurement invariance through MG-LCA, first, the number of latent classes is identified for Turkey, USA and Canada. The obtained statistics starting from 1 up to the 4-class model are examined to identify the number of latent classes in countries. The number of latent classes obtained for each country and the assessment criteria for classes are provided in Table 3.

Table 3. Latent Classes and Information Criteria Values by Countries

	Estimated number of parameters	<i>sd</i>	L ²	BIC	AIC	AIC3	CAIC
Turkey							
1-class	9	5732	78573.875	28961.187	67109.875	61377.875	23229.187
2-class	19	5722	73184.813	23658.679	61740.813	56018.813	17936.679
3-class	29	5712	72997.478	23557.898	61573.478	55861.478	17845.898
4-class	39	5702	72997.477	23644.451	61593.477	55891.477	17942.451
5-class	49	5692	72997.478	23731.005	61613.478	55921.478	18039.005
Canada							
1-class	9	8059	108154.927	35658.895	92036.927	83977.927	27599.895
2-class	19	8049	102234.857	29828.782	86136.857	78087.857	21779.782
3-class	29	8039	102043.556	29727.438	85965.556	77926.556	21688.438
4-class	39	8029	102043.556	29817.395	85985.556	77956.556	21788.395
5-class	49	8019	102043.556	29907.351	86005.556	77986.556	21888.351
US							
1-class	9	9500	135863.083	48843.140	116863.083	107363.083	39343.140
2-class	19	9490	126847.106	39918.763	107867.106	98377.106	30428.763
3-class	29	9480	126565.792	39729.049	107605.792	98125.792	30249.049
4-class	39	9470	126565.762	39820.649	107625.762	98155.792	30350.649
5-class	49	9460	126565.793	39912.249	107645.793	98185.793	30452.249

Table 3 shows that the three-class model has the lowest values for L2, BIC, AIC, AIC3 and CAIC in each country. In this context, it can be said that the latent variable of liking mathematics has three latent classes for the research sample. During the testing of the measurement invariance, analyses were performed based on the three-class model. Accordingly, first, the heterogeneous model, in which fixed and slope parameters are freely estimated, then, the partial homogeneous model in which slope parameters in both datasets are accepted equal and finally, the homogeneous model in which fixed parameters are also equalized in addition to slope parameters were created. First, the measurement invariance between the USA and Canada, where the same language is spoken, was tested. Accordingly, MG-LCA results for the USA-Canada sample are as shown in Table 4.

Table 4. MG-LCA Results Obtained for the USA – Canada and USA- Turkey

	Steps	Estimated number of parameters	sd	L ²	BIC	AIC	AIC3	CAIC
USA-Canada	Heterogeneous Model	166	17411	57088.965	-113092.181	22266.965	4855.965	-130503.181
	Partial Homogeneous Model	112	17465	57446.451	-113262.510	22516.451	5051.451	-130727.510
	Homogeneous Model	85	17492	58554.107	-112418.761	23570.107	6078.107	-129910.761
USA - Turkey	Heterogeneous Model	166	15084	52902.889	-92391.248	22734.889	7650.889	-107475.248
	Partial Homogeneous Model	112	15138	53877.805	-91936.478	23601.805	8463.805	-107074.478
	Homogeneous Model	85	15165	58080.332	-87994.024	27750.332	12585.332	-103159.024

Based on the comparison of the USA and Canada samples, it can be said that the most appropriate model according to BIC and CAIC is the partial homogeneous model (Kankaras & Moors, 2011). Comparison of USA-Turkey reveals that BIC and CAIC values are the lowest for the heterogeneous model. Thus, concerning the MG-LCA results for the USA-Turkey sample it can be said that the measurement invariance cannot be established.

3.3. Findings Obtained from MRM

In order to test the measurement invariance through MRM, first, the most appropriate number of latent classes to establish model-data fit for the USA-Canada and USA-Turkey were set. 400 bootstrap samples were used in each analysis to decide the number of the appropriate latent classes. The appropriate number of classes is decided considering the biggest insignificant p-value of Bootstrap Pearson χ^2 above 0.05. The number of latent classes and fit assessment criteria for the samples of USA-Canada and USA-Turkey are shown in Table 5.

Table 5. Fit Statistics for the Mixed Rasch Model

	Estimated number of parameters	BIC	Geometric Mean LL	Cressie Read (Bootstrap p-value)	Pearson χ^2 (Bootstrap p-value)
USA-Canada					
1-class	28	313816.58	0.37120018	0.000	0.000
2-class	57	301013.13	0.38687637	0.000	0.097
3-class	86	297434.88	0.39162740	0.000	0.010
USA-Turkey					
1-class	28	283209.56	0.35674063	0.000	0.000
2-class	57	269961.05	0.37476185	0.000	0.008
3-class	86	266247.67	0.38025276	0.000	0.022
4-class	115	264500.81	0.38306995	0.003	0.500
5-class	144	263102.19	0.38541874	0.000	0.013

According to the model assessment criteria in Table 5, one-class models in both samples, USA-Canada and USA-Turkey, are not appropriate. In this case, it can be claimed that the measurement invariance is not established for both samples. Once the establishment of the measurement invariance is failed, the appropriate number of classes to establish the model-data

fit is tried to be set. In the USA-Canada sample, in which the same language is spoken, only the p -value for Bootstrap Pearson χ^2 value of the two-class model is not significant ($p > 0.05$). In this case, the 2-class model was decided to be the most appropriate model for the USA-Canada sample. In the USA-Turkey sample, in which different languages are spoken, it is the four-class model in which the Bootstrap Pearson χ^2 value is not significant ($p > 0.05$).

Since the measurement invariance could not be established, item-based measurement invariance in MRM was examined. In this regard, first, the measurement invariance of nine items in the Mathematics Liking Scale was examined in the USA-Canada sample. The model establishing the model-data fit for the USA-Canada sample is the two-class model.

As for DIF, it emerges when differences take place between the difficulty parameters in classes. Item difficulty parameters obtained for each class are shown in Table 6. Comparison of the classes between rows allows identifying the items which are disproportionately easy or difficult and thus coming up with a clearer interpretation of each class.

Table 6. Item Difficulty Estimations for Two-Class Model in the USA-Canada Sample

Items	Class 1	Class 2
Item 1	0.949	0.408
Item 2	0.062	0.048
Item 3	-0.482	-0.138
Item 4	1.147	0.602
Item 5	0.546	0.233
Item 6	-0.600	-0.514
Item 7	-0.213	-0.155
Item 8	-0.769	-0.382
Item 9	-0.639	-0.102

Based on Table 6, Item 1 and Item 4 in the Latent Class 1 can be said to be more difficult than those in the Latent Class 2, in other words, individuals who are in Class 2 like mathematics less compared to the individuals in the Latent Class 1. On the other hand, it is seen that Item 8 and Item 9 are more difficult for the Latent Class 2, in other words, individuals who are in Class 1 like mathematics less compared to the individuals in the Latent Class 2. Some items were identified to have DIF as a result of the differentiation of difficulty parameters related to them into two latent classes. χ^2 test statistics is adopted to find out the source of DIF. Accordingly, since this study employs students from different countries, χ^2 analysis is performed between the students' latent classes and countries. 54% and 46% of the USA-Canada sample are made up of American and Canadian students, respectively. Results of the χ^2 test analysis performed between countries and class membership are shown in Table 7.

Table 7. Results of χ^2 Analysis Between Latent Classes and Countries

Country	Latent Class		Total	χ^2	p
	1	2			
U.S.A	5154 (54.2%)	4355 (45.8%)	9509 (54%)	102.90	0.00*
Canada	4985 (61.8%)	3083 (38.2%)	8068 (46%)		
Total	10139 (57.7%)	7438 (42.3%)	17577 (100.0%)		

* $p \leq .05$

Table 7 suggests a significant relationship between students' coming from different countries and latent class membership ($\chi^2 = 102.90$; $p \leq 0.05$). In this regard, DIF is considered to arise from students' coming from different countries. The rates of the American and Canadian students in Latent Class 1 are 54.2% and 61.8%, respectively. The rates in the second latent class are 61.8% for American students and 38.2% for Canadian students.

The measurement invariance of nine items in the Mathematics Liking Scale was examined for the USA and Turkey, where different languages are spoken. The model establishing the model-data fit for the USA-Turkey sample is the four-class model. Since a four-class construct emerged in the USA-Turkey sample speaking different languages, the measurement invariance could not be established. In this regard, in order to identify which items in the Mathematics Liking Scale prevent the measurement invariance from being established, in other words, display DIF, item difficulty parameters for each class were calculated and are presented in Table 8.

Table 8. Item Difficulty Estimations for Four-Class Model in the USA-Turkey Samples

Items	Class 1	Class 2	Class 3	Class 4
Item 1	1.066	0.711	0.872	0.264
Item 2	0.559	-1.041	-0.717	0.921
Item 3	-0.004	-0.742	-0.705	0.775
Item 4	1.218	1.525	0.584	-0.399
Item 5	0.696	0.359	1.002	0.093
Item 6	-2.021	-0.240	0.214	-0.461
Item 7	-0.577	0.058	0.181	-0.218
Item 8	-0.596	-0.472	-0.818	-0.637
Item 9	-0.340	-0.157	-0.613	-0.338

Examination of Table 8 reveals that item difficulty parameter values of the Latent Class 4 for Item 1, Item 4 and Item 5 are lower than the item difficulty values in other latent classes. Difficulty indices of the Latent Class 2 and the Latent Class 3 for Item 2 are observed to reflect quite low values as opposed to the difficulty indices of the Latent Class 1 and the Latent Class 4, which display very high values. For Item 3, the value of the difficulty parameter of the Latent Class 4 is much higher than that of the other latent classes. For Item 6, the item difficulty parameter value of the Latent Class 1 is much lower when compared to the other latent classes.

Considering that the difficulty parameters for some items are very different across four latent classes, the items can be claimed to have DIF. In MRM, χ^2 test statistics are used to identify the DIF source. The χ^2 analysis is carried out between the students' latent classes and countries in order to examine whether or not there is DIF with respect to coming from countries speaking different languages. 62% and 38% of the USA-Turkey sample are made up of American and Turkish students, respectively. Results of the χ^2 analysis performed between countries and class membership are shown in Table 9.

Table 9. Results of χ^2 Analysis Between Latent Classes and Countries

Country	Latent Class				Total	χ^2	<i>p</i>
	1	2	3	4			
U.S.A	4324 (45.5%)	3146 (33.1%)	1311 (13.8%)	728 (7.7%)	9509 (62%)	1,363.13	0.00*
Turkey	992 (17.3%)	2560 (44.6%)	1641 (28.6%)	548 (9.5%)	5741 (38%)		
Total	5316 (34.9%)	5706 (37.4%)	2952 (19.4%)	1276 (8.4%)	15250 (100.0%)		

* $p \leq 0.05$

Table 9 suggests a significant relationship between students' coming from different countries and latent class membership ($\chi^2=1363.13$; $p \leq 0.05$). In this regard, students' coming from different countries can be suggested as a DIF source. The rates of American and Turkish students in Latent Class 1 are 45.5% and 17.3%, respectively. The rates in the second latent class are 33.1% for American students and 44.6% for Turkish students. The rates of American and Turkish students in Latent Class 3 are 13.8% and 28.6%, respectively. The rates in Latent Class 4 are 7.7% for American students and 9.5% for Turkish students.

4. DISCUSSION and CONCLUSION

Cross-cultural studies enable us to explore the universality of social and psychological laws and the cultural differences in people's characteristics, views and behaviors. A number of generalizations are made through comparison studies regarding the differences between the cultural groups. Thus, the validity of the results of the cross-cultural comparisons gains importance. Proving the validity comparison results necessitates testing the measurement invariance of measurement instruments because although the original measurement instrument can be translated into the languages of other cultures "flawlessly", it is not possible for each culture to interpret the questions in the same way (Hui & Triandis, 1985).

This study aims to examine the measurement invariance of the data obtained from the "Mathematics Liking Scale", which was administered to the students in TIMSS 2015 assessment by means of different methods, in countries, speaking the same and different languages. To this end, MG-CFA, MG-LCA and MRM methods which have different theoretical foundations were adopted.

As a result of the study, all steps of the measurement invariance was established when MG-CFA was employed for the analyses performed for the USA and Canada where the same language is spoken. In other words, data on these countries are comparable. When the measurement invariance was examined using the same data, it was seen that partial homogeneity was achieved by the MG-LCA. This step corresponds to the metric invariance in MG-CFA. In the MRM, another method used in the study, the measurement invariance for the USA-Canada sample could not be established and some items were found to have DIF. Country differences were examined to identify the possible cause of DIF and the results were found to be significant.

The examination of the measurement invariance of the data obtained from the American and Turkish students who speak different languages revealed that the steps up to the scalar invariance in the MG-CFA were established. This result parallels with the measurement invariance results for the "Support for Scientific Inquiry" questionnaire for students, which was administered within the scope of PISA 2006, in the study conducted by Asil and Gelbal (2012). In the analyses, it was found out that none of the items disturbed the invariance in samples of countries having a similar culture (Australia-New Zealand); that two items disturbed the invariance in the samples of the countries speaking the same language but having different

cultures (Australia-USA); and nine items disturbed the invariance in the samples of countries having different languages and cultures (Australia-Turkey). When MG-LCA was used to examine the measurement invariance of data obtained from American and Turkish students who speak different languages, the measurement invariance remained in the heterogeneity step. This step corresponds to the configural invariance in MG-CFA. In the MRM, on the other hand, the measurement invariance for the USA-Turkey samples could not be established and some items were found to have DIF. The country variable was examined to find out the possible causes of DIF and country difference was found to be a possible cause. In the study, Yandı, Köse, Uysal and Oğul (2017) obtained similar results and found that the measurement invariance could not be established when the countries with different cultures as well as different languages were compared. Köse (2015) also came up with a similar result. According to the results obtained from the study, while the individual parameter estimates obtained by MRM were good in heterogeneous data sets, it was observed that MRM was not successful in determining the reason for the difference in item function in data sets with multi-category and small sample. Sırgancı (2019) examined the effect of the covariant (common) variable in determining the changing item function with the Mixed Rasch Model. According to the results obtained from the study, MRM's latent DMF determination power and correct decision percentage increased significantly when the covariant variable was included in the model.

In conclusion, MG-LCA can be claimed to be a good alternative to MG-CFA in cases where the data structure is continuous. The differences detected between MG-CFA and MG-LCA are also similar to the results of the study carried out by Yandı, Köse and Uysal (2017). Moreover, the results obtained from this study coincide with the results of the studies conducted by Kankaras, Vermunt and Moors (2011) in which the methods based on IRT, SEM and LCA were compared. The advantage of the Mixed Rasch Model is that it allows not only detecting the DIF but also interpreting its possible cause more directly. Thus, unlike MG-CFA, MRM provides very detailed information for item response profiles. Therefore, it was found that MRM would be helpful especially in examining the invariance of the measurement instruments if used in combination with MG-CFA (Quandt, 2011).

According to the results obtained from this study, first of all, it is recommended to test the invariance of the structures to be compared in the comparison studies of the countries participating in large-scale exams. In this study, methods with different theoretical foundations were used to test the measurement invariance at the scale and item level. Future studies can test the measurement invariance with IRT-based methods in addition to these methods.

There are studies in the literature testing the measurement invariance (Eid, Langeheine & Diener, 2003; Kankaras & Moors, 2010; Somer, Korkmaz, Sural & Can, 2009; Yandı, 2017; Yandı, Köse & Uysal, 2017). The common finding in these studies is that the measurement invariance results obtained by different methods differ from each other. Since each method has its own assumptions and statistical backgrounds and is based on its own data structure different results can be obtained. In conclusion, it is recommended to provide evidence for measurement invariance by means of different methods in future studies (Kankaras, Vermunt & Moors, 2011).

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Authors are expected to present author contributions statement to their manuscript such as;

Zafer Erturk: Investigation, %60, Methodology, %65, Resources, %50, Visualization, %60, Software, %55, Formal Analysis, %60, and Writing, %50, Supervision, **Esra Oyar:** Investigation, %40, Methodology, %35, Resources, %50, Visualization, %40, Software, %45, Formal Analysis, %40, and Writing, %50.

ORCID

Zafer ERTÜRK  <https://orcid.org/0000-0003-3651-7602>

Esra OYAR  <https://orcid.org/0000-0002-4337-7815>

5. REFERENCES

- Anderson, J. C., & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2), 155-173. <https://doi.org/10.1007/BF02294170>
- Arim, R. G., & Ercikan, K. (2014). Comparability between the American and Turkish versions of the TIMSS mathematics test results. *Education & Science*, 39(172), 33-48.
- Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing*, 15(3), 216-238. <https://doi.org/10.1080/15305058.2015.1004409>
- Aryadoust, V., & Zhang, L. (2016). Fitting the mixed rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing*, 33(4), 529-553. <https://doi.org/10.1177/0265532215594640>
- Asil, M., & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği [Cross-cultural equivalence of the PISA student questionnaire]. *Eğitim ve Bilim*, 37(166), 236-249.
- Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, 18. 1-13. <https://doi.org/10.7275/n191-pt86>
- Bahadır, E. (2012). *Uluslararası Öğrenci Değerlendirme Programı'na (PISA 2009) göre Türkiye'deki öğrencilerin okuma becerilerini etkileyen değişkenlerin bölgelere göre incelenmesi* [According Programme for International Student Assessment (PISA 2009), investigation of variables that affect Turkish students' reading skills by regions]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.
- Başusta, N. B., & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement invariance at groups' comparisons: a study on PISA student questionnaire]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 30(4), 80-90.
- Bilir, M. K. (2009). *Mixture item response theory-mimic model: simultaneous estimation of differential item functioning for manifest groups and latent classes* (Unpublished doctoral dissertation). Florida State University.
- Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement*, 71(1), 186-199.
- Brien, M., Forest, J., Mageau, G. A., Boudrias, J. S., Desrumaux, P., Brunet, L., & Morin, E. M. (2012). The basic psychological needs at work scale: measurement invariance between Canada and France. *Applied Psychology: Health and Well-Being*, 4(2), 167-187.
- Bryne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, 34(2), 155-175. <https://doi.org/10.1177/0022022102250225>

- Buchholz, J., & Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*, 43(3), 241-250. <https://doi.org/10.1177/0146621617748323>
- Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabı istatistik, araştırma deseni SPSS uygulamaları ve yorum*. [Data analysis handbook statistics for social sciences, research design spss applications and interpretation.] Ankara: Pegem Akademi Yayıncılık.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466. <https://doi.org/10.1037/0033-2909.105.3.456>
- Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, 13(2), 287-321. <https://doi.org/10.1207/s15328007sem1302>
- Cho, S. J. (2007). *A multilevel mixture IRT model for DIF analysis* (Doctoral dissertation, uga).
- Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133-148. <https://doi.org/10.1111/j.1745-3984.2005.00007>
- Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, 55(4), 1019-1031. <https://doi.org/10.1037/0022-006X.55.4.584>
- Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley & Sons, Inc.
- Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology*, 34(2), 195-210. <https://doi.org/10.1177/0022022102250427>
- Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, 16(1), 20-30. <https://doi.org/10.1027//1015-5759.16.1.20>
- Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). *Rasch models: Foundations, recent developments, and applications*. New York: Springer Science & Business Media.
- Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, 75(2), 208-234. <https://doi.org/10.1177/0013164414536183>
- Goodman L. (2002) Latent class analysis In, Hagenaars J., McCutcheon A. (Ed.), *Applied latent class analysis* (pp. 3-18). Cambridge University Press: New York.
- Gülleroğlu, H. D. (2017). PISA 2012 matematik uygulamasına katılan Türk öğrencilerin duyuşsal özelliklerinin cinsiyete göre ölçme değışmezliğinin incelenmesi. [An investigation of measurement invariance by gender for the turkish students' affective characteristics whotook the PISA 2012 math test]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, 37(1), 151-175.
- Güngör, D., Korkmaz, M., & Somer, O. (2013). Çoklu-grup örtük sınıf analizi ve ölçme eşdeğerliği. [Multi-Group Latent Class Analysis and Measurement Equivalence]. *Türk Psikoloji Dergisi*, 28(72), 48-57.
- Güzeller, O.C. (2011). PISA 2009 Türkiye örnekleminde öğrencilerin bilgisayar özyeterlik inançları ve bilgisayar tutumları arasındaki ilişkinin incelenmesi. [An investigation of the relationship between students' computer self-efficacy beliefs and their computer attitudes in PISA 2009 turkey sampling] *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, 12(4), 183-203.

- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, 6(1) 53 – 60.
- Horn, J. L., McArdle, J.J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist*, 1(4), 179-188.
- Hui, C.H., & Triandis, H.C. (1985). Measurement in cross-cultural psychology: a review and comparison of strategies. *Journal of Cross-cultural Psychology*, 16(2), 131–152. <https://doi.org/10.1177/0022002185016002001>
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.
- Kankaras, M. (2010). *Essays on measurement equivalence in cross-cultural survey research: A latent class approach* (Unpublished doctoral dissertation).
- Kankaras, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. *Serbian Psychological Association*, 43(2), 121-136.
- Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, 40(2), 279-310. <https://doi.org/10.1177/0049124111405301>
- 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. <https://doi.org/10.14689/ejer.2016.66.22>
- Karasar, N. (2013). *Bilimsel araştırma yöntemi*. [Scientific research methods]. Ankara: Nobel Yayınevi.
- 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. <https://doi.org/10.1111/j.1745-3984.1990.tb00751.x>
- Köse, İ. A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (q32-q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi. [Examining the differential item functioning in the PISA 2009 student survey subscales (q32-q33)] *Kastamonu Eğitim Dergisi*, 23(1), 227-240.
- Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24(4), 492-516. <https://doi.org/10.1177/0049124196024004004>
- Li, M., Wang, M. C., Shou, Y., Zhong, C., Ren, F., Zhang, X., & Yang, W. (2018). Psychometric properties and measurement invariance of the brief symptom inventory-18 among chinese insurance employees. *Frontiers in psychology*, 9, 519. <https://doi.org/10.3389/fpsyg.2018.00519>
- Liang, L., & Lee, Y. H. (2019). Factor structure of the ruminative response scale and measurement invariance across gender and age among chinese adolescents. *Advances in Applied Sociology*, 9, 193-207. <https://doi.org/10.4236/aasoci.2019.96016>
- Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological methodology*, 31(1), 223-264. <https://doi.org/10.1111/0081-1750.00096>
- McCutcheon, A. L., & Hagenars, J. A. (1997). Comparative social research with multi-sample latent class models. *Applications of latent trait and latent class models in the social sciences*, 266-277.

- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <https://doi.org/10.1007/BF02294825>
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2), 195-215. <https://doi.org/10.1007/BF02295283>
- Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, 20(4), 303-320. <https://doi.org/10.1093/esr/jch026>
- Moors, G., & Wennekers, C. (2003). Comparing moral values in Western European countries between 1981 and 1999. A multiple group latent-class factor approach. *International Journal of Comparative Sociology*, 44(2), 155-172. <https://doi.org/10.1177/002071520304400203>
- Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26(3), 573-596. <https://doi.org/10.1057/palgrave.jibs.8490187>
- Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Oon Pey Tee., & R. Subramaniam (2018) Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education*, 40(3), 268-290. <https://doi.org/10.1080/09500693.2017.1413717>
- Ölçüoğlu, R. (2015). *TIMSS 2011 Türkiye sekizinci sınıf matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi* [The investigation of the variables that affecting TIMSS 2011 Turkey eight grade math achievement according to regions]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.
- Ölmez, İ. B., & Cohen, A. S. (2018). A mixture partial credit analysis of math anxiety. *International Journal of Assessment Tools in Education*, 5(4), 611-630. <https://doi.org/10.21449/ijate.455175>
- Önen, E. (2009). *Ölçme değişmezliğinin yapısal eşitlik modelleme teknikleri ile incelenmesi* [Examination of measurement invariance with structural equation modelling techniques]. Unpublished doctoral thesis, Ankara University, Ankara.
- Pishghadam, R., Baghaei, P., & Seyednozadi, Z. (2017). Introducing emotioncy as a potential source of test bias: A mixed Rasch modeling study. *International Journal of Testing*, 17(2), 127-140. <https://doi.org/10.1080/15305058.2016.1183208>
- Quandt, M. (2011). Using the mixed Rasch model in the comparative analysis of attitudes. *Cross-cultural analysis: Methods and applications*, 433-460.
- Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44(1), 75-92. <https://doi.org/10.1111/j.2044-8317.1991.tb00951.x>
- Rost, J., Carstensen, C., & Von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. *Applications of latent trait and latent class models in the social sciences*, 324-332.
- Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In *Rasch models* (pp. 257-268). Springer: New York, NY.

- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. <https://doi.org/10.1177/0013164413498257>
- Schnabel, D. B., Kelava, A., Van de Vijver, F. J., & Seifert, L. (2015). Examining psychometric properties, measurement invariance, and construct validity of a short version of the Test to Measure Intercultural Competence (TMIC-S) in Germany and Brazil. *International Journal of Intercultural Relations, 49*, 137-155. <https://doi.org/10.1016/j.ijintrel.2015.08.002>
- Sırgancı, G. (2019). *Karma rasch model ile değişen madde fonksiyonunun belirlenmesinde kovaryant (ortak) değişkenin etkisi*. [The effect of covariant (common) variable in determining the changing item function with mixed rasch model]. Unpublished doctoral thesis, Ankara University, Faculty of Education, Ankara.
- Silvia, P. J., Kaufman, J. C., & Pretz, J. E. (2009). Is creativity domain-specific? Latent class models of creative accomplishments and creative self-descriptions. *Psychology of Aesthetics, Creativity, and the Arts, 3*(3), 139-148. <https://doi.org/10.1037/a0014940>
- Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliliğinin yapısal eşitlik modellenmesi ve madde tepki kuramı kapsamında incelenmesi. [Examining measurement invariance with structural equation modeling and item response theory]. *Türk Psikoloji Dergisi, 24*(64), 61-75. <https://doi.org/10.14527/9786053188407.23>
- Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90. <https://doi.org/10.1086/209528>
- Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. [Structural equation modeling: basic concepts and lisrel applications]. *Türk Psikoloji Yazıları, 3*(6), 49-74.
- Şen, S. (2016). Applying the mixed Rasch model to the Runco ideational behavior scale. *Creativity Research Journal, 28*(4), 426-434. <https://doi.org/10.1080/10400419.2016.12299858>
- TIMSS (2011). TIMSS 2011 international database. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands. Retrieved January 10, 2020 from <http://timss.bc.edu/timss2011/international-database.html>.
- Tucker, K. L., Ozer, D. J., Lyubomirsk, S., & Boehm, J. K. (2006). Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Social Indicators Research, 78*(2), 341-360. <https://doi.org/10.1007/s11205-005-1037-5>
- Uyar, Ş. & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample]. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2*(3), 30-43
- Uzun, N. B. (2008). *TIMSS-R Türkiye örnekleminde fen başarısını etkileyen değişkenlerin cinsiyetler arası değişmezliğinin değerlendirilmesi* [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey sample]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.
- Uzun, B. & Ogretmen T. (2010). Fen başarısı ile ilgili bazı değişkenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi. [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey Sample]. *Eğitim ve Bilim, 35*(155), 26-35.

- von Davier M. (2001). WINMIRA 2001: Software and user manual. Available from: <http://208.76.80.46/~svfklumu/wmira/index.html>.
- von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. *In Rasch models* (pp. 371-379). Springer, New York, NY.
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, 12(3), 1-26. <https://doi.org/10.7275/mhqa-cd89>.
- Yalçın, S. (2019). Use of mixed item response theory in rating scales. *International Electronic Journal of Elementary Education*, 11(3), 273-278.
- Yandı, A. (2017). Ölçme eşdeğerliğini incelemede kullanılan yöntemlerin farklı koşullar altında istatistiksel güç oranları açısından karşılaştırılması [Comparison of the methods of examining measurement equivalence under different conditions in terms of statistical power ratios]. Unpublished doctoral thesis, Ankara University, Institutes of Social Sciences, Ankara.
- Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: Pisa 2012 örneği. [Examining the measurement invariance with different methods: Example of Pisa 2012] *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 13(1), 243-253. <https://doi.org/10.17860/mersinefd.305952>
- Yandı, A., Köse, İ. A., Uysal, Ö., & Oğul, G. (2017). PISA 2015 öğrenci anketinin (st094q01nast094q05na) ölçme değişmezliğinin farklı yöntemlerle incelenmesi. [Investigation of the PISA 2015 student survey (ST094Q01NA-ST094Q05NA) with the different methods of measurement]. Ankara: Pegem
- Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62(3), 297-330. <https://doi.org/10.1007/BF02294554>
- Yüksel, S. (2015). Ölçeklerde saptanan madde işlev farklılığının karma rasch modelleri ile incelenmesi [Analyzing differential item functioning by mixed rasch models which stated in scales]. Unpublished doctoral thesis, Ankara University, Institutes of Health Sciences, Ankara.
- Yüksel, S., Elhan, A. H., Gökmen, D., Küçükdeveci, A. A., & Kutlay, Ş. (2018). Analyzing differential item functioning of the Nottingham Health Profile by mixed rasch model. *Turkish Journal of Physical Medicine & Rehabilitation*, 64(4), 300-307. <https://doi.org/10.5606/tftrd.2018.2796>

6. APPENDIX

Table A1. *Items in the Mathematics Liking Scale*

Codes	Items - English	Items - Turkish
BSBM17A	I enjoy learning mathematics	Matematik öğrenirken eğleniyorum.
BSBM17B	I wish I did not have to study mathematics*	Keşke matematik çalışmak zorunda olmasam.*
BSBM17C	Mathematics is boring*	Matematik sıkıcıdır.*
BSBM17D	I learn many interesting things in mathematics	Matematik dersinde ilginç şeyler öğrenirim.
BSBM17E	I like mathematics	Matematiği severim.
BSBM17F	I like any schoolwork that involves numbers	Sayıların dâhil olduğu her okul işini severim.
BSBM17G	I like to solve mathematics problems	Matematik problemlerini severim.
BSBM17H	I look forward to mathematics class	Matematik derslerini dört gözle beklerim.
BSBM17I	Mathematics is one of my favorite subjects	Matematik favori dersimdir.

*Reverse scored items (TIMSS, 2015).

Investigation of measurement invariance of PISA 2015 collaborative problem solving skills: Turkey, Norway and Singapore

Yusuf Taner Tekin ^{1,*}, Derya Cobanoglu Aktan ¹

¹Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Feb 18, 2020

Revised: Sep. 08, 2020

Accepted: Jan. 06, 2021

Keywords

PISA Collaborative
problem solving skills,
Measurement invariance,
Multi-group confirmatory
factor analysis,
Cultural comparisons

Abstract: The purpose of this research is to examine measurement invariance of collaborative problem solving skills measured by PISA 2015 Xandar subtest for Singapore, Norway, and Turkey. The research was conducted with 2990 participants' data obtained from Turkey (1032), Norway (923), and Singapore (1035) on PISA 2015 collaborative problem solving study. In the first part of the study, exploratory factor analysis was performed to obtain the factor structure of the Xandar subtest. Then, the model data fit was checked by confirmatory factor analysis via χ^2 / df (3.127), RMSEA (0.027), CFI (0.987) and TLI (0.979) values. Multi-group confirmatory factor analysis was used in invariance analyses. The findings show that the collaborative problem solving model met only configural invariance across the countries and has not met the metric, scale, and strict invariance stages. The results show that meaningful comparisons cannot be made between the countries, because the factor loadings, variances, error variances, and covariances differ among countries.

1. INTRODUCTION

Global developments, demographic changes, and technological progress require certain changes in individuals' lives and and specific skills are needed in every field. These skills include communication, teamwork, leadership, taking initiative, literacy in mother tongue and a foreign language, competence in science, mathematics, and problem solving. Having these skills will enable individuals to be more successful in their daily, business and social lives. The acquisition of the mentioned skills can occur spontaneously in social life and is also acquired through education. However, these skills to be acquired through educational institutions should be transferred to daily life situations. At the same time, it is necessary to measure the level of acquisition of these skills and to plan educational policies according to these measures.

International exams such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS) aim to assess the transfer of acquired knowledge and skills in the fields of science, mathematics, and reading to daily life situations. PISA which is implemented by the Organization for Economic Cooperation and Development (OECD), is a pioneer test in this field. Moreover, PISA has developed different

CONTACT: Yusuf Taner Tekin ✉ tanerrr.tekin@gmail.com 📍 Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

assessment applications in recent years. In 2012, PISA started to assess individuals' financial literacy skills and in 2015 collaborative problem solving skills. The reason for that is nowadays especially the labor market requires individuals who are in dialogue with others, can communicate and solve problems collaboratively. The increasing demand for highly qualified individuals also emphasizes those who have these skills. With this in mind, the results obtained from PISA, also provide resources for developing specific policies for the countries on the quality of their education and their students' collaborative problem solving skills.

1.1. Collaborative Problem Solving Skills

PISA 2015 defines collaborative problem solving as “the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.” (OECD, 2017, p.34). According to Demir and Seferoğlu (2017) subjective structuring and transfer of knowledge, increasing emphasis on authentic learning and producing knowledge have led to the emergence of collaborative problem solving as well as problem solving skills. Nelson (2009) argued that collaborative problem solving has two structural components such as cooperative learning and problem-based learning. The author also states that collaborative problem solving provides students with experiences that create an intrinsic motivation for learning, questioning, collaborating and problem solving (Nelson, 2009). The nature of collaborative problem solving goes back to the work that O’Neil and his colleagues started in the 1990s to evaluate concepts in the best way and develop a theoretical framework and methodology. O’Neil, Chuang, and Chung (2003) have defined competencies similar to those used by PISA today. These competencies are grouped under five categories and expressed as the use of resources, interpersonal relations, information, systems, and technology. Thus, collaborative problem solving is process based on the contribution of both the cognitive and social skills of individuals involved in an activity (Hesse, 2017). In light of such developments, PISA implemented collaborative problem solving in 2015 and focused on solving the problem situations presented to individuals in a computer-assisted environment on a common understanding with one's teammates. It is meant that the computer accompanies the people participating in the application as virtual individuals (OECD, 2017). In the process of collaborative problem solving, PISA defined the following competences;

- establishing and maintaining shared understanding,
- taking appropriate action to solve the problem,
- establishing and maintaining team organization.

Furthermore, the capacity of the individual, doing the work with at least two or more people, and attempting to solve problems were identified as key competences (OECD, 2017). The theoretical development of the competencies identified by PISA is based on the topics of “computer-assisted collaboration, team discourse analysis, information sharing, individual problem solving, organizational psychology, and business context assessment”.

Collaborative problem solving research gained popularity in recent years. The recent examination of the concept is closely related to the fact that it is one of the skills sought after today. A recent study (Erkoç, 2018) investigated the effects of collaborative game design on various skills (-settings-) such as critical thinking, problem solving and algorithm development. Erkoç (2018) found that there is a significant difference in terms of the problem solving skills in favor of the group in which the collaborative game development approach is applied. At the same time, it was observed that there was also a significant difference in favor of the collaborative group between self-control factor, which is one of the problem solving skills. Uzunosmanoğlu (2013) conducted a study on the computer-assisted collaborative problem solving processes with dual eye-tracking. The study was conducted with 18 university students and focused on the participants' ability to discuss geometry problems with their teammates

using a collaborative approach. When the results are obtained, it was seen that the team members who collaborate more often achieve better results than the team members who collaborate less. In another study conducted by Özdemir (2005), the effects of individual and collaborative problem-based learning on critical thinking skills, academic achievement and attitude towards internet use were examined among 70 university students. It was found that there was a significant difference between the scores of using critical thinking skills according to the students' groups, and the researcher reported that this difference was in favor of the collaborative group. The results of these and similar studies show that collaborative problem solving is important for solving complex problems and critical skill for individuals to have.

1.2. Measurement Invariance

In the PISA studies, it was found that students' achievements are associated with certain variables. These are the variables that can directly or indirectly affect the achievement of individuals such as their socio-economic status, equality of opportunity, time devoted to learning, future academic expectations, and pre-school education. However, when the results obtained from these variables are compared, it is not right to attribute the differences that arise only to the characteristics of individuals and to environmental factors. Because these differences among individuals may not stem from the individuals themselves, but also the measurement tool too. Even though the language experts in different countries have made efforts to eliminate language-related differences, it is not guaranteed that the measurement tool will have the same meaning and be interpreted by individuals in different countries (Başusta, and Gelbal, 2015). Hence, this situation will make it impossible to carry out generalizability studies on the groups for the measurement results.

It is not desirable that the other traits interfere with the measurement results other than the trait that is intended to be measured. Otherwise, this can cause validity problem for the measurement results. The items in measurement instruments are expected to be interpreted in the same way without being affected by the other variables. When the studies conducted in Turkey were examined, it was observed that the studies on measurement invariance have increased in recent years. Invariance means that measurements administered to the different groups show equivalent or similar psychometric properties (Başusta, Gelbal, 2015). Uyar (2011) conducted a measurement invariance study on gender, statistical area, and school types by using the learning strategies model for PISA 2009 Turkey data. Bahadır (2012), on the other hand, used structural equation modeling (SEM) to examine the differences among the seven geographical regions of the reading skills model of PISA 2009 and concluded that the model was in good agreement with the data. In another study, Başusta and Gelbal (2015) examined the factor structure of the science technology-related items in the PISA 2009 student questionnaire and they tested these factors for measurement invariance in terms of gender. Research on the measurement invariance (cultural and country invariance) studies for PISA tests were also reported in the literature. For instance, Kibrıslıoğlu (2015) examined the measurement invariance based on culture and gender for PISA 2012 mathematics test for Turkey, China (Shanghai) and Indonesia data. The results of the research showed that the model holds the configural invariance stage among countries but does not hold the metric, scalar and strict invariance stages. Lately, Karakoç Alatlı (2016) studied the measurement invariance for Australia, France, Shanghai-China, and Turkey for PISA 2012 literacy test.

In terms of the measurement invariance studies that were carried out of Turkey, Greif, Wüstenberg, Molnar, et al (2013) studied the measurement invariance of complex problem-solving skills models over the grade level by using the Hungarian students' data. Oliden and Lizaso (2013), on the other hand, examined the measurement invariance of four different language forms, Spanish, Galician, Catalan, and Basque, on the data from the Spanish sample of PISA 2009 reading skills. Findings showed that the scores obtained from different language

forms do not exhibit invariance property. Another study by Wu, Liu and Zumbo (2007) tried to explain why the strict invariance stage is necessary for measurement invariance. For this purpose, the authors examined the countries such as the United States of America, Canada, Australia, New Zealand, and also the countries with similar cultures like Taiwan, Korea, and Japan by using TIMSS 1999 math tests. In this context, the researchers examined the measurement invariance by making 21 comparisons among and within various cultures.

The review of the previous studies shows that the measurement invariance is not always met and therefore before making comparisons, invariance studies should be performed. In particular, it is suggested and important to examine invariance if the results of the research are going to be/expected to be used in shaping educational policies. In addition, as in the case of PISA in the international arena, it is also necessary to show how different groups interpret the test applied to collaborative problem solving skills, which are among the essential critical skills of our time.

When a measurement tool is applied to groups with different characteristics, errors can be encountered in interpreting the results obtained if the effects of the demographic characteristics cannot be eliminated. However, errors encountered here cannot be attributed to only a single group membership. This could originate from the measurement tool. Cheung and Resvold (2002) state that differences can be explained not only by individual characteristics but also by measurement tools. The basic problem that is desired to be solved in the measurement invariance is whether the measurements of the same properties, measured with the same measurement tool, could change in different observations and working conditions of a given situation. If there is no such evidence of measurement invariance, it would not be right to make a scientific inference. In such a case, hence, it would not be correct to interpret the findings of differences between individuals and groups clearly (Mark and Wan, 2005). To make a comparison between groups on measurement results, measurement invariance must exist. To have a measurement invariance, the relationship between observed and latent variables must be the same in different groups (Karakoç Alatlı, 2016). According Millsap and Kwok (2004), to meet the invariance, the likelihood of getting a certain score is equal for individuals belonging to different groups whose similar characteristics are measured in the test. However, the most important feature sought in a measurement tool is validity and validity evidence. Therefore, accurate evidence on the validity of the scores obtained from measurement tools also necessitates measurement invariance studies (Yandı, Köse & Uysal, 2017).

Different definitions of measurement invariance can be found in the literature. Bryne and Watkins (2003) define invariance as the interpretation and perception of the scale by individuals of different groups in the same way. On the other hand, Raju, Laffittle and Byrne (2002) define invariance as getting the same score by different groups in terms of the characteristics measured by the scale. In other respects, measurement invariance can be realized in different cases or comparison of sub-sample groups of the same population. That is, the measurement invariance shows the comparability of the same structure in different cultural groups, the variance of the variables can be estimated independently from the group and the comparability of latent mean, variance and covariances of different groups (Bahadır, 2012). The comparisons here test the hypothesis of intergroup differences rather than the intra-group invariance of the model (Lance and Vandenberg, 2000). The main purpose of such studies is to use the measurements based on equality between groups. However, the measurement tools are prepared with the assumption of 'different groups measure the same property'. If this assumption is confirmed, the accuracy of scoring and analysis will be meaningful. If this assumption cannot be verified, the analysis and the results obtained will lose their significance. In other words, the measurement model shows the same structure in more than one group. This means that the factor loadings of the scale

items, the correlations between the factors and the error variances are the same (Bollen, 1989; Byrne, 2004; Jöreskog & Sörbom, 1993).

Collaborative problem solving skills were highlighted and stated as critical skills for today's well educated students in the 2015 PISA assessment. Thus, collaborative problem solving skills were important components of the PISA 2015 collaborative problem solving test. From this point of view, it is critical to test the validity of the results and the comparability of the measurement model formed by the collaborative problem solving skills in the light of the PISA data, which offers a large sample and cross-country comparisons. The countries (Singapore, Norway, Turkey) in this study were selected based on their high, medium and low scores respectively in the PISA 2015 collaborative problem solving test. To solve the problems and sub-problems determined within the scope of this research, the steps of measurement invariance by Multi Group Confirmatory Factor Analysis (MGCFA) method were examined for the paired country groups respectively. So, we aimed to answer the problem of "do the collaborative problem solving PISA 2015 data hold the measurement invariance for the countries (Singapore, Norway, Turkey)?" Moreover, the following sub-problems were also examined in this study.

1) Do Singapore - Norway, Singapore - Turkey, and Norway - Turkey measurement models show;

- (a) configural invariance,
- (b) metric invariance,
- (c) scalar invariance, and
- (d) strict invariance?

2) If the invariance cannot be achieved, what are the relevant parameters for the invariance stages?

2. METHOD

This study is carried out to examine whether measurement invariance for collaborative problem solving PISA 2015 Xandar subtest data is met for Singapore, Turkey, and Norway groups. In this study, the data obtained from the OECD official website (<https://www.oecd.org/pisa/data/2015database/>) were used and no data collection was performed. According to the data characteristics, the research is a quantitative and a correlational study because it examines the relationship of observed variables with latent variables.

2.1. Data Characteristics

PISA 2015, the sixth round of PISA, was implemented in 2015 with the participation of approximately 540,000 students, representing approximately 29 million students in 72 countries and economies. 35 of these participating countries are OECD members. Within the scope of the study, the Xandar subtest, which is one of the six different subtests in which the cooperative problem solving skills are measured, was selected by purposeful sampling method as one of the non-probable sampling methods. As seen in [Table 1](#) the number of individuals who answered the Xandar subtest was 1035 for Singapore, 923 for Turkey, and 1032 for Norway, and a total of 2990 individuals. Testing whether the measurement model created with collaborative problem-solving skills in the light of PISA data has the same structure for different countries will ensure the validity of the results and the significance of the comparisons. Here, we mean the countries with high, medium and low scores in the collaborative problem solving test scores of the PISA 2015 application. Therefore Singapore, Norway and Turkey have been selected.

Table 1. Number of PISA 2015 and Xandar Subtest Participants by Country.

Country	PISA 2015		Xandar Subtest	
	Number of participants	Percentage	Number of participants	Percentage
Norway	5.456	31.2	923	30.9
Singapore	6.115	35.0	1.035	34.6
Turkey	5.895	33.8	1.032	34.5
Total	17.466	100.0	2.990	100.0

In this study, the Xandar subtest was selected because its questions were published as examples, explanations were made according to the proficiency levels of these questions and had a sufficient sample size of data. The Xandar section starts with a general explanation. In this explanation, it is stated that each person will work with three teammates. However, the teammates expressed here are virtual persons. At the next stage, it is aimed to determine how individuals understand and solve the problem together with their team members in the face of three different situations. Following the instructions in the introduction of the test, participants are expected to proceed to the next stage by selecting one of the possible answers that appear on the screen, based on the comments they made by the virtual teammates. Here, one of the expressions chosen from the possible answers is correct (1) and the others are incorrect (0). The second screen, according to the individual's response to the event on the first screen and the views of the virtual persons about the event, appears on this screen. Thus, the individual completes the section each time by selecting one of the possible answers to continue the plot. More information about the Xandar subtest can be found on the OECD's website. (<https://www.oecd.org/pisa/test/cps-xandar-scoring-guide.pdf>).

Table 2. Collaborative Problem Solving Competencies for Items.

Item	CPS	
	Skills	Description
m1	C3	Following rules of engagement (e.g., prompting other team members to perform their tasks)
m2	C1	Communicating with team members about the actions to be/being performed
m3	B1	Building a shared representation and negotiating the meaning of the problem (common ground)
m4	B1	Building a shared representation and negotiating the meaning of the problem (common ground)
m5	B3	Describing roles and team organization (communication protocol/rules of engagement)
m6	A1	Discovering perspectives and abilities of team members
m7	B3	Describing roles and team organization (communication protocol/rules of engagement)
m8	C3	Following rules of engagement (e.g., prompting other team members to perform their tasks)
m9	D1	Monitoring and repairing the shared understanding
m10	D2	Monitoring results of actions and evaluating success in solving the problem
m11	D3	Monitoring, providing feedback and adapting the team organization and roles

The Xandar subtest includes 12 items, but one of the items is scored differently from the “1-0” form and therefore it was not included in the study. The study was performed with 11 items. Table 2 contains the levels and descriptions of these items. They were coded as CC100101 in the original data set and these codes were changed to m1, m2, ... and m11 for the convenience of the analysis and interpretation of the data. The collaborative problem solving competencies of these items are as follows:

- At level 1, the items (m2, m3, m4, m6, and m9), establishing and maintaining shared understanding
- At Level 2, item 10 (taking appropriate action to solve the problem)
- At level 3, the items (m1, m5, m7, m8, and m11), establishing and maintaining team organization.

Students who respond correctly to the items in level 1 are expected to explore the perspectives and abilities of their teammates, discuss a problem on shared ground, and communicate with

team members about the actions to be taken, and monitor and evaluate the actions they take in this direction. Students who answer the items in level 2 are expected to discover the type of communication they will perform to solve the problem, define the tasks to be completed, and monitor and evaluate the actions they perform as in the first level. Students who respond to the top 3 and top-level items correctly are expected to understand the roles for solving the problem, define the roles, follow the agreement rules set out in this direction, and follow and evaluate the team organization and roles, and give feedback.

2.2. Data Analysis

Measurement invariance is analyzed in stages. Four stages need to be tested to ensure that the invariance is fully achieved. These stages are configural invariance, metric invariance, scalar invariance, and strict invariance (Meredith, 1993). Configural invariance; tests for identical factor structures for different groups; metric invariance checks equality of the factor loadings; scalar invariance tests equality of intersection points at regression equation; strict invariance refers to the invariance of residual load variance (Brown, 2015). The invariance stages were tested with the Mplus 7 analysis program and it was decided whether the invariance stages were achieved by taking the fit indices χ^2 , RMSEA, CFI and TLI as reference. While conducting the MGCFA, one of the groups was taken as a basis and the values of the group were fixed at each stage, and the level of adaptation of the values of the other group to the fixed group was examined. The group whose values are kept constant is called the reference group, and one of the countries was chosen as the reference group in each analysis for the paired groups in the study. In addition to examining whether the fit indices are within the accepted range, the difference of CFI and TLI values compared to the less constrained model in the invariance stages were examined. If this difference is between -0.01 and 0.01, it has been taken into account that it is acceptable level for transition to the next stage (Cheung & Resvold, 2002). The invariance phases start with the structural invariance phase and if the fit indices are at an acceptable level, the next analyses were done. The level of change in chi-square, CFI and TLI values compared to the previous stage is discussed in the next stages after structural invariance. Before doing these analyses, the assumptions necessary for the analyses were checked. After that, the factor structure of the problem solving data were examined. After analyzing the factor structure, the collaborative problem solving model was confirmed by confirmatory factor analysis, and finally, the measurement invariance of the model was tested through Multiple Groups Confirmatory Factor Analysis (MGCFA).

In terms of assumptions, the missing values and multicollinearity were examined. For multicollinearity, tolerance and variance inflation were examined. After that exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed Kaiser-Meyer Olkin (KMO) and Barlett Sphericity Tests were used to investigate the suitability of the data set to EFA. In the CFA MGCFA model and data fit levels were examined by χ^2 / df , RMSEA, CFI and TLI indices.

Before analyzing the data, it is important to consider whether the data set is suitable for the analysis and whether missing data and multicollinearity are affecting the data set. The analysis of CFA and EFA were done by using MPLUS packages (WLSMV) which are employed with dichotomous (1-0) data. There were no missing data. For the multicollinearity assumption, tolerance values and variance inflation factor values (VIF) were examined, separately for each factor. These values are given in Table 3. When Table 3 was examined, it is seen that all tolerance values are greater than 0.01, and variance inflation factor values are less than 10, which shows that there was no multicollinearity.

After checking assumptions, EFA was employed with 11 items of the Xandar subtest of collaborative problem solving skills. The distribution of the items to the factors and the

corresponding collaborative problem solving competencies indicated in the PISA Final Report were examined with EFA analysis.

Table 3. *Tolerance and Variance Inflation Values.*

Factor	Item	VIF	Tolerance
f1	m2	1.193	0.838
	m3	1.122	0.891
	m4	1.119	0.894
	m6	1.103	0.907
f2	m5	1.072	0.933
	m7	1.087	0.920
	m8	1.042	0.960

KMO and Barlett Sphericity Tests were used to determine the suitability of the data set for the EFA. The KMO value indicates whether the data matrix is suitable for factor analysis and is expected to be greater than 0.60. The Barlett sphericity test examines whether there is a relationship between variables based on partial correlations and the chi-square value calculated here is expected to be significant (Çokluk, Şekercioğlu, & Büyüköztürk, 2015). KMO and Barlett's values indicate that the data set is appropriate for EFA. EFA is an analysis based on correlation or covariance matrix. For this reason, when the EFA with 1-0 data patterns is desired, the correlation matrix should be tetrachoric. Since the data characteristics in this study were of 1-0 structure, an analysis was performed by the tetrachoric correlation matrix. EFA analysis started with 11 items, but the items (m1, m9, m10 and m11) with low factor loadings (<0.3), were excluded from the analysis. The analysis was continued with the remaining seven items. The analysis results in Table 4 show that seven items were collected in two factors. The items in the first factor (f1) are m2, m3, m4, and m6. The items in the second factor are m5, m7, and m8. The item distributions obtained in the factors also align with the competencies in the PISA final report. The PISA report is also used for naming the factors. Accordingly, f1 is called as “Common Understanding”, and f2 is “Team Organization. Factor loadings of the items collected under the Common Understanding factor and the Team Organization factor are presented in Table 4.

Table 4. *Item Factor Loadings.*

Item	Factor Loadings	
	f1	f2
m2	0.703	0.285
m3	0.502	0.277
m4	0.537	0.168
m6	0.512	0.153
m5	0.198	0.503
m7	0.269	0.655
m8	0.129	0.424

Collaborative problem solving model, which was put forward by EFA, was confirmed by CFA. The obtained model is shown in Figure 1.

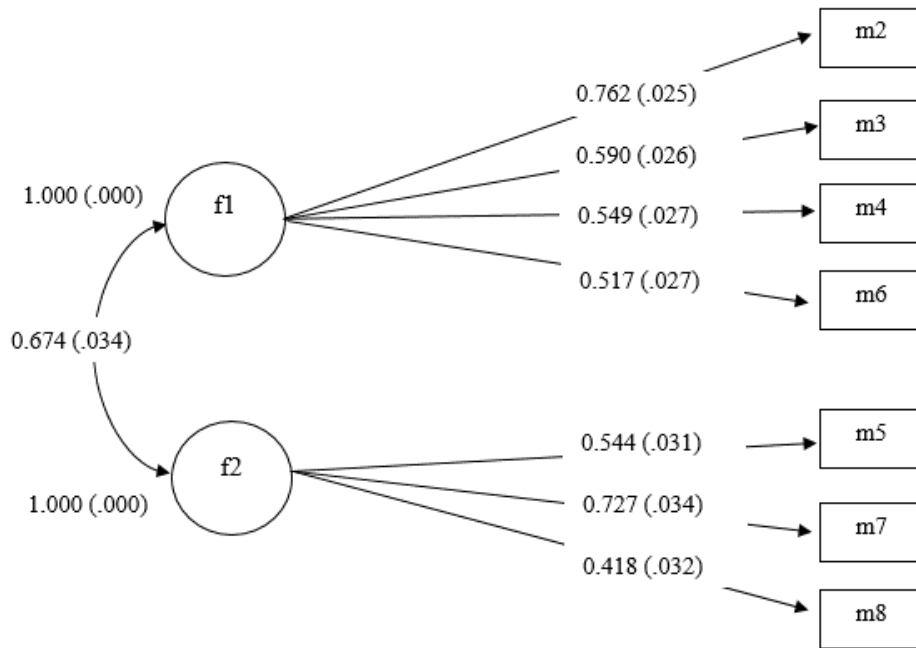


Figure 1. Collaborative Problem Solving Model.

The CFA was performed with Mplus 7 program and the model data fit was examined by referring to the indices indicated in Table 5.

Table 5. Acceptable Levels of Goodness of Fit Indices.

Fit Indices	Acceptable fit	Good fit
χ^2	$2df < \chi^2 \leq 3df$	$0 \leq \chi^2 \leq 2df$
χ^2 / df	$2 < \chi^2 / df \leq 8df$	$0 \leq \chi^2 / df \leq 2$
RMSEA	$0.05 < RMSEA \leq 0.08$	$0 \leq RMSEA \leq 0.05$
TLI	$0.95 \leq NNFI < 0.97$	$0.97 \leq NNFI \leq 1.00$
CFI	$0.95 \leq CFI < 0.97$	$0.97 \leq CFI \leq 1.00$

(Schermelleh and Moosbrugger, 2003; Tabachnick and Fidell, 2007)

Collaborative problem solving skills model and tested CFA model results for Singapore, Norway and Turkey subgroups are listed in Table 6.

Table 6. Collaborative Problem Solving Model and Model Fit Indices of Subgroups.

Models	χ^2	$\chi^2 (p)$	χ^2 / df	RMSEA	CFI	TLI
Collaborative Problem Solving Model	40.657	0.000	3.127	0.027	0.987	0.979
Singapore	20.509	0.083	1.577	0.024	0.987	0.979
Norway	18.743	0.131	1.441	0.022	0.990	0.984
Turkey	22.363	0.050	1.720	0.026	0.961	0.936

When Table 6, which includes model fit indices for collaborative problem solving model, is examined, it can be said that model data fit level shows a good fit for $p = 0.05$ significance level. When subgroups elaborated separately, chi-square value $p = 0.05$ level of significance for Singapore is $0.083 > 0.05$ for Norway is $0.131 > 0.05$ not meaningful, but Turkey = $0.05 = 0.05$ is significant. In addition, to control the effect of sample size χ^2 / df , and goodness of fit indices RMSEA, CFI and TLI were also examined. While each of the discussed indices showed

a good fit for Singapore and Norway, for Turkey χ^2/df and RMSEA showed a good fit, CFI and TLI values indicated acceptable fit.

In the next step of the study, MGCFA was used to reveal the effects of unobservable structures on observable variables. One of the groups was considered as a reference at the MGCFA, and the values of this group were fixed at each stage and the level of fit of the values of the other groups was examined accordingly. The group whose values were constant is called the reference group, and in each analysis, one of the countries was selected as the reference group for the binary groups. In addition to examining whether the fit indices were within the acceptable range, the differences between CFI and TLI values were examined according to the less restricted model at the invariance stages. If this difference is between -0.01 and 0.01, it is considered to be an acceptable level for the transition to the next stage (Cheung and Resvold, 2002).

3. RESULT / FINDINGS

In this section, the findings are presented in the order of research problems. Findings related to the configural, metric, scalar, and strict invariance of PISA 2015 collaborative problem solving model are presented respectively. Firstly, the findings for the related countries regarding configural invariance are shown in [Table 7](#).

Table 7. *Configural Invariance Findings.*

Configural invariance	χ^2	<i>df</i>	RMSEA	CFI	TLI
Singapore-Norway	1224.328	42	0.025	0.986	0.979
Norway-Turkey	882.754	42	0.026	0.978	0.967
Singapore-Turkey	879.747	42	0.025	0.978	0.967

The configural invariance of the collaborative problem solving measurement model was tested at this stage. When [Table 7](#) is examined, it is seen that for Singapore and Norway RMSEA = 0.025 < 0.05. 0.97 < CFI = 0.986 < 1 and 0.97 < TLI = 0.979 < 1 and these values show good fit levels. For Norway and Turkey while RMSEA = 0.026 < 0.05. 0.97 < CFI = 0.978 < 1 and 0.95 < TLI = 0.967 < 0.97 values show a good fit for RMSEA and CFI, for TLI index, the fit is considered acceptable. Lastly for Singapore and Turkey RMSEA = 0.025 < 0.05. 0.97 < CFI = 0.978 < 1 and 0.95 < TLI = 0.967 < 0.97, RMSEA, and CFI show a good fit. However, for TLI index, it is only at the acceptable level. These findings for Singapore-Norway, Norway-Turkey and Singapore-Turkey groups demonstrate that the model met the configural invariance. Since the configural invariance is a prerequisite for metric invariance, then the next stage for metric invariance has been tested for all three groups. The fit indices are presented in [Table 8](#) for this purpose.

In order to obtain evidence of metric invariance, item factor loadings were examined in addition to item factor structures. Singapore-Norway, Norway-Turkey, and Singapore-Turkey group analyses results were presented separately. As indicated in [Table 8](#), for Singapore and Norway, RMSEA = 0.031 < 0.05. 0.97 < CFI = 0.974 < 1, 0.95 < TLI = 0.967 < 0.97, RMSEA and CFI indices show a good fit and an acceptable fit for TLI. For Norway and Turkey RMSEA = 0.044 < 0.05. CFI = 0.924 < 0.95. TLI = 0.903 < 0.95 while RMSEA show a good fit. CFI and TLI indices are only at acceptable level. For Singapore and Turkey, as indicated in [Table 8](#) RMSEA = 0.028 < 0.05, 0.95 < CFI = 0.967 < 0.97, 0.95 < TLI = 0.958 < 0.97 as in the previous comparison RMSEA showed a good fit but CFI and TLI indices were only at acceptable level.

Table 8. Metric Invariance Findings.

Scalar invariance	χ^2	df	RMSEA	CFI	TLI	χ^2 diff. test	Δdf	ΔCFI	ΔTLI
Singapore-Norway	1224.328	42	0.031	0.974	0.967	15.691 ($p=0.0078$)	0	-0.012	-0.012
Norway-Turkey	882.754	42	0.044	0.924	0.903	38.078 ($p=0.000$)	0	-0.054	-0.064
Singapore- Turkey	879.747	42	0.028	0.967	0.958	13.504 ($p=0.0191$)	0	-0.011	-0.009

For Singapore-Norway, although the fit indices were found to be a good fit for RMSEA and CFI, and acceptable for TLI index, chi-square ($\Delta\chi^2$) difference test results between the two models were found to be $p = 0.0078 < 0.05$. In other words, models for Singapore and Norway groups differ significantly from each other. In addition, when ΔCFI and ΔTLI values are examined, it is observed that they are not in the range of -0.01 to 0.01, which is accepted for the transition to the next stage (scalar invariance). The obtained ΔCFI and ΔTLI values were found to be the same and -0.012. For Norway-Turkey RMSEA, although they present a good level of fit CFI and TLI has presented index values outside the acceptable range. Additionally, chi-square ($\Delta\chi^2$) $p = 0.000 < 0.05$ of the difference test is significant thus the models for Norway and Turkey have been found to significantly differ from each other. When ΔCFI and ΔTLI values were examined, it was observed that they were not in the range of -0.01 to 0.01. The obtained ΔCFI and ΔTLI values are -0.054 and -0.064 respectively. For Singapore and Turkey RMSEA showed a good level of fit, but CFI and TLI indices are only at the acceptable level. However, as is clear from Table 8, chi-square ($\Delta\chi^2$) $p = 0.0191 < 0.05$ of the difference test result is significant therefore; the model for Singapore and Turkey group has been found to significantly differ from each other. When ΔCFI and ΔTLI values were examined, it was found that these values are not in the specified range of -0.01 and 0.01. The obtained ΔCFI and ΔTLI values are -0.011 and -0.009 respectively.

The chi-square difference test results presented in the findings were obtained by a two-step approach using the DIFFTEST option in the Mplus analysis program (Wang and Wang. 2012). The findings show that models for Singapore-Norway. Turkey-Norway Singapore-Turkey groups did not show the metric invariance step. This reveals that the PISA 2015 collaborative problem solving test might have been affected by the other variables for these countries.

Table 9. Item Factor Loadings and Thresholds for Singapore and Norway (Configural).

Item	Factor Loadings		Item	Thresholds	
	Singapore	Norway		Singapore	Norway
M2	0.703	0.707	M2\$1	-0.683	-0.160
M3	0.423	0.625	M3\$1	-0.755	-0.254
M4	0.496	0.585	M4\$1	-0.316	0.151
M5	0.618	0.707	M5\$1	-0.423	-0.736
M6	0.498	0.571	M6\$1	-0.339	0.384
M7	0.700	0.527	M7\$1	-0.928	-1.100
M8	0.368	0.253	M8\$1	-0.788	-0.957

Considering that metric invariance is a prerequisite for scalar invariance and the findings are significant at 0.05 level and metric invariance does not hold, the analysis did not proceed to the next stage of invariance. In the second stage of the study, to investigate which items differ from

each other, the factor loadings of the items and the threshold values for country groups were examined and the findings are presented in Tables 9, 10, and 11.

When the factor loadings were examined for Norway and Singapore, it was observed that the differences were large for items m3, m4, m5, m6, m7, and m8. When we consider the content of these items, the participants of these two countries; “discuss the meaning of the problem on a common basis for the solution of an existing problem” (m3 and m4); “establish team organization and team rules” (m5 and m7); “explore different team members' perspectives and abilities” (m6) and “ask other team members to perform their duties” (m8). On the other hand, the item thresholds in Table 9 also contain interesting findings. The magnitude of the negative item thresholds shows that item's easiness and for positives vice versa. In this respect, when the threshold values in the table above are examined, it is observed that the items m2, m3, m4 in the instrument were easier for Singapore and the other items were easier for Norway.

Table 10. *Item Factor Loadings and Thresholds for Norway and Turkey (Configural).*

Item	Factor Loadings		Item	Thresholds	
	Norway	Turkey		Norway	Turkey
M2	0.707	0.683	M2\$1	-0.160	0.156
M3	0.625	0.337	M3\$1	-0.254	0.367
M4	0.585	0.471	M4\$1	0.151	0.316
M5	0.707	0.336	M5\$1	-0.736	-0.075
M6	0.571	0.402	M6\$1	0.384	0.301
M7	0.528	0.869	M7\$1	-1.100	-0.200
M8	0.253	0.252	M8\$1	-0.957	-0.160

When Table 10 for Norway and Turkey is examined, it is observed that factor loadings for items m3, m4, m5, m6, and m7 differ from each other in a relatively big magnitude. In terms of content of the items, it was noted that the participants of the two countries differed in their interpretations regarding “discussing the meaning of the problem on a common basis” (m3 and m4); “establishing team organization and team rules” (m5 and m7); and “exploring the perspectives and abilities of different team members for the solution of an existing problem” (m6). At the same time, when the item thresholds are examined items m2, m3, m4, m5, m7, and m8 are quite easy for the Norwegian participants than participants in the Turkey sample. The only item, which is easy for Turkey sample participants was item m6.

Table 11. *Item Factor Loadings and Thresholds for Singapore and Turkey (Configural).*

Item	Factor Loadings		Item	Thresholds	
	Singapore	Turkey		Singapore	Turkey
M2	0.707	0.683	M2\$1	-0.683	0.156
M3	0.422	0.337	M3\$1	-0.755	0.368
M4	0.497	0.471	M4\$1	-0.316	0.316
M5	0.707	0.336	M5\$1	-0.423	-0.075
M6	0.497	0.402	M6\$1	-0.339	0.301
M7	0.692	0.870	M7\$1	-0.928	-0.201
M8	0.367	0.252	M8\$1	-0.788	-0.160

When we compare factor loadings for Singapore and Turkey, significant differences are observed at items m3, m5, m6, m7, and m8. This finding is similar to the findings of the comparisons of two groups (Singapore- Norway, and Norway-Turkey). In addition, in terms of

the item thresholds, all items were easier for the participants of the Singapore than that of Turkey.

When a general evaluation was made on the differences of the items, it was observed that items m3, m5, m6, and m7 differed in all three comparisons. In other words, it can be said that there are differences in terms of discussing the meaning of the problem on common ground, establishing roles and team organization, exploring team members' perspectives and following the rules of the agreement.

4. DISCUSSION and CONCLUSION

According to the findings, while the configural invariance was achieved in all three groups, the metric invariance could not be achieved. Since the metric invariance stage was not achieved, scalar and strict invariance stages were not tested. Therefore, it was concluded that factor structures were the same in all three groups but factor loadings, variances, error variances, and covariances differed. This result shows that the participants of the countries (Singapore, Norway, and Turkey) interpreted the Xandar subtest of the collaborative problem solving skills test differently.

To be able to compare country scores, the established model must hold measurement invariance. However, the findings show that measurement invariance does not hold for the data in this study. The findings show that in country comparisons, factor loadings of m3, m5, m6, and m7 differed from each other. It can be said that these differences can be one of the reasons for not completing all the stages of measurement invariance. These differences in factor loadings may mean that there is a difference in participants' interpretations of these items. The competencies measured in these items are: understanding and discussing the meaning of an existing problem, establishing team rules, and exploring the perspectives and abilities of team members. The differences in the results obtained from the measurement tool show that the participants of this country interpret the items related to these competencies differently.

Considering that, information is globalized and individuals with critical skills are sought after, countries need to become equivalent in this field with other countries. However, the PISA 2015 results show that the scores among the top, middle and low group countries differ significantly from each other in terms of collaborative problem solving skills (OECD, 2017). The fact that the invariance stages cannot be fully achieved is another indication of this. There is also variability between these countries due to unobservable variable(s). This situation leads to the differentiation of the countries in this field due to different reasons and the result that some countries raise competent individuals in terms of collaborative problem solving skills while others are left behind in terms of these skills.

An important contribution of this study to the literature is that its contribution to the collaborative problem solving on the literature. Therefore, there is no measurement invariance on collaborative problem solving research that can be compare to our results with the literature. For the first time in 2015, the OECD conducted a collaborative problem-solving study. Therefore, the results obtained by comparisons of different countries that are made within the scope of this research are of particular importance. On the other hand, although this is the first study in the field, studies are documenting that measurement invariance is not achieved in large scale studies such as PISA and TIMSS. For instance, Kırışlıoğlu (2015) found that only configural invariance stage was achieved for mathematics literacy in PISA 2012 Turkey, China-Shanghai, and Indonesia data. Similarly, Karakoç Alatl (2016) for PISA 2012 mathematical literacy and scientific literacy data of Australia, France, China-Shanghai, and Turkey sample only met the configural stage. As a final example, Wu, Liu, and Zumbo (2007) conducted a study using TIMSS data from the USA, Canada, Australia, New Zealand, Taiwan, Korea, and

Japan and their results showed that only structural and metric invariance stages hold for the data.

Especially in the studies carried out with many countries, the invariance stages must be fully hold for comparisons to be meaningful. For this reason, researchers should examine not only descriptive statistics but also invariance. This study was conducted with the countries in the upper, middle, and lower groups. In addition, invariance studies should be conducted for countries whose scores are not very different from each other. On the other hand, when the literature is examined, the financial literacy test administered in PISA 2012 application is as important as collaborative problem solving skills. In this sense, it is important for the researchers to examine the state of invariance related to financial literacy test on a country-by-country basis and to conduct cross-cultural invariance studies.

Within the scope of this study, only Xandar subtests were examined from six different subtests for collaborative problem solving skills. For this reason, researchers can conduct invariance studies of the other five subtests on different subgroups belonging to different countries and within the same country will contribute to both measurement invariance and collaborative problem solving literature. Another important point is that item bias should be examined in addition to invariance studies. Identifying the factors that cause bias will allow for the purely measurement applications of these factors and to give reliable results.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Authors are expected to present author contributions statement to their manuscript such as; **Yusuf Taner Tekin**: Investigation, Data Analysis, Methodology, Resources, Writing. **Derya Çobanoğlu-Aktan**: Investigation, Data Analysis, Methodology, Resources, Writing, Supervision and Validation.

ORCID

Yusuf Taner TEKİN  <https://orcid.org/0000-0001-9068-7894>

Derya ÇOBANOĞLU AKTAN  <https://orcid.org/0000-0002-8292-3815>

5. REFERENCES

- Bahadır, E. (2012). *According programme for international student assessment (PISA 2009), investigation of variables that affect Turkish students' reading skills by regions* [Master Thesis]. Hacettepe University, Ankara.
- Başusta, N.B., & Gelbal, S. (2015). Examination of Measurement Invariance at Groups' Comparisons: A Study on PISA Student Questionnaire. *Hacettepe University Journal of Education*, 30 (4), 80-90
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley- Interscience Publication.
- Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural equation modeling*, 11(2), 272-300.
- Bryne, B. M. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal Of Cross-Cultural Psychology*, 34, (2), 155-175.
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. (Second Edition). The Guilford Press.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-Fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.

- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Pegem.
- Demir, Ö., & Seferoğlu, S.S. (2017). İşbirlikli Problem Çözmenin Kodlama Eğitime Yansımaları Olarak Eşli Kodlamanın İncelenmesi. *5th International Instructional Technologies & Teacher Education Symposium*. 11-13 October 2017, İzmir.
- Erkoç, M.F. (2018). *The effect of collaborative game design on critical thinking, problem solving and algorithm development skills* [Doctoral Dissertation]. İstanbul University.
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts - something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105(2), 364-379.
- Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assesment. *Journal of Educational Measurement*, Spring, 54(1), 12-35.
- Jöreskog, K. G., & Sörbom, D. (1993). *Lisrel 8: Structural equation modeling with the simplis command language*. Lincolnwood: Scientific Software International, Inc.
- Karakoç Alatlı, B. (2016). *Investigation of measurement invariance of literacy tests in the Programme for International Student Assessment (PISA - 2012)* [Published Doctoral Dissertation]. Ankara University.
- Kıbrıslıoğlu, N. (2015). *The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey - China (Shangai) – Indonesia* [Master Thesis] Hacettepe University.
- Lance, C. E., & Vandenberg, R. J., (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1), 4-70.
- Mark, B. A., & Wan, T.T.H (2005). Testing measurement equivalence in a patient satisfaction instrument. *Western Journal of Nursing Research*, 27(6), 772-787.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Pyschometrika*, 58, 525-543.
- Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological methods*, 9(1), 93.
- Nelson, L.M. (2009). *Instructional-design theories and models: A new paradigm of instructional theory, Volume II*. Routledge Publishing, (ISBN 978-0-8058-2859-7)
- OECD, 2017. *PISA 2015 results (Volume 5): Collaborative problem solving*. OECD Publishing.
- O'Neil, H. F., Chuang, S. H., & Chung, G. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education: Principles, Policy & Practice*, 10(3), 361-373.
- Oliden, P., E., & Lizaso, J, M. (2013) Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain, *Psicothema*, 25(3), 390-395.
- Özdemir, S. (2005). *The effects of individual and collaborative problem-based learning using an online asynchronized learning tool on critical thinking abilities, academic achievements, and attitudes toward internet use* [Doctoral Dissertation]. Gazi University.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, 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.

- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Boston, MA: Pearson.
- Uyar, Ş. (2011). *An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample* [Master Thesis] Hacettepe University.
- Uzunosmanoğlu, S.D. (2013). *Examining computer supported collaborative problem solving processes using the dual-eye tracking paradigm* [Master's Thesis]. Orta Doğu Teknik Üniversitesi.
- Wang, J., & Wang, X. (2012). *Structural equation modelling: Applications using mplus. (First edition)*. UK: Wiley Publications.
- Wu, D. A., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26.
- Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı Yöntemlerle Ölçme Değişmezliğinin İncelenmesi: Pisa 2012 Örneği. *Mersin University Journal of the Faculty of Education, 13*(1), 243-253.

Teachers' Attitudes and Opinions about Design and Skill Workshops and Ranking of Workshops by Teachers

Aysegul Bayraktar ¹, Seher Yalcin ^{1,*}

¹Ankara University, Faculty of Education, Ankara, Turkey.

ARTICLE HISTORY

Received: Mar. 19, 2020

Revised: Dec. 09, 2020

Accepted: Jan. 06, 2021

Keywords:

Design and skill workshops,
Teachers' attitudes,
Rankings of workshops
based on their importance.

Abstract: In this study, the aim was to both develop a valid and reliable measurement tool for determining teachers' attitudes as well as to determine their opinions towards design and skill workshops (DSW). In addition, the researchers aimed to determine how teachers rank design and skill workshops based on their importance. Since an attempt was made to describe the existing situation at just one point in time, a cross-sectional survey model was used. Teachers working at schools with DSW located in the cities of Ankara and Istanbul, Turkey were chosen as participants. Criterion sampling method, which is a purposeful sampling method was used for determining the sample of this study. A total of 123 teachers working at four primary schools in Ankara as well as 99 teachers working at three secondary schools, one primary school, and two high schools in Istanbul during February 2019 participated to this current study. As a result of interviews with the teachers and members of the General Directorate of Teacher Training in the Turkish Ministry of National Education (MoNE) the scale items were written. As a result of the principal components analysis for attitude scale, a valid and reliable measurement tool with 10 items emerged after removing overlapping items. According to rankings of the various workshops' importance, the teachers were divided into two groups, including those who thought that software and science were more important and those who thought that drama and visual arts were more important.

1. INTRODUCTION

In 2018, the Ministry of National Education (MoNE) in Turkey planned to establish "Design-Skill Workshops" (DSW) based on the agenda for their "Educational Vision of 2023" (Ministry of National Education, 2018). The purpose of opening these DSW was to provide opportunities for acquiring skills associated with the interests and abilities of students at all levels of education starting from primary school. These workshops were structured to focus on 21st century skills such as science, art, sports, and culture (MoNE, 2018). Workshops provided students with multiple and optional learning opportunities to reveal students' creativity in every field. Thus, through workshop participation, the students can create original works (Öztütüncü, 2016).

CONTACT: Seher Yalcin ✉ yalcins@ankara.edu.tr 📍 Ankara University, Faculty of Education, Ankara, Turkey.

ISSN-e: 2148-7456 /© IJATE 2021

In the second period of 2019, teachers who were planning to be involved in workshops where the pilot applications were being started, were provided training by the authors of this study in collaboration with the MoNE. During these training seminars, it was observed that the teachers were concerned about the use of skill workshops and that there were differences in determining the order of workshops based on their importance. A review of related literature showed that teachers' opinions regarding science, technology, and/or coding are frequently investigated (Bakırcı & Kutlu, 2018; Göksoy & Yılmaz, 2018; Hsu, Purzer, & Cardella, 2011). In addition, teachers' opinions regarding drama activities are investigated in Güler and Kandemir (2015). In research by Kurt and Duran (2019), where teachers are asked how they view the establishment of design and skill workshops in the MoNE Educational Vision of 2023, it is observed that teachers reflect on deficiencies in physical infrastructure and their anxieties regarding teaching in these design and skills workshops. In a study by Gündoğan and Can (2020), aimed at determining the opinions of teachers about design-skill workshops, it is found that elementary school teachers have some expectations from other teachers, parents, and school principals regarding design skill workshops. According to these teachers, workshops contribute to students in terms of their development periods, self-perception, choosing a profession, having positive attitudes towards schools, and using their spare time effectively. Factors such as having teachers who are incompetent, increased workload of teachers, parents seeing workshops as requiring extra expenses, and limited physical infrastructure are among the difficulties listed that may be experienced through this process (Gündoğan & Can, 2020).

As seen in meta-analysis studies (Allen, Witt, & Wheelless, 2006; Witt, Wheelless, & Allen, 2004) regarding teacher behaviors in education and teaching, teachers heavily affect student achievement. Teachers' attitudes are important for increasing students' self-development, academic achievement, and creativity skills (Erdoğan, 2006). In this context, the aim of this current study was to develop an attitude scale to determine the attitudes of teachers towards workshops since teachers' attitudes may affect their behaviors, and in turn, their behaviors can influence students' learning. It is known that one of the most important factors affecting individuals' behavior towards an event or situation is the attitude towards that particular event or situation (Fishbein & Ajzen, 1975). Attitudes can also make it easier for individuals to adapt to their environment. Teachers' attitudes and behaviors affect students' learning and motivation as well as their academic, social, and emotional development (Sezer, 2018). Therefore, in this study, the aim was to determine the attitudes of teachers towards the DSW opened in pilot schools. To predict teachers' attitudes, it is crucial that their attitudes be measured with a valid and reliable tool (Tavşancıl, 2010). Since there was no such tool in the literature, the need to develop a valid and reliable measurement tool aimed at understanding teachers' attitudes towards DSW was determined in this current study. In addition, to determine the effects of DSW workshops on all stakeholders, the opinions of teachers regarding the effects of these workshops on students, school principals, and parents were determined through a questionnaire. Finally, another aim was to determine how teachers ranked workshops according to their importance for students' development. The workshops targeted to be opened by the MoNE were: I) Outdoor Sports (football, basketball, tennis, etc.), II) Wood and Metal (wood carving, wire bending, wood design, etc.), III) Garden and Animal Care (agricultural production, planting, maintenance, landscaping, etc.), IV) Language and Critical Thinking, V) Drama (theater, diction, pantomime, etc.), VI) Science, Technology, Engineering and Mathematics, VII) Visual Arts (painting, sculpture, ceramic, etc.), VIII) Music, IX) Indoor Sports (karate, gymnastics, dance, etc.), X) Life Skills (simple repairs, using small appliances, etc.), and XI) Software and Design (robotics and coding, software, design, etc.).

In the literature, while there are studies in which teachers' opinions regarding some workshops are considered, there was no study found that determines teachers' attitudes towards the 11 workshops listed in this current study as well as which workshops are in general found as most

important by teachers. Therefore, in this study, the aim was to determine the attitudes of teachers about design and skill workshops, which were carried out as pilot studies in several schools in 2019 according to the MoNE (2018) Educational Vision of 2023. Thus, it was hoped that the attitudes of teachers about DSW, which will gradually spread to other schools, would be determined, and if necessary, MoNE would take measures regarding these issues. In addition, it is important to determine the teachers' priorities about the workshops to be opened. Also, it is extremely important that the students who are currently growing up in Turkey, gain 21st century skills to increase their competitive potential for the overall good of the country. Thus, in this study, another aim was to determine teachers' attitudes and opinions about the importance of these design and skills workshops functioning in pilot schools. In this context, the questions to be answered were as follows:

1. Is the attitudes scale of the teachers towards the design and skill workshops valid and reliable?
2. What are the opinions of the teachers towards the design and skill workshops?
3. What are the teachers' profiles according to the importance order of the workshops to be opened?

2. METHOD

In this current study, attempts were made to determine the attitudes of teachers regarding DSW and the order of importance of these workshops at one point in time. Thus, a cross-sectional survey model was utilized in this study (Fraenkel, Wallen, & Hyun, 2012).

2.1. Participants

Criterion sampling method, which is a purposeful sampling method, was used in the scope of this study. The participating teachers were chosen based on their working in schools selected as pilot schools where the DSW were implemented in 2019. Overall, a total of 222 teachers, with 123 teachers working at four primary schools in Ankara, Turkey as well as 99 teachers working at three secondary schools, one primary school, and two high schools in Istanbul, Turkey during February 2019 were chosen as participants. Among the 222 participating teachers, 156 were women (70.3%), 61 were men (27.5%) and 2.2% did not provide information regarding their sex. The teachers' seniority ranged from one year of teaching to 44 years, and the average years of experience was 17.73 years ($SD = 10.97$). Although teachers from a variety of disciplines including mathematics, science, and language took part in the scope of this study, it was recognized that most of the participating teachers were elementary school teachers ($f = 102, 45.9\%$). While, 6.3% ($f = 14$) of the participants were mathematics teachers, 5.9% ($f = 13$) were Turkish language teachers, 5.4% ($f = 12$) were English language teachers, 5% ($f = 11$) were Science teachers, 5% ($f = 11$) were Religion teachers, and the remaining 26.5% taught other subjects.

2.2. Data Collection Process and Tools

While preparing the measurement tool, first, what the researchers wanted to measure was clearly determined (DeVellis, 2003). Since the attitudes and opinions of the teachers were what the researchers were seeking to examine, a total of 41 items were prepared, as measures of teachers' attitudes and opinions. Thus, scale and questionnaire items were created based on interviews conducted with the teachers as well as the Teacher Training Directorate working within the Turkish Ministry of National Education. After the item writing stage was completed, in the following third stage, the format of the measurement tool was determined (DeVellis, 2003). While preparing the scale and questionnaire, 41 items were written using a five-point Likert type scale (*strongly disagree, disagree, undecided, agree, and strongly agree*). Structuring of the rating option of items as either verbal or numerical can differentiate a

respondent's attitude due to the confounding effect of the meaning attributed to numbers. For example, when each rating is clearly expressed verbally, the reliability coefficient gets higher (Uyumaz & Çokluk, 2016). Therefore, ratings are better expressed verbally. Subsequently, the items should be reviewed by experts (DeVellis, 2003). As a result, to determine the suitability of items in this study, the opinions of experts were received from one Turkish language instructor and three instructors working at a public university in the Department of Measurement and Evaluation. While 11 items of the scale consisted of teachers' attitudes towards DSW, 30 items in the questionnaire reflected teachers' opinions in terms of the effects of the workshops on students, teachers, parents, and administrators. Thus, the data were collected through the finalized questionnaire and scale items developed by the researchers.

In addition, the prepared questionnaire included the names of 11 workshops that MoNE was planning to open in piloted schools, and the participating teachers were asked to rank the workshops beginning with 1 which in their opinion was the most important for increasing the competitive potential of our country and so forth to 11, which they consider as the least important.

2.3. Data Analysis

Before the analysis, the collected data of this study were analyzed in terms of the required assumptions. Firstly, the data set were examined in terms of missing values, and it was found that no more than five people failed to respond to an item. This was a rate of approximately 2%, which is less than 5%, thus, the listwise deletion method was utilized (Tabachnick & Fidell, 2001). Next, the data set was examined in terms of the univariate outlier and multivariate outlier values, and as a result, the outlier values were deleted and a total of 201 teachers' responses remained. Also, the skewness and kurtosis coefficients were examined to meet the normal distribution assumption of the data, and it was observed that they ranged from -3 to +3. In addition, Bartlett's sphericity test result was found to be less than .05, thus, it was understood that the data came from a multivariate normal distribution (Çokluk, Şekercioğlu & Büyüköztürk, 2010). Finally, correlations between the items were examined and they were determined to not be above .90. Therefore, the data were found to provide the necessary assumptions for analysis. Also, six items in the scale (items 1, 2, 3, 4, 5 and 6) were coded in reverse.

To determine the construct validity of the attitude scale of the DSW, a principal components analysis (PCA) was conducted. In this study, the reason of using principal components analysis instead of exploratory factor analysis (EFA) was that the observed measurements did not have a quantification appropriate to the theoretical model (Hovardaoğlu, 2000). In addition, the Promax rotation method can be used for investigating the relationship between the components, since it is an economical and quick methodology (Tabachnick & Fidell, 2001). For the validity of items, the Pearson product-moment correlation coefficient was utilized to calculate the item total test correlation as well as to calculate the relationship between the sub-dimensions. According to Büyüköztürk (2013), a correlation coefficient, as absolute value, between 0.00 and 0.30 indicates low level of correlation while a value between 0.31 and 0.69 is accepted as a medium level of correlation, and a value between 0.70 and 1.00 indicates a high level of correlation. The Cronbach's alpha coefficient was calculated to test reliability of all the scale sub-dimensions, and as Alpar (2010) points out, a Cronbach's alpha internal consistency coefficient between 0.00 and 0.40 indicates the scale is unreliable, when it is between 0.41 and 0.60 the scale reliability is low, between 0.61 and 0.80 the scale is moderately reliable, and between 0.81 and 1.00 the scale is considered highly reliable. In addition, the frequency and percentages were calculated for the survey items. The analyses regarding the validity and reliability of the developed measurement tool as well as the frequencies and percentages of the

questionnaire items in this current study were conducted with the assistance of the SPSS software program.

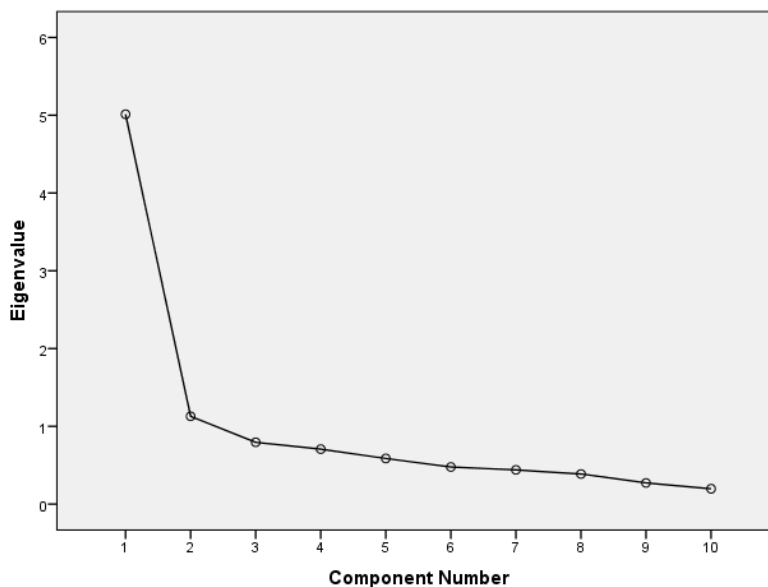
As part of the third sub-purpose of the study, data were obtained from the teachers' ranking of different workshops from the most important to the least important. Also, to reveal the profiles of teachers who chose diverse workshops as the most important, a latent class analysis (LCA) was used. According to Vermunt and Magidson (2004), LCA, which began with the use of categorical data analysis, has developed to be applied to data such as continuous and ordinal data. The purpose of LCA is to identify latent variables that explain the relationships between observed data. The in the latent class analysis, all observed variables are the cause of an unobserved latent variable. Thus, by trying different models, the model that best fits the data set is determined (Vermunt & Magidson, 2004). As a result, the BIC value is used to select the model that best fits the data (Lukočienė, Varriale, & Vermunt, 2010). The latent class analysis in this current study was conducted using the Latent Gold 5.1 package program (Vermunt & Magidson, 2013).

3. RESULT / FINDINGS

3.1. Findings Related to the Validity and Reliability of the Attitude Scale

In the scale, initially 11 items were written. The result of the Kaiser-Meyer-Olkin (KMO) test performed for the adequacy of the sample size consisting of 222 teachers before PCA was found to be .891. According to Çokluk et al. (2010), when this value is above .80, it can be interpreted as good. In addition, since Bartlett's sphericity test result ($\chi^2_{(55)} = 1004.785; p < .01$) was less than .05, it was understood that the data came from a multivariate normal distribution. As a result of the principal components analysis, the scree plot was examined, and it was observed that the items were collected in two components. The variance rate explained by these components was 59.956%. Thus, when the component loads of the items were examined, it was seen that one item was overlapping. The factor loadings of this item were more than .32 in two factors and the difference between the values of the loading was less than .10. The overlapping item was removed from the scale, and 10 items remained in the scale. As a result of the principal component analysis for the remaining 10 items, it was seen that the scale was formed under two components. The scree plot is presented in Figure 1.

Figure 1. Scree plot for the scale with 10 items.



As seen in Figure 1, 10 items were collected under two components. The load values of items, total test correlation, and Cronbach's alpha values of the items under each component are presented in Table 1.

Table 1. Loads of items under the components, item total test correlation, and Cronbach's alpha values.

Items	Loads of components		Item total test correlation
	1	2	
1. I think workshops are unnecessary.	.891	-.131	.698**
2. It makes me stressed that the workshops will be opened in my school.	.834	-.032	.711**
3. I do not want to take part in workshops.	.795	-.025	.711**
4. I do not like workshops at all.	.704	.185	.778**
5. Workshops are an extra workload for us.	.625	.053	.663**
6. Instead of taking parts in the workshops, I prefer to teach in the classroom.	.514	.276	.733**
7. I do not get bored because there is a continuous activity in the workshops.	-.039	.935	.746**
8. Opening workshops makes me happy.	-.063	.875	.697**
9. I am interested in workshops.	.072	.805	.740**
10. It makes me happy to take part in workshops.	.041	.530	.545**
The explained variance (%)	50.124	11.286	Total 61.410
Cronbach's alpha	.837	.784	.875

** $p < .01$

As seen in Table 1, the first component consisted of six items, the loads belonging to this component were between .514 and .891, and the Cronbach's alpha coefficient for this component was .837. The second component consisted of four items, component loads varied between .530 and .935, and the Cronbach's alpha coefficient was .784. The variance explained for the entire scale was 61.410% and the Cronbach's alpha internal consistency coefficient was .875. It was shown in the results that the reliability coefficients of all sub-dimensions and all of scale were acceptable and reliable (Alpar, 2010). In addition, average variance extracted (AVE) and composite reliability (CR) values were calculated within the scope of construct validity. AVE was calculated as 0.54 for the first factor and 0.64 for the second factor. Composite reliability was calculated as .87 for each factor. According to Hair, Black, Balin, and Anderson (2010), if the AVE is higher than .50 and CR is higher than .60, the validity of the construct should be considered as acceptable. In addition, as seen in Table 1, the total test scores of the items correlated with .663 to .778 for the first component and .545 to .746 for the second component. Correlation values must be .30 and above to express that the items adequately measure the desired property to be measured (Field, 2009). In this context, it can be stated that all items were particularly moderate and highly related, which was the intention of the measurement.

When the items in the first component were examined for naming the components, it was determined that there were items for teachers who had a negative attitude towards design and skill workshops. When the items in the second component were examined, it was seen that there were items for teachers who had a positive attitude towards design and skill workshops as well as wanted to work in these workshops. In order to determine the relationship among the two components in the scale, the correlation coefficients among the components were examined and it was seen that the correlation coefficient among the sub-components of the scale was .627 and

was significant ($p < 0.05$). In general, it can be said that the correlations were at a medium level. When the average of both components was examined, the average of the first component was $\bar{x} = 4.29$ (SD = 0.83), and the average of the second component was $\bar{x} = 4.17$ (SD = 0.78). When the averages were analyzed based on subcomponents, both components seemed to have high component scores. While the first component expressed teachers' negative attitudes towards the workshops, the items in the second component showed their positive attitudes towards the workshops.

3.2. Opinions of Teachers towards the Design and Skill Workshops

The frequency, percentages, mean, and standard deviation of 30-items regarding teachers' opinions on design and skill workshops are presented in Table 2.

Table 2. Opinions of teachers towards the design and skill workshops.

Items	Strongly Disagree		Disagree		Undecided		Agree		Strongly Agree		\bar{x}	ss
	f	%	f	%	f	%	f	%	f	%		
1. Workshops improve students' self-confidence.	1	.5	1	.5	2	.9	72	32.4	145	65.3	4.62	0.58
2. Workshops reveal students' different talents.	1	.5	-	-	3	1.4	67	30.2	151	68	4.65	0.56
3. Through workshops, students acquire the skills targeted to be taught in a practical way.	1	.5	1	.5	12	5.4	89	40.1	117	52.7	4.45	0.66
4. Workshops increase the collaboration among students.	1	.5	-	-	7	3.2	75	33.8	139	62.6	4.58	0.60
5. Workshops develop students' critical thinking skills.	2	.9	2	.9	13	5.9	99	44.6	106	47.7	4.37	0.72
6. Workshops develop students' skills about working in groups.	2	.9	5	2.3	6	2.7	76	34.2	126	56.8	4.48	0.75
7. Students are reluctant to participate in workshops.	72	32.4	95	42.8	32	14.4	12	5.4	8	3.6	2.04	1.01
8. Workshops increase students' commitment to school.	1	.5	8	3.6	28	12.6	106	47.7	79	35.6	4.14	0.81
9. Workshops increase students' academic success.	1	.5	4	1.8	34	15.3	111	50	68	30.6	4.11	0.76
10. Workshops improve students' self-expression skills.	2	.9	2	.9	6	2.7	110	49.5	101	45.5	4.38	0.68
11. Workshops are an essential key for students to acquire the skills required for this generation.	-	-	4	1.8	20	9	103	46.4	95	42.8	4.30	0.71
12. Workshops enable students to recognize professions.	1	.5	3	1.4	14	6.3	101	45.5	103	46.4	4.36	0.70
13. Workshops increase students' motivation towards the lesson.	2	.9	4	1.8	19	8.6	113	50.9	83	37.4	4.23	0.75
14. Workshops reinforce students' entrepreneurial behavior.	1	.5	2	.9	13	5.9	108	48.6	98	44.1	4.35	0.67
15. Workshops increase students' leadership skills.	1	.5	5	2.3	24	10.8	101	45.5	89	40.1	4.24	0.77
16. Workshops improve students' empathy building skills.	1	.5	5	2.3	30	13.5	110	49.5	75	33.8	4.14	0.77
17. Workshops improve students' communication skills.	-	-	3	1.4	6	2.7	119	53.6	93	41.9	4.37	0.61
18. In workshops, it is difficult to protect students' health and ensure their safety.	11	5	35	15.8	87	39.2	61	27.5	24	10.8	3.24	1.02

Table 2. Continued: Opinions of teachers towards the design and skill workshops.

19. The workshops provide practice of explained theoretical information.	-	-	11	5	42	18.9	118	53.2	50	22.5	3.94	0.78
20. I find it difficult to direct the students to the workshops according to their interests.	38	17.1	91	41	56	25.2	29	13.1	5	2.3	2.42	1.00
21. I do not have the skills to teach students something in workshops.	49	22.1	82	36.9	58	26.1	20	9	10	4.5	2.36	1.07
22. Workshops are not applicable to the school where I work.	71	32	96	43.2	42	18.9	9	4.1	3	1.4	1.99	0.89
23. Each school will not have enough workshops that students can attend according to their interests.	22	9.9	27	12.2	83	37.4	67	30.2	16	7.2	3.13	1.06
24. With the organization and management of workshops, the workload of school administrators will increase.	17	7.7	22	9.9	37	16.7	102	45.9	41	18.5	3.58	1.14
25. For the workshops to be implemented in all schools in Turkey is impossible.	16	7.2	41	18.5	100	45	38	17.1	25	11.3	3.07	1.05
26. The success of schools with different workshops increases.	3	1.4	4	1.8	44	19.8	106	47.7	63	28.4	4.01	0.83
27. Some parents do not want their children to be in the workshops for security reasons.	10	4.5	35	15.8	88	39.6	67	30.2	21	9.5	3.24	0.98
28. Students choose workshops not according to their interests, but according to parents' demands.	22	9.9	47	21.2	65	29.3	67	30.2	21	9.5	3.08	1.13
29. Parents want their children to benefit from all workshops.	2	.9	17	7.7	88	39.6	87	39.2	28	12.6	3.55	0.84
30. Parents want their children to go to schools with different workshops.	6	2.7	13	5.9	72	32.4	90	40.5	41	18.5	4.62	0.58

When the questionnaire items in Table 2 are examined, the first 17 items include the opinions of teachers on the effect of the workshops on students. Items 18, 19, 20, and 21 are related to the teachers' own competences regarding the workshops. While, items 22, 23, 24, 25, and 26 are related to teachers' opinions on the effects of workshops on schools. The items 27, 28, 29, and 30 refer to teachers' opinions regarding the parents' responses to the workshops. When the responses provided for the 30 questionnaire items were evaluated, it was seen that the 2nd, 1st, and 4th items had the highest average. The item that teachers thought workshops had the most impact on students was the second item (Workshops reveal students' different talents). 98.2% of the teachers responded to this item as "agree" and "strongly agree". The next item in which the teachers thought workshops had the most impact on students was the first item (Workshops improve students' self-confidence). 97.7% of the teachers responded to this item as "agree" and "strongly agree". The item in which teachers thought workshops had the least impact on students was the 7th item (Students are reluctant to participate in workshops). 75.2% of the teachers expressed positive responses by providing a "strongly disagree" and "disagree" response to this negatively stated item. Then, the item in which teachers thought workshops had the least impact on students was the 9th item (Workshops increase students' academic success). 80.6% of the teachers responded to this item as "agree" and "strongly agree". Although this rate was lower than other items, it was a high rate. Therefore, when the items in this category were analyzed in general, it can be stated that teachers thought that workshops would have positive effects on several student skills such as cooperation, communication, and leadership.

When the items related to teachers' own competencies regarding workshops (items 18 to 21 and above) are examined, the item with the highest average ($\bar{x} = 3.94$) was the 19th item (The workshops provide practices of the explained theoretical information). The item that teachers felt most inadequate about workshops was the 18th item (In workshops, it is difficult to protect students' health and ensure their safety). Only 20.8% of the teachers stated that they could protect the health and safety of the students by responding to this negatively stated item by choosing “*strongly disagree*” and “*disagree*”. When the teachers' views on the effects of the workshops on schools (between the item 22nd and 26th), the 26th item (The success of the schools with different workshops increases) had the highest average ($\bar{x} = 4.01$). Subsequently, the item with the highest average among other negatively stated items was item 24 (With the organization and management of workshops, the workload of school administrators will increase). In other words, 64.4% of teachers thought that workshops would increase their workload.

When the opinions of the teachers about the responses of the parents regarding the workshops are examined (items between 27 to 30), the item they think is the most positive was item 30 (Parents want their children to go to schools with different workshops). 59% of the teachers responded to this item as “*agree*” and “*strongly agree*”. That means, most teachers thought that the workshop would affect parents' school preferences. On the other hand, it was shown in item 28 that teachers' opinions regarding parents' responses to workshops as the least positive (Students choose workshops not according to their interests, but according to parents' demands). Only 39.7% of the teachers responded to this item as “*agree*” and “*strongly agree*”. In other words, teachers thought that students would mostly choose workshops according to their interests.

3.3. Profiles of Teachers Based on How They Rank Workshops in Terms of Their Importance

As a result of the analysis made to determine the profiles of teachers according to their importance order regarding DSW, the results of compliance from the tested models are presented in Table 3.

Table 3. Teacher profiles according to the teachers' importance order regarding DSW.

		LL	BIC(LL)	Npar	L ²	Class.Err.
Model1	1-C	-5241.0445	11075.0882	111	8251.7659	0.0000
Model2	2-C	-5160.0887	10977.2844	123	8089.8542	0.0681
Model3	3-C	-5128.2635	10977.7421	135	8026.2038	0.1005
Model4	4-C	-5092.6479	10970.6190	147	7954.9727	0.0956
Model5	5-C	-5056.1109	10961.6529	159	7881.8986	0.0933

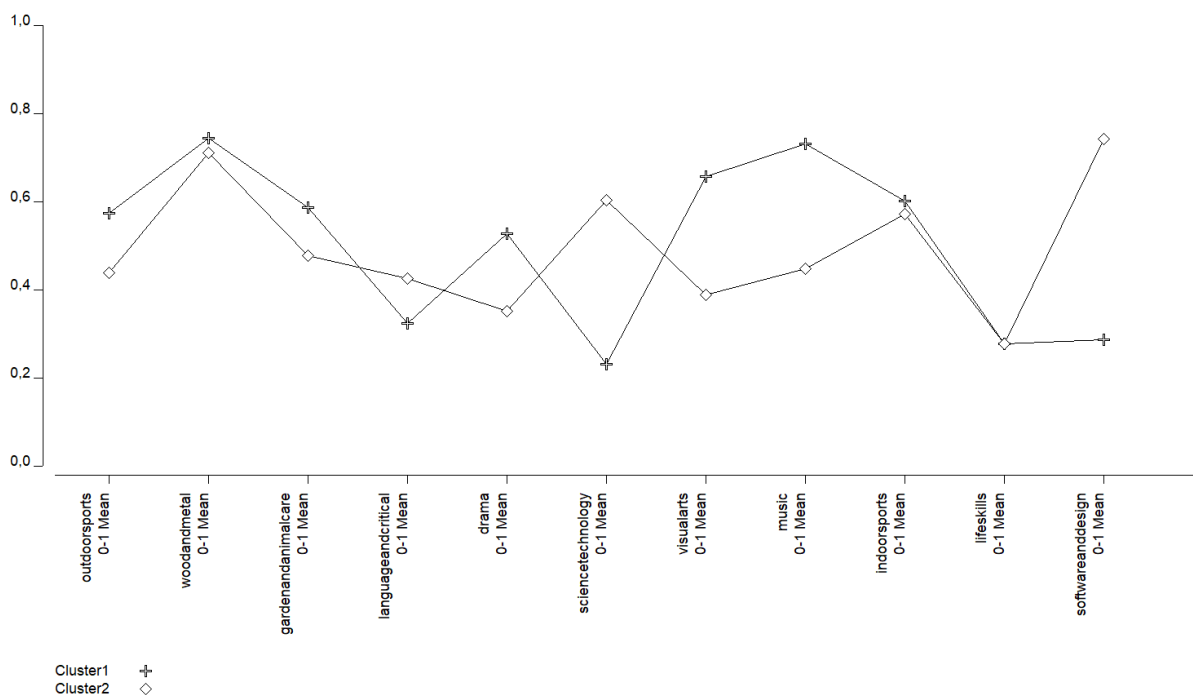
Note= C: Cluster

The statistics used to examine the compatibility of the data with the model regarding the ordering of the 11 workshops according the teachers' opinions of their importance as well as the different models attempted to determine the teacher profiles related to these rankings are presented in Table 3. As seen in Table 3, the classification error received the lowest value in a two-class model. The class with the lowest BIC value was the second model, that is, the two latent class models fit the data best. When the results of the two-class model were examined, 53.7% ($f = 115$) of teachers were in cluster-1 and 43.9% ($f = 94$) in cluster-2. 67.8% of the teachers in cluster-1 were women and 30.4% were men. The average years of their seniority were 15.54. 74.5% of the teachers in cluster-2 were women and 23.4% were men. The average years of their seniority was 19.38. According to the importance of the 11 workshops to be

opened, it was seen that teachers were divided into two groups. The status of the workshops in the two clusters based on teachers' ranking is presented in Figure 2.

As shown in Figure 2, teachers' rankings of workshops including wood and metal (wood carving, wire bending, wood design, etc.), garden and animal care (agricultural production, planting, maintenance, landscaping, etc.), indoor sports (karate, gymnastics, dance, etc.) and life skills workshops (simple repairs, using small appliances, etc.) were close to each other in terms of their importance. Among these workshops, the teachers, in turn, considered life skills, garden and animal care, indoor sports, and wood and metal workshops as the most important ones.

Figure 2. Status of the workshops based on teachers' ranking.



The workshops where the teachers were divided into two groups in terms of their importance were outdoor sports (football, basketball, tennis, etc.), language and critical thinking, drama (theater, diction, pantomime, etc.), science, technology, engineering and mathematics, visual arts (painting, sculpture, ceramics, etc.), music and software and design (robotics and coding, software, design, etc.) workshops. The importance ranking of the teachers in the first cluster of these workshops was as follows, starting from the most important: science, technology, engineering and mathematics; software and design; language and critical thinking; drama; outdoor sports; visual arts; and music. The order of importance of teachers in the second cluster was as follows, starting from the most important: drama; visual arts; language and critical thinking; music; outdoor sports; science, technology, engineering and mathematics; and software and design.

4. DISCUSSION and CONCLUSION

In this study, the aim was to determine the attitudes of teachers regarding design and skill workshops, which were carried out as pilot studies in several schools during 2019 according to the Educational Vision of 2023 as well as to develop a valid and reliable measurement tool aimed at measuring these attitudes. As a result of the analysis, a two-component scale with 10 items emerged. The two components that occurred in the scale were: i) teachers who have a negative attitude towards design and skill workshops, and ii) teachers who have a positive

attitude towards design and skill workshops and want to take part in these workshops. While the reliability coefficients for the sub-dimensions of the scale and for the entire scale were good, the variance explained for the entire scale was sufficient. As a result, a valid and reliable measurement tool consisting of 10 items with two components was developed to determine teachers' attitudes towards design and skill workshops. Six items in the scale (items 1, 2, 3, 4, 5 and 6) were coded in reverse. According to the obtained results, it was determined that the general attitude of teachers was high.

Also, in this current study, the aim was to determine the opinions of teachers regarding the effects of design and skill workshops on students, teachers, parents, and school administrators. As a result of the research, it was determined that almost all teachers who participated in the study thought that the workshops would reveal students' unique abilities as well as help to develop the students' self-confidence. In addition, it was determined that they thought workshop participation would contribute to the development of several skills such as student communication, cooperation, and leadership. In parallel to this current study, another study aimed at determining the opinions of elementary school teachers regarding design-skill workshops (Gündoğan & Can, 2020), finds that workshops positively affect students' development periods, self-perception, choice of profession, positive attitude towards school, and being active in their spare time.

As a result of this current research, it was determined that according to teachers' responses, protecting the health and safety of students and the organization and management of workshops were both difficult. In addition, the workload of school administrators would increase. Also, some teachers stated not having the skills to teach these workshops as well as they would have difficulty in placing students in workshops according to the students' interests. Furthermore, in line with the findings obtained in this current study, another study to determine the opinions of elementary school teachers regarding design and skill workshops, factors such as inadequate teachers, increased workload, parents seeing workshops as an extra expense, and limited physical infrastructure of schools are among the difficulties that can be experienced within the workshop process (Gündoğan & Can, 2020). In order to eliminate these negative opinions regarding workshops, investigating teachers', students' and parents' opinions about their needs and interest can be helpful before designing and operating workshops in educational settings (Öztürk, 2020).

In this context, it is recommended that MoNE create opportunities for teachers to improve their knowledge and experience regarding the management, number, safety, and implementation of design and skill workshops. It is also recommended that MoNE determine the attitudes and opinions of teachers in this process by applying this developed scale and questionnaire. In addition, MoNE should also conduct interviews with teachers for gaining teacher insight regarding the elimination of workshop shortcomings if any appear prior to the pilot schools expanding in number.

Finally, in this current study, the aim was to determine the teachers' profiles according to the order of importance among the workshops to be opened. As a result, teachers were divided into two groups as those who thought science, technology, and software were more important, and teachers who thought drama and visual arts were more important. It was a remarkable finding that the opinions of teachers were extremely divergent from one another. Examination of the related literature showed that there was no study comparing the importance of design and skill workshops. However, it was seen that the opinions of teachers regarding the importance of different fields had been considered in past research. For example, in studies in which teachers' opinions regarding science, technology, and software related topics were obtained (Bakırcı & Kutlu, 2018; Göksoy & Yılmaz, 2018; Hsu et al., 2011), it was recognized that teachers consider these fields important since they contribute to the development of many 21st century skills for

students. For example, in a study in which teachers' opinions about robotics and coding course were obtained through interviews, Göksoy and Yılmaz (2018) find that most teachers think robotics and coding lessons are very successful and beneficial in terms of developing students' analytical thinking skills, understanding the logic of algorithms, gaining coordination skills, and increasing their multi-faceted thinking skills. In addition, it is determined in Göksoy and Yılmaz (2018) that all of the teachers queried agreed upon the necessity of teaching robotics and coding lessons to students in all grade levels. While Hsu et al. (2011) conducted a survey investigation of 192 elementary school teachers' opinions and familiarity regarding the use of design, engineering, and technology within their classrooms. As a result, these teachers think that providing room for design, engineering, and technology lessons within their curriculum is important for students so that they can gain experiences and follow new developments in the age of science and technology. However, when it came to their familiarity with using technology in their classrooms, the teachers stated that they are not familiar and need in-service training to better prepare students for the future as well as not lose motivation within the classroom. Similarly, teachers in this current study also highlighted their concerns regarding how to manage workshops as well as having a lack of experience and knowledge regarding certain design and skill workshops they expected to be opened within their schools. Furthermore, in a study in which opinions of science teachers were obtained through interviews regarding the approach of science, technology, engineering and mathematics (STEM), Bakırcı and Kutlu (2018) state that according to teachers, STEM will develop students' research, inquiry, and creativity skills; design products suitable for solving the determined problem situation; and increase their scientific process skills. The teachers also state that the STEM approach will increase the motivation and interest of students towards the lesson, allow students to create products, and increase laboratory use in schools. In another study, Güler and Kandemir (2015), state that the use of drama method by teachers in their lessons positively affects both the academic success and social skills of students. In this context, it can be stated that the findings in this current study for teachers having two diverse ranking choices is consistent with the research in the field of education. However, since there was no study in the literature comparing the importance or the effects of workshops according to teachers' opinions, it is recommended that further studies be carried out to reveal how teachers views regarding workshops and interviews be conducted to better understand the reasoning behind teachers decisions as well as compare any diverse or similar findings. Also, it was revealed through the results of this current study that teachers were divided into two clusters. The examination of these clusters showed that teachers in cluster-2 had more years of seniority and that a higher proportion of them were women. Considering that the teachers in the second cluster were composed of teachers who believed that drama and visual arts were more important, may be due to female teachers utilizing drama more frequently in their classrooms as a result of greater confidence about using it as part of their instruction. Therefore, it can be stated that female teachers may see workshops regarding drama and visual arts as more important. For example, Güler and Kandemir (2015) find that female teachers have higher self-efficacy in using the drama method than male teachers.

Thus, in this current study, teachers' opinions and attitudes were investigated regarding the workshops that were slated to be opened in their pilot schools. Importantly, it was believed that identifying the workshops not deemed as important by teachers, could be a useful guide for MoNE in the preparation process for future school workshops. At the same time, the number of workshops which teachers considered as important could also increase. Therefore, along with the teachers, the order of importance of workshops presented in schools can also be determined according to the opinions of other stakeholders such as the students, parents, and administrators. As a result, after considering these views, important decisions can be made regarding the future establishment, opening and management of design and skill workshops within Turkish schools.

Acknowledgments

The first draft of the paper was presented at VIth International Eurasian Educational Research Congress, Ankara, Turkey.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Aysegül Bayraktar: Investigation, research design, literature review, data collection, and writing the manuscript. **Seher Yalcin:** Research design, literature review, methodology, data collection, data analysis, and writing the manuscript.

ORCID

Aysegül Bayraktar  <https://orcid.org/0000-0002-1700-8899>

Seher Yalcin  <https://orcid.org/0000-0003-0177-6727>

5. REFERENCES

- Allen, M., Witt, P. L., & Wheelless, L. R. (2006). The role of teacher immediacy as a motivational factor in student learning: Using meta-analysis to test a causal model. *Communication Education*, 55(1), 21-31.
- Alpar, R. (2010). *Uygulamalı istatistik ve geçerlik, güvenirlik [Applied statistics and validity, reliability]*. Detay Yayıncılık.
- Bakırcı, H., & Kutlu, E. (2018). Fen bilimleri öğretmenlerinin FeTeMM yaklaşımı hakkındaki görüşlerinin belirlenmesi [Determination of science teachers' views on the STEM approach]. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 9(2), 367-389. <https://doi.org/10.16949/turkbilmat.417939>.
- Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı [Handbook of data analysis for social sciences]* (18. baskı). Pegem Akademi.
- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve Lisrel uygulamaları [Multivariate statistics for social sciences: SPSS and Lisrel applications]*. Pegem Akademi.
- DeVellis, R. F. (2003). *Scale development theory and applications* (2nd ed.). SAGE Publication, Inc.
- Erdoğan, M. Y. (2006). Yaratıcılık ile öğretmen davranışları ve akademik başarı arasındaki ilişkiler [Relationships between creativity, teacher behaviours and academic success]. *Electronic Journal of Social Sciences*, 5(17), 95-106.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage Publications Ltd.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.
- Göksoy, S., & Yılmaz, İ. (2018). Bilişim teknolojileri öğretmenleri ve öğrencilerinin robotik ve kodlama dersine ilişkin görüşleri [The opinions of information relations teachers and their students with regard to lessons of robots and decoding]. *Journal of Düzce University Institute of Social Sciences*, 8(1), 178-196.
- Güler, M., & Kandemir, Ş. (2015). Öğretmenlerin drama yöntemine yönelik görüşleri ve öz yeterlik düzeyleri [Teachers' opinions on drama method and self-efficacy levels]. *Journal of Kırşehir Education Faculty*, 16(1), 111-130.

- Gündođan, A., & Can, B. (2020). Sınıf öğretmenlerinin tasarım-beceri atölyeleri hakkındaki görüşleri [Pre-service teachers' views on design-skill ateliers]. *Turkish Studies-Educational Sciences*, 15(2), 851-876. <https://dx.doi.org/10.29228/TurkishStudies.40357>
- Hair, J. F., Black, W. C., Balin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Maxwell Macmillan International Editions.
- Hovardaođlu, S. (2000). *Davranış bilimleri için istatistik* [Statistics for behavioral sciences]. Hatipođlu Yayınları.
- Hsu, M., Purzer, S., & Cardella, M. E. (2011). Elementary teachers' views about teaching design, engineering, and technology. *Journal of Pre-College Engineering Education Research (J-PEER)*, 1(2), 31-39.
- Kurt, M., & Duran, E. (2019). 2023 eğitim vizyonuna ilişkin öğretmen görüşleri [Teachers' view about the 2023 education vision]. *International Journal of New Approaches in Social Studies*, 3(1), 90-106.
- Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, 40(1), 247-283. <https://doi.org/10.1111/j.1467-9531.2010.01231.x>.
- Ministry of National Education [Milli Eğitim Bakanlığı-MEB]. (2018). *2023 Eğitim Vizyonu*. [Educational vision of 2023], Retrieved January 20, 2019, from http://2023vizyonu.meb.gov.tr/doc/2023_EGITIM_VIZYONU.pdf
- Öztürk, Z. (2020). Tasarım ve beceri atölyelerine yönelik uygulamalar - Almanya örneđi [Applications for design and skill laboratories - example of Germany]. *Milli Eğitim Dergisi* [Journal of National Education], 49(227), 141-158.
- Öztütüncü, S. (2016). Disiplinlerarası atölye dersleri üzerine bir değerlendirme [An evaluation on interdisciplinary workshop courses]. *Akdeniz Sanat* [Mediterranean Art], 9(19), 15-28.
- Sezer, Ş. (2018). Öğretmenlerin sınıf yönetimi tutumlarının öğrencilerin gelişimi üzerindeki etkileri: Fenomenolojik bir çözümleme [The effects of teachers' classroom management attitudes on students' development: A phenomenological analysis]. *Hacettepe University Journal of Education*, 33(2), 534-549. <https://doi.org/10.16986/HUJE.2017031319>.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.), Allyn and Bacon,.
- Tavşancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi* [Measuring attitudes and data analysis with SPSS]. Nobel Yayın Dağıtım.
- Uyumaz, G., & Çokluk, Ö. (2016). An investigation of item order and rating differences in likert-type scales in terms of psychometric properties and attitudes of respondents. *Journal of Theoretical Educational Science*, 9(3), 400-425.
- Vermunt, J. K. & Magidson, J. (2004). Latent class analysis. In M. S. Lewis-Beck, A. Bryman, and T. F. Liao (Eds.), *The sage encyclopedia of social sciences research methods* (pp. 549-553). Sage Publications.
- Vermunt, J. K., & Magidson, J. (2013). *Latent GOLD 5.0 upgrade manual*. Statistical Innovations.
- Witt, P. L., Wheelless, L. R., & Allen, M. (2004). A meta-analytical review of the relationship between teacher immediacy and student learning. *Communication Monographs*, 71(2), 184-207.

Homework Process in Higher Education Scale (HPHES): A Validity and Reliability Study

Veda Yar Yildirim ^{1,*}

¹Kahramanmaraş Sutcu Imam University, Faculty of Education, Department of Educational Sciences, Kahramanmaraş, Turkey

ARTICLE HISTORY

Received: May 27, 2020

Revised: Nov. 08, 2020

Accepted: Jan. 07, 2021

Keywords:

Homework,
Scale,
Higher Education,
Validity,
Reliability.

Abstract: The aim of this study was to develop a scale to measure the process of receiving and completing homework from the perspective of university students, and to conduct its validity and reliability analyses. Two different sample groups were formed in order to develop the Homework Process in Higher Education Scale (HPHES). Students studying in different faculties in four different universities in the 2019-2020 academic year were included in the sample. The sample consisted of 368 students for Exploratory Factor Analysis (EFA) and 400 students for Confirmatory Factor Analysis (CFA). In the EFA, it was determined that the scale had a five-factor structure with 28 items. This structure was evaluated using CFA. When the fit indices of the resulting model were examined, the following results were obtained: $\chi^2/df = 2.36 < 4$; CFI= 0.91; TL= 0.90; RMSEA= 0.05; SRMR = 0.05. The structure was confirmed using CFA. Cronbach's Alpha reliability coefficient results calculated for the scale were verified with composite reliability coefficients. The convergent validity was tested by calculating average variance extracted (AVE) of each factor. The results of validity and reliability study of the HPHES showed that it was a valid and reliable measurement tool with five factors and 28 items. The subject of homework in higher education can be examined in terms of different variables using the HPHES.

1. INTRODUCTION

Homework has always been a high priority in the education system. In particular, homework may be more important than it was in the past due to recently increasing chaos in the external world. Homework generally indicates a task, duty or behavior which must be carried out according to a set of rules and instructions (Turkish Language Association, 2020). The concept of homework in education can be defined as the tasks given to students by teachers to complete in their extra-curricular time (Cooper, 1989; Li, Bennett et al., 2018). In one study on higher education, the students perceived their own independent studying as homework (Murtagh, 2010). Other students have been unhappy about having too much homework and thus not being able to participate in leisure activities (Núñez et al., 2015). Although homework is one of the key and indispensable elements in learning and teaching processes, students have often complained about it (Ünal et al., 2018; Hyman et al., 2005).

CONTACT: Veda Yar Yıldırım ✉ vedayaryildirim@gmail.com 📍 Kahramanmaraş Sutcu Imam University, Faculty of Education, Department of Educational Sciences, Kahramanmaraş, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

Baran (2019) drew attention to the history of the homework and stated that homework has been part of the education system for more than a century. Both its good and bad sides have been discussed; for example, pioneering educators carried out campaigns in America in the 1940s on the basis of the idea that homework harmed children and their families. It has also been stated that giving students less homework after the 1980s caused problems in the education system, and that this was the reason for its subsequent increase thereafter. Nowadays, distance learning is becoming more and more common in higher education. However, online homework can have disappointing results (Xu et al., 2018). These results show that, considering the processes related to homework given in higher education, the subject of homework is not given enough importance.

The positive effects of homework are not just academic. Homework also has positive non-academic effects on students, including improved self-management and self-discipline, better time management, more curiosity, and more independence in problem-solving (Cooper, 1989; 2001). Studies have shown that well-prepared homework positively affects students' skills of self-regulation, their academic self-efficacy, responsibility for their own learning, high-level thinking skills, effective learning strategies, and that it develops the habit of independent study (Duru & Segmen, 2017, as cited. Ünal, Yıldırım, and Sürücü, 2018). In another study conducted with middle school students, homework was found to improve academic achievement and produce a number of benefits (Yar Yıldırım, 2018). Similarly, Murillo and Martinez-Garrido (2014) stated that doing homework increased students' academic success.

There is various advice in the literature about the nature of giving homework. This can be summarized as follows:

- 1) Students like research-type homework. This is explained by the fact that they can easily find the information online that they need to complete their homework (Çakır & Ünal, 2019).
- 2) When giving homework, the students' interests, level of development, access to materials, and how they will be supervised should all be taken into account (Arıkan & Altun, 2007).
- 3) When giving homework, it is necessary to explain the content of the homework, and it should be interesting, stimulating, well-defined, and encourage creativity (Yapıcı, 1995; Türkoğlu et al., 2007).
- 4) Homework is a cause of stress for both students and other interested parties (parents, teachers, etc.) (Baran, 2019).
- 5) Cooper and Kalish (2015) emphasized that there is a moderate relationship between homework and achievement, and stated that homework should not be privileged above other learning activities, such as playing games and learning social skills, in order to prevent homework from having any negative effects.
- 6) The continuity of education is important for students. Public or private holidays disturb the unity of learning. Any periods not spent in education can lead to forgetting past learning and a lack of motivation. When students return to school, they then have to repeat what they have already learned. To prevent this, a higher number of homework is usually given during holiday periods. However, this does not produce the desired effect (Cooper & Kalish, 2015). Therefore, giving a lot of homework during the holidays does not serve any purpose.
- 7) Homework should be marked, and these marks should be given in a way which increases the success of students (Yapıcı, 1995; Türkoğlu et al., 2007).
- 8) Homework in higher education differs from homework in primary and secondary education. Homework in higher education is not intended to complete classroom learning (Bembenutty, 2005).
- 9) The student's attitude is important for homework to serve a purpose (Reisimer, 1999).

Studies in higher education point out various issues that should be taken into consideration in terms of homework. University students are generally encouraged to do homework by educators. If, for example, they are required to spend two hours preparing for each one-hour course they attend, this could mean that they should be studying for 40 hours per week. However, the majority of students spend less than 15 hours a week on lessons and homework (Young, 2002). This can negatively affect their level of achievement. Low effort and little time for students to do homework are associated with low motivation and a low sense of responsibility (Flunger et al., 2017). In a study conducted with university students, the students stated that they had done most of their homework, and they attributed incomplete homework to external factors such as sickness, adapting to the course, and the difficulty of the homework (Li et al., 2018). One reason for not completing homework was excessive smartphone usage (Furst et al., 2018).

The most important challenge with regard to homework occurs in the evaluation process. Students may share their work with others before handling them in. This makes it difficult to evaluate their performance. The issue of academic honesty has thus been the subject of a number of studies on homework (Balbuena & Lamela, 2015).

Homework is used as a form of learning at all levels of education. The quality of homework and the curricula implemented in higher education are assured by the European Credit Transfer System (ECTS). The ECTS is a student-centered system. According to the study by Şen et al., (2016), the time spent by students on homework is recorded in the system, but it is often not taken into account. Homework is a process, not a result. The perception of homework, the acts of setting the homework and completing it, as well as its benefits and the feedback provided to students are among the components of this process. A number of different scales in the literature have focused on homework and attitudes related to doing homework. Studies have also focused on its functionality. However, the studies assessing the use of homework in higher education by analyzing students' perceptions are not common. It is thought that the scale that will emerge with this research will contribute to the development of homework processes in higher education since processes that cannot be measured can be difficult to develop.

This scale, developed in this context, measures the perceptions of students attending higher education regarding the homework process. Perception is that people organize and interpret data transported to the sense organs through stimuli (Arkonaç, 1998: 65). Homework is a process that occurs with many sense organs, as explained above. Perceptions affect attitudes. Especially recently, in pandemic processes where homework is more involved in education, it is important to measure the perceptions of homework that may affect students' attitudes with different processes. The Homework Process in Higher Education Scale (HPHES) can be used to evaluate homework and all the processes involved from the perspective of the students. In this context, the aim of the study was to conduct validity and reliability analyses of the HPHES.

2. METHOD

This section provides information about the study groups, the process of developing the scale, and the data analysis.

2.1. Sample

To develop the HPHES, convenience sampling method was used to determine the sample of the study. This method allows data collection to be conducted more easily (Balıcı, 2004). For this purpose, the sample consisted of the students studying at different faculties at Mustafa Kemal University, Nevşehir Hacı Bektaş Veli University, Tokat Gaziosmanpaşa University (TOGU) and Firat University in the 2019-2020 academic year. In the study, the data collected from two separate groups were analyzed. The sample groups consisted of 368 students for Exploratory Factor Analysis (EFA) and 400 students for Confirmatory Factor Analysis (CFA).

Table 1 shows the data from the sample groups for EFA and CFA during the process of developing the HPHES.

Table 1. Data from the sample groups for EFA (N = 368) and CFA (N= 400) of the HPHES.

Data from the sample for EFA				Data from the sample for CFA				
Variables		N	%	Variables		N	%	
Universities	Mustafa Kemal	242	65.8	Universities	Firat	258	64.5	
	Hacı Bektaş Veli	126	34.2		TOGU	142	35.5	
	Total	368	100		Total	400	100	
Year	1	48	13.0	Year	1	99	24.8	
	2	122	33.2		3	234	58.5	
	3	121	32.9		4	67	16.8	
	4	77	20.9		Total	400	100	
	Total	368	100					
Gender	Male	144	39.1	Gender	Male	142	35.5	
	Female	224	60.9		Female	258	64.5	
	Total	368	100		Total	400	100	
Faculty	Education	50	13.6	Faculty	Education	61	15.3	
	Science and Literature	61	16.6		Science and Literature	172	43.0	
	Theology	27	7.3		Theology	31	7.8	
	Economics	51	13.9		Economics	30	7.5	
	Fine Arts	26	7.1		Sports Sciences	97	24.3	
	Architecture	28	7.6		Other	9	2.3	
	Dentistry	24	6.5		Total	400	100	
	Health Sciences	26	7.1					
	Conservatory	26	7.1					
	Veterinary	24	6.5					
	Total	368	100					

As shown in Table 1, the study had two different samples. The first sample group was the group in which data were collected for EFA during the development of the HPHES. This group included 368 students studying at Mustafa Kemal University and Hacı Bektaş Veli University. The total number of the students from Mustafa Kemal University was 242 while 126 of the sample were studying at Hacı Bektaş Veli University. In the sample, 224 of these students were female and 144 were male. There were students in all years of study (first, second, third and fourth years). As seen in Table 1, the students were studying in 11 different faculties and colleges in the group in which the data were collected for EFA during the development of the HPHES. The second sample group was the group in which the data were collected for CFA during the HPHES's development. In this group, the 400 students were studying at Firat University and Tokat Gaziosmanpaşa University. The students who were studying at Firat University were 258 while 142 of them were studying at Tokat Gaziosmanpaşa University. In the same group, 258 of these students were female and 142 of them were male. There were students in the first, second and fourth years of study. As shown in Table 1, the students were studying at seven different faculties in this group in which the data were collected for CFA during the development of the HPHES.

As seen in Table 1, data were collected from 368 students for EFA and 400 students for CFA. This number was sufficient to develop a scale. According to the literature, a sample size larger than 300 is considered sufficient to obtain consistent results (Field, 2005; Tabachnick & Fidell, 2001), and it is also stated that the number of samples should be above 100 or five times higher the number of items (Ho, 2006). In this study, the number of students from whom data were collected was nine times higher than the total number of items for EFA. The number of students

from whom data were collected for CFA was approximately 13 times higher than the total number of items. When the data collected after the application were examined, 26 forms with problems such as missing information, giving two or more responses for one item, and giving the same response for each item were excluded from the evaluation.

2.2. Development of the Scale

Studies on the subject in question were reviewed during the development of the HPHES (Murillo & Martinez-Garrido, 2014; Núñez et al., 2015; Flunger et al., 2017; Gündüz, 2005; Cooper, 1989; Türkoğlu et al., 2007; Çakır & Ünal, 2019; Yapıcı, 1995; Yar Yıldırım, 2018; Edinsel, 2008). The parts of the European Credit Transfer System (ECTS) related to the process of evaluating homework in higher education were also examined. An item pool was created in line with the literature and expert opinions. The scale is a five-point Likert-type scale and consists of the following options: “*strongly agree*” (5), “*agree*” (4), “*partly agree*” (3), “*disagree*” (2), and “*strongly disagree*” (1). To develop the HPHES the content validity and face validity were tested by obtaining the opinions of one expert in the field of Curriculum and Instruction, one expert in the field of Educational Administration and Supervision, and one expert in the field of Measurement and Evaluation. To determine whether the items in the pretest form developed in line with the opinions and suggestions of the experts were understandable to the students, a pre-application session was conducted with 20 students. These applications were carried out by the researcher; the feedback of the students was also evaluated. As a result of the expert opinions, the 40-item scale was finalized.

2.3. Data Analysis

At this stage, the construct validity and reliability studies of the scale were conducted. For the construct validity of the scale, the structure of the scale was first examined using EFA and then CFA was applied to determine whether the resulting structure was confirmed. In EFA analysis, maximum likelihood estimation method and direct oblimin rotation were applied. Maximum likelihood estimation method is one of the most preferred factoring techniques. With this analysis, it is possible to see the correlation coefficients between the factors and to test whether the factor loads are significant (Çokluk et al., 2010). The oblique rotation technique direct oblimin was used because the purpose of the research was to reveal a structure consisting of interrelated factors theoretically and the relationship between the factors was expected. This rotation technique is the only oblique rotation technique in SPSS (Can, 2014).

The reliability of the scale was tested with Cronbach’s Alpha reliability coefficient and composite reliability coefficient. It is claimed that in multi-dimensional scales, the composite reliability gives a stronger reliability value than the alpha value (Şencan, 2005). For the internal validity of the items, the item-total correlations and 27% low and high groups item analysis were examined. The relationship between the factors of the scale was also examined. In addition, convergent validity of the scale was tested by calculating average variance extracted (AVE) of each factor. The SPSS 22 program was used to analyze the data, and the Mplus 7.4 program was used for CFA. For EFA and CFA, analyses were conducted with two separate data sets. Composite reliability and average variance extracted (AVE) were calculated on excel 2010.

3. FINDINGS

This section presents the findings related to the validity and reliability analysis of the HPHES.

3.1. Findings regarding the Content Validity of the Scale

3.1.1. Exploratory Factor Analysis

Before the EFA was carried out to develop the HPHES assumptions that the absence of extreme values that may affect the results, fitting to the normal distribution, and the suitability of the sample size to factoring were tested. The Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test results are given in [Table 2](#).

Table 2. Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test results.

Kaiser-Meyer-Olkin sample suitability measure	.94
Bartlett's sphericity test chi-squared value	7837.54
Degree of freedom	780
Significance level	0.00

According to the results of the KMO test, the KMO value was .94 and thus higher than .60. This finding showed that the sample size was perfectly sufficient for factor analysis (Büyüköztürk, 2019). When the Bartlett's sphericity test results given in [Table 2](#) were examined, it was determined that the test result is statistically significant. The chi-squared value was significant and at the level of .01, and the data had a multivariate normal distribution. In addition, Q-Q plots and histograms were also examined to test the normal distribution of the items. Boxplot is examined for extreme values. Skewness (-.30) and kurtosis (-.28) values between -1 and +1 is an indicator of normal distribution. The results of the analysis showed that the measurement tool was suitable for EFA (Büyüköztürk, 2019). Accordingly, data analysis was carried out with the entire data set.

In the EFA, eigenvalues, variances, and scree plots were examined to determine the number of factors related to the 40-item scale. A relationship between factors is expected in the study. In addition, it is also aimed to reveal a structure consisting of theoretically related factors. Therefore, the oblique rotation technique direct oblimin is used. This rotation technique is the only oblique rotation technique in SPSS (Can, 2014). In the first analysis, using the Direct Oblimin oblique rotation technique, seven factors with an eigenvalue greater than one and with a contribution to the variance of 60.256 were found.

However, when the items were evaluated in terms of the degree of cyclicity and factor loads, some items were cyclical (Çokluk et al., 2010) and some items had one or two other items in the factor that they depended on. It is stated in the literature that each factor should consist of at least three items for the factor to be stable (Velicer & Fava, 1998). Therefore, a total of 12 items (29, 30, 19, 31, 32, 21, 22, 20, 14, 11, 10, and 7) were excluded, and the analysis was conducted again. Eigenvalues, explained variances, factor loads, the reliability coefficient, and item-total correlations for the final form of the factor structure determined by EFA are given in [Table 3](#) below.

Table 3. *EFA results, reliability coefficient, item-total correlations for the HPHES.*

Factors	New Item No	Item no	Items	Factor Loads					Item- Total Correlation
				1. Factor	2. Factor	3. Factor	4. Factor	5. Factor	
Process of doing homework	1	37	Feedbacks on the homework should be positive.	.65					.60
	2	35	In the process of doing the homework, the teachers should allocate sufficient time for the students for the necessary feedback.	.52					.59
	3	36	The process of doing homework brings along other gains.	.52					.65
	4	38	After the homework, students should feel pleased about their achievement.	.51					.64
	5	40	Homework should be applicable to daily life after education process.	.51					.63
	6	33	Feedback should be given from time to time while the homework is being completed.	.50					.63
	7	39	The energy and work spent evaluating homework should be reflected in the results.	.50					.66
	8	34	In the process of doing the homework, students should constantly interact with their teachers.	.46					.58
Form of the homework	9	25	Homework promotes creativity in students.		.90				.64
	10	24	Homework should be interesting.		.84				.60
	11	26	Homework should be given with clear, well-defined instructions.		.66				.64
	12	27	When giving homework, its difficulty level should be appropriate for the students.		.65				.61
	13	23	When giving homework, the teacher should talk with the students.		.58				.58
	14	28	Students should be motivated about the outcomes of the homework when it is being given to them.		.53				.59
Benefits of the homework	15	9	Homework improves self-respect.			.88			.65
	16	8	Homework increases self-confidence.			.70			.60
	17	5	Homework contributes to socialization.			.40			.57
Outcomes of the homework	18	18	Homework improves the ability to use resources.				.73		.60
	19	15	Homework helps to consolidate prior learning.				.60		.61
	20	17	Homework improves the ability to access information.				.55		.67
	21	16	Homework improves the ability to study independently.				.46		.63
	22	6	Homework develops a sense of responsibility.				.39		.61

Table 3. *Continues*

Characteristics of the homework	23	4	Homework contributes to make the lessons permanent.	.62	.63		
	24	12	Homework contributes to life-long learning.	.61	.58		
	25	2	Homework supports learning.	.59	.59		
	26	13	Homework completes learning functions in teaching.	.58	.64		
	27	1	Homework increases the time spent on courses.	.50	.44		
	28	3	Homework increases the desire to study.	.48	.61		
Eigenvalue			10.76	2.64	1.49	1.18	1.09
Explained variance (Total: 61.14%)			38.24	9.45	5.33	4.21	3.90
Cronbach's alpha (Total: .94)			.86	.89	.79	.86	.82

As seen in [Table 3](#), when the item-total correlation values for the items in the scale were analyzed, there was no item below .30 in the scale. When the items of the scale were analyzed individually, it was seen that the item-total correlations ranged between .44 and .67. This result is one of the proofs that the items on the scale have high validity. When interpreting the item-total correlation, it can be said that items with .30 and the higher item-total correlation distinguish individuals well in terms of the measured feature (Büyüköztürk, 2019). Therefore, no items needed to be discarded in terms of item-total correlation values.

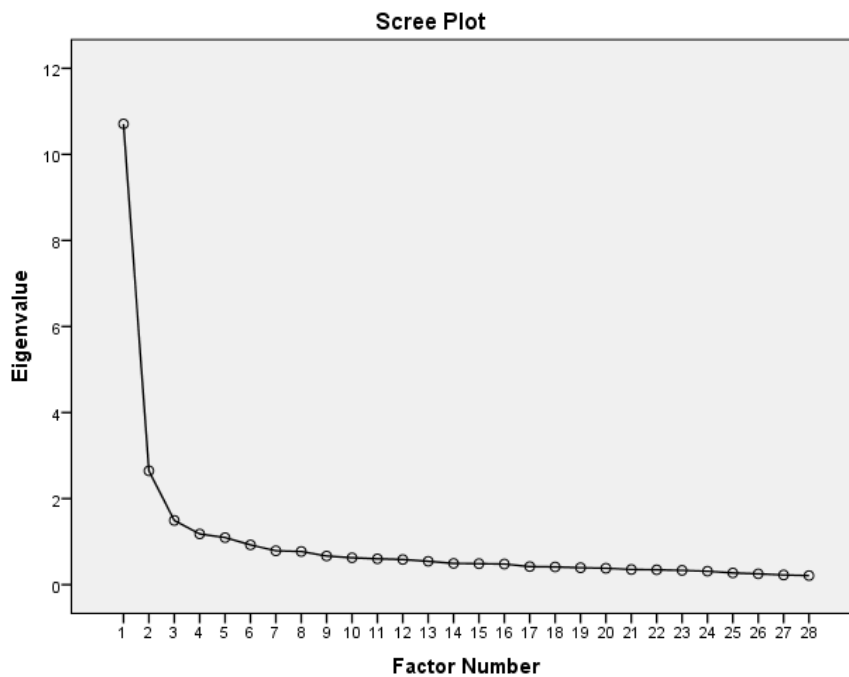
The scale in which the Direct Oblimin rotation technique was used had a five-factor structure. The contribution of the factors of the scale to the total variance was 38.24% for the first factor (process of doing homework), 9.45% for the second factor (form of the homework), 5.33% for the third factor (benefits of the homework), 4.21% for the fourth factor (outcomes of the homework), and 3.90% for the fifth factor (characteristics of the homework). The total contribution of the five factors in the scale to the variance was calculated as 61.14%.

The first factor (process of doing homework) of the HPHES consists of eight items (37, 35, 36, 38, 40, 33, 39, and 34) and the factor load values range from .46 to .65. The second factor (form of the homework) consists of six items (25, 24, 26, 27, 23, and 28) and the factor load values range from .53 to .90. The third factor (benefits of the homework) consists of three items (9, 8, and 5) and the factor load values range between .40 and .88. The fourth factor (outcomes of the homework) consists of five items (18, 15, 17, 16, and 6), and the factor load values vary between .39 and .73. The fifth factor (characteristics of the homework) consists of six items (4, 12, 2, 13, 1, and 3) and the factor load values range between .48 and .62. In scale development studies, items with factor loads of .45 and higher in the scale are accepted as a good measure (Büyüköztürk, 2019). However, it is stated that items above 0.30 can be included in the scale (Kline, 2014). In terms of factor load values, the factor loads in the HPHES were .39 and higher.

The Cronbach's alpha reliability coefficients were .86 for the first factor (process of doing homework), .89 for the second factor (form of the homework), .79 for the third factor (benefits of the homework), .86 for the fourth factor (outcomes of the homework), and .82 for the fifth factor (characteristics of the homework). When all the items in the scale were evaluated together, the Cronbach's alpha reliability coefficient was .94. These values showed that the data collected by the scale had internal consistency.

The scree plot for the HPHES, which has a five-factor structure with a total of twenty-eight items, was also examined since the number of samples was over 300 (Field, 2005). [Figure 1](#) shows the scree plot of the HPHES.

Figure 1. Scree Plot.



In Figure 1 (the scree plot), the slope reaches a plateau after the fifth point. There are five factors with eigenvalues above 1 and the scree plot supports this finding.

Table 4. Arithmetic mean, standard deviation values and correlation coefficients for factors (n=368)

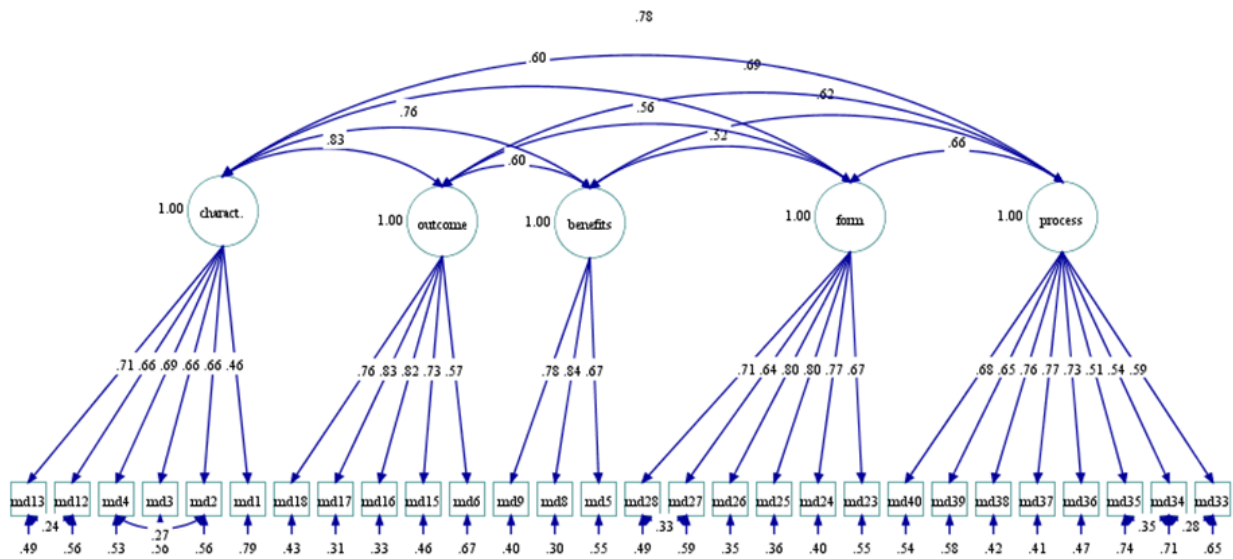
Factor	\bar{X}	<i>df</i>	Total	Process	Form	Benefits	Outcomes	Characteristics
Total	3.29	.73	1	.84**	.73**	.79**	.81**	.81**
Process	3.25	.79		1	.65**	.57**	.57**	.60**
Form	3.24	.99			1	.40**	.44**	.41**
Benefits	3.01	1.04				1	.56**	.59**
Outcomes	3.69	.87					1	.66**
Characteristics	3.27	.86						1

As seen in Table 4, the correlation values between the score for the whole scale and the five factors, and between the factors, were high and there was a significant relationship between these values at a level of .01. Correlation coefficients varied between .40 and .84. These results demonstrate that all of the factors and the scale measured a similar structure.

3.1.2. Confirmatory Factor Analysis

The scale was tested with CFA in order to verify the 28 item and five-factor structure. The diagram obtained as a result of CFA is given in Figure 2 below. As a result of the model obtained, the compatibility index of the scale was examined. According to the findings, the model can be accepted because the RMSEA and SRMR values are lower than 0.08 while the CFI and TLI values are higher than 90 (Kline, 2015).

Figure 2. CFA Diagram for the HPHEs.



As a result of the CFA of the scale, the model can be accepted because the RMSEA and SRMR were lower than 0.08 and the CFI and TLI were higher than 90 ($\chi^2/df= 2.38 <4$; CFI=0.92; TLI=0.91; RMSEA=0.05; SRMR=0.05). Figure 2 shows the factor loads of each item. Since there was a high correlation between some items related to the same factor in the model, the error measurements of the items were linked. As a result of the model, it was observed that the factor loads of each item were significant.

To determine the item discrimination of the items in the scale, the mean scores of the items were determined and item analysis was performed on the low 27% group and high 27% group. The difference between the mean group scores was analyzed using the independent groups t-test. The analysis is given in Table 5.

Table 5. Item analysis results for low 27% and high %27 groups' means.

Item No	t (Low 27%-high %27)	Item No	t (Low 27%-high %27)	Item No	t (Low 27%-high %27)	Item No	t (Low 27%-high %27)
1	8.15*	9	14.40*	23	11.02*	34	11.91*
2	13.41*	12	14.67*	24	14.44*	35	10.11*
3	14.14*	13	14.54*	25	13.91*	36	13.51*
4	11.85*	15	13.10*	26	14.02*	37	13.72*
5	14.09*	16	14.48*	27	11.97*	38	16.12*
6	11.12*	17	13.20*	28	15.32*	39	12.65*
8	13.92*	18	12.01*	33	11.63*	40	14.20*

(¹n= 400 ²n₁=n₂=108 *p< .001)

As seen in Table 5, there is a statistically significant difference between the upper and lower groups of 27% for all items in the scale, and it is seen that t-values are significant ($p <.001$). These results show that scale items have high item discrimination, high validity and are items to measure the same behavior.

Another operation after verifying the structure of the scale with CFA; in addition to Cronbach's Alpha internal consistency coefficient, the reliability of the scale is tested with a composite reliability coefficient. The composite reliability coefficients were .90 for the first factor (process of doing homework), .95 for the second factor (form of the homework), .92 for the third factor

(benefits of the homework), .90 for the fourth factor (outcomes of the homework), and .86 for the fifth factor (characteristics of the homework). When all the items in the scale were evaluated together, the composite reliability coefficient was .94. Composite reliability is calculated by factor loads and error rates obtained as a result of confirmatory factor analysis. It is suggested that compound reliability should be .70 and above (Hair, Anderson, Tatham, & Black, 1998). In addition, the AVE value calculated for each factor of the scale is over .05.

The average variance extracted (AVE) value was .52 for the first factor (process of doing homework), .77 for the second factor (form of the homework), .80 for the third factor (benefits of the homework), .64 for the fourth factor (outcomes of the homework), and .51 for the fifth factor (characteristics of the homework). An AVE value at least 0.5 indicates sufficient convergent validity (Henseler, Rinle, & Sinkovics, 2009). Convergent validity is important in terms of showing that a certain structure has emerged (Şencan, 2020).

4. DISCUSSION, CONCLUSION and RECOMMENDATIONS

Homework is a task given to students to complete in their extra-curricular time (Cooper, 1989; Li et al., 2018) which increase their self-management, self-discipline, time management and independent problem-solving skills, and curiosity (Cooper, 1989; Li et al, 2018). Doing homework is considered important in higher education due to its effect on the educational process. This scale, which was specifically developed for university students, will contribute to the literature on homework in higher education.

The HPHES has a five-factor structure with twenty-eight items. The scale's factors are "process of doing homework", "form of the homework", "benefits of the homework", "outcomes of the homework", and "characteristics of the homework". The first factor (the process of doing homework) consists of eight items, the second factor (the form of the homework) consists of six items, the third factor (benefits of the homework) consists of three items, the fourth factor (outcomes of the homework) consists of five items, and the fifth factor (characteristics of the homework) consists of six items. The total contribution of the factors of HPHES to variance is 61.14%.

When all the factors in the scale were evaluated together, the Cronbach's alpha reliability coefficient calculated was found to be .94. Accordingly, the data collected with the scale has internal consistency. It was concluded that the correlation values between the score for the whole HPHES and the five factors, and between the factors, were high and that there was a significant relationship between these values at the level of .01. The correlation coefficients varied between .40 and .84. These results indicate that all of the factors and the scale measure a similar structure.

The model can be accepted as the RMSEA and SRMR are lower than 0.08 and the CFI and TLI values are greater than 90 ($\chi^2/df= 2.36<4$; CFI=0.91; TLI=0.90; RMSEA=0.05; SRMR=0.05) according to the CFA which was conducted to confirm the five-factor, 28-item structure of the HPHES as a result of EFA.

A statistically significant difference was found between the groups in the 27% low and high analysis for the scale items and the *t* value was significant ($p<.001$). The item-total correlations of the items on the scale ranged from .44 to .67. These results showed that the scale items have high item discrimination and high validity, and that they measure the same behavior.

Cronbach's Alpha reliability coefficient results calculated for the scale were verified with composite reliability coefficients. Composite reliability coefficient calculated for the whole scale was found to be .94. In addition, the AVE value calculated for each factor of the scale is over .05.

These analyses were carried out to demonstrate the validity and reliability of the HPHES. Its structure was determined to be that of a scale with 28 items and five factors. The findings showed that the scale can provide valid and reliable results. The Turkish version of scale is given in [Table A1](#) in the appendix part.

In the chaotic atmosphere caused by the recent coronavirus pandemic, the homework given at universities has gained importance. Distance education includes both homework and exams. The HPHES developed within the scope of this study will contribute to providing feedback on how homework is perceived by students. This feedback could also be used to improve the application. The subject of homework in higher education can also be examined using the HPHES in terms of different variables.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Veda Yar Yıldırım: All the research process (Investigation, Resources, Data collecting, Visualization, Software, Formal Analysis, and Writing the original draft, Methodology, Supervision, and Validation).

ORCID

Veda Yar Yıldırım  <https://orcid.org/0000-0002-2129-4189>

5. REFERENCES

- Arıkan, Y. D., & Altun, E. (2007). Sınıf ve okul öncesi öğretmen adaylarının çevrimiçi ödev sitelerini kullanımına yönelik bir araştırma [A research on preschool and primary studentteachers' use of online homework sites]. *Elementary Education Online*, 6(3), 366-376.
- Arkonaç, S. A. (1998). *Psikoloji (zihin süreçleri bilimi)* [Psychology (science of mind processes)]. Alfa.
- Balbuena, S. E., & Lamela, R. A. (2015). Prevalence, motives, and views of academic dishonesty in higher education. *Online Submission*, 3(2), 69-75.
- Balcı, A. (2004). *Sosyal bilimlerde araştırma* (4. Baskı) [Research in social sciences]. Pegem.
- Baran, A. (2019). Home improvement: Look at the historical role of homework in education, where we are today, and what schools need to consider as they evaluate their approach. *Independent School*, 78(2), 44.
- Bembenutty, H. (2005). *Predicting homework completion and academic achievement: The role of motivational beliefs and self-regulatory processes* [Unpublished doctoral dissertation]. City University of New York.
- Beumann, S., & Wegner, S.-A. (2018). An outlook on self-assessment of homework assignments in higher mathematics education. *International Journal of STEM Education*, 5, 55. <https://doi.org/10.1186/s40594-018-0146-z>
- Büyükköztürk, Ş. (2019). *Sosyal bilimler için veri analizi el kitabı* (25. Baskı) [Manual of data analysis for social sciences]. Pegem.
- Çakır, E., & Ünal, A. (2019). An investigation into middle school students' conceptions of homework. *Language Teaching and Educational Research (LATER)*, 2(1), 41-56. <https://doi.org/10.35207/later.535622>
- Can, A. (2014). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* [Quantitative data analysis in the scientific research process with SPSS]. Pegem.

- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik [Multivariate statistics for social sciences]*. Pegem.
- Cooper, H. (1989). *Homework*. Longman.
- Cooper, H. (2001). Homework for all-in moderation. *Educational Leadership*, 58(7), 34.
- Cooper, H., & Kalish, N. (2015). Should schools give summer homework? *New York Times Upfront*, 147(13), 22.
- Edinsel, K. (2008). *Bologna Süreci'nin Türkiye'de Uygulanması "Bologna Uzmanları Ulusal Takımı Projesi" 2007-2008 Sonuç Raporu [Implementation of the Bologna Process in Turkey "Bologna Experts National Team Project" 2007-2008 Final Report]*. <http://w3.gazi.edu.tr/~gyavuzcan/documents/1.pdf>
- Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage.
- Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., & Schnyder, I. (2017). A person-centered approach to homework behavior: Students' characteristics predict their homework learning type. *Contemporary Educational Psychology*, 48, 1-15.
- Furst, R. T., Evans, D. N., & Roderick, N. M. (2018). Frequency of college student smartphone use: impact on classroom homework assignments. *Journal of Technology in Behavioral Science*, 3(2), 49–57. <https://doi.org/10.1007/s41347-017-0034-2>.
- Gündüz, Ş. (2005). Geleneksel çevrimiçi ve bireysel işbirliğine dayalı ödev uygulamalarının lisans öğrencilerinin akademik başarılarına ve ödevle ilişkin tutumlarına etkisi [The effects of traditional online and individual cooperative homework on undergraduate students' academic achievement and attitude toward homework] [Unpublished doctoral dissertation]. Anadolu University.
- Hair, J. F., Anderson, E. R., Tatham, L. R., & Black, W. C. (1998). *Multivariate data analysis*. Prentice Hall.
- Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In Rudolf R. Sinkovics & Pervez N. Ghauri (Eds.), *New Challenges to International Marketing* (C. 20, ss. 277–319). Emerald Group Publishing Limited. [https://doi.org/10.1108/S1474-7979\(2009\)0000020014](https://doi.org/10.1108/S1474-7979(2009)0000020014)
- Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. CRC Press.
- Hyman, L.M., Superville, C.R., Ramsey, V.J., & Williams, J.H. (2005). Using control charting to evaluate and reinforce student learning in accounting. *International Journal of Management*, 22(1), 41-48.
- Kline, P. (2014). *An easy guide to factor analysis*. Routledge.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
- Li, W., Bennett, R. M., Olsen, T., & McCord, R. (2018). Engage engineering students in homework: attribution of low completion and suggestions for interventions. *American Journal of Engineering Education*, 9(1), 23–38.
- Murillo, F. J., & Martinez-Garrido, C. (2014). Homework and primary-school students' academic achievement in Latin America. *International Review of Education*, 60(5), 661-681. <https://doi.org/10.1007/s11159-014-9440-2>
- Murtagh, L. (2010). They give us homework! Transition to higher education: The case of initial teacher training. *Journal of Further and Higher Education*, 34(3), 405–418. <https://doi.org/10.1080/0309877X.2010.484057>
- Núñez, J. C., Suárez, N., Rosário, P., Vallejo, G., Cerezo, R., & Valle, A. (2015). Teachers' feedback on homework, homework-related behaviors, and academic achievement. *The Journal of Educational Research*, 108(3), 204-216.

- Reisimer, E. L. (1999). *The relationship between parental attitudes on homework and homework return rates in kindergarten* [Unpublished master thesis]. University of Wisconsin – Stout the Graduate College.
- Şen, Z., Uludağ, G., Kavak, Y., & Seferoğlu, S. S. (2016). Bologna süreciyle ilgili bir inceleme: Öğrenci başarısını değerlendirme yöntemleri ile öğrenci iş yükünün karşılaştırılması [An investigation regarding the Bologna process: Comparison of student assessment methods and student workload]. *Journal of Higher Education*, 6(2), 84–94.
- Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik* [Reliability and validity in social and behavioral measurements]. Seçkin.
- Şencan, H. (2020, 25 Mayıs). Veri analizi [Data analysis]. https://ders.es/tez/gecerlilik_analizleri.html
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics (4th ed.)*. Allyn and Bacon.
- Turkish Language Association, (2020). *Turkish language association dictionaries*. Retrieved February 20, 2020, from <https://sozluk.gov.tr/>
- Türkoğlu, A., İflazoğlu, A., & Karakuş, M. (2007). İlköğretimde ödev [Homework in primary education]. Morpa Kültür.
- Ünal, A., Yıldırım, A., & Sürücü, A. (2018). Eğitim Fakültesi Öğrencilerinin etkili olarak kabul ettikleri ödevler [Homework accepted as effective by the students of the faculty of education]. *Mehmet Akif Ersoy University Journal of Education Faculty*, 48, 555-574.
- Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, 3(2), 231-251. <https://doi.org/10.1037/1082-989X.3.2.231>
- Xu, J., Fan, X., & Du, J. (2018). A study of the validity and reliability of the online homework emotion regulation scale. *Measurement*, 115, 243-248. <https://doi.org/10.1016/j.measurement.2017.10.050>
- Yapıcı, N. (1995). İlkokullarda öğretmen-öğrenci ve velilerinin ev ödevi konusundaki görüşlerinin belirlenmesi [Determining the opinions of teachers, students, and parents about homework in primary schools] [Unpublished master's thesis]. Ankara University.
- Yar Yıldırım, V. (2018). Öğretmen, öğrenci ve velilerin ortaokul düzeyinde verilen günlük ödevler hakkındaki görüşleri [The opinions of the students, teachers, and parents about the daily assignments given at secondary school level]. *Milli Eğitim Dergisi*, 220, 201-224.
- Young, J. R. (2002). Homework? What homework. *The Chronicle of Higher Education*, 49(15), A35–A37.

6. APPENDIX

Table A1. Turkish version of the scale.

Yükseköğretimde Ödev Süreci Ölçeği (YÖSÖ)							
Son madde no	İlk madde no	MADDELER	Kesinlikle katılmıyorum	Katılmıyorum	Orta düzeyde katılıyorum	Katılıyorum	Kesinlikle katılıyorum
			1	2	3	4	5
1	37	Yapılan ödevlere ilişkin dönütler yapıcıdır.	1	2	3	4	5
2	35	Ödevin yapılma sürecinde gerekli dönütler için hocalar öğrencilere yeterli zamanı ayırmaktadırlar.	1	2	3	4	5
3	36	Ödev yapma süreci başka kazanımları da beraberinde getirmektedir.	1	2	3	4	5
4	38	Ödev süreci sonunda öğrencide başarı hazzı oluşmaktadır.	1	2	3	4	5
5	40	Ödevler öğretim süreci sonunda tüm yaşamda kullanılmaktadır.	1	2	3	4	5
6	33	Ödevin yapılma sürecinde zaman zaman dönütler verilmektedir.	1	2	3	4	5
7	39	Ödevlerin değerlendirilmesinde gösterilen emek, çaba, sonuca yansımaktadır	1	2	3	4	5
8	34	Ödevin yapılma sürecinde öğrenciler hocalarla sürekli etkileşim halindedirler.	1	2	3	4	5
9	25	Ödevler verilirken öğrencide yaratıcılığa teşvik edici nitelikte olması dikkate alınmaktadır.	1	2	3	4	5
10	24	Ödevler ilgi çekici nitelikte verilmektedir.	1	2	3	4	5
11	26	Ödevler açık, iyi tanımlanmış bir yönergeyle verilmektedir.	1	2	3	4	5
12	27	Ödevler verilirken öğrencinin yapabileceği zorlukta olması dikkate alınmaktadır.	1	2	3	4	5
13	23	Ödev verme sürecinde öğrenciyle istişare edilmektedir.	1	2	3	4	5
14	28	Ödevler verme sürecinde ödevin kazanımları konusunda öğrenciler motive edilmektedir.	1	2	3	4	5
15	9	Ödev, özsaygıyı artırmaktadır.	1	2	3	4	5
16	8	Ödev, özgüveni artırmaktadır.	1	2	3	4	5
17	5	Ödev, sosyalleşmeye katkıda bulunmaktadır.	1	2	3	4	5
18	18	Ödev, mevcut kaynakları kullanma becerisini geliştirmektedir.	1	2	3	4	5
19	15	Ödev öğrenilenleri pekiştirmektedir.	1	2	3	4	5
20	17	Ödev, bilgiye ulaşma becerisini geliştirmektedir.	1	2	3	4	5
21	16	Ödev, bağımsız çalışma becerisini geliştirmektedir.	1	2	3	4	5
22	6	Ödev, sorumluluk duygusu kazandırmaktadır.	1	2	3	4	5
23	4	Ödev, derste yapılanları kalıcı hale getirmekte katkıda bulunmaktadır.	1	2	3	4	5
24	12	Ödev, yaşam boyu öğrenmeye katkıda bulunmaktadır.	1	2	3	4	5
25	2	Ödev, öğrenmeyi desteklemektedir.	1	2	3	4	5
26	13	Ödev, öğretimde öğrenme fonksiyonlarını tamamlama özelliği bulunmaktadır.	1	2	3	4	5
27	1	Ödev, ders için ayrılan zamanı çoğaltmaktadır.	1	2	3	4	5
28	3	Ödev, çalışma isteğini artırmaktadır.	1	2	3	4	5

Kirkpatrick Model: Its Limitations as Used in Higher Education Evaluation

Michael B. Cahapay ^{1,*}

¹College of Education, Mindanao State University, General Santos City, Philippines

ARTICLE HISTORY

Received: May 28, 2020

Revised: Oct. 29, 2020

Accepted: Jan. 02, 2021

Keywords:

Kirkpatrick model,
Program evaluation,
Higher education,
Limitation.

Abstract: One of the widely known evaluation models adapted to education is the Kirkpatrick model. However, this model has limitations when used by evaluators especially in the complex environment of higher education. Addressing the scarcity of a collective effort on discussing these limitations, this review paper aims to present a descriptive analysis of the limitations of the Kirkpatrick evaluation model in the higher education field. Three themes of limitations were found out: propensity towards the use of the lower levels of the model; rigidity which leaves out other essential aspects of the evaluand; and paucity of evidence on the causal chains among the levels. It is suggested that, when employing the Kirkpatrick model in higher education, evaluators should address these limitations by considering more appropriate methods, integrating contextual inputs in the evaluation framework, and establishing causal relationships among the levels. These suggestions to address the limitations of the model are discussed at the end of the study.

1. INTRODUCTION

Evaluation is an essential phase of curriculum and program development in education. Morrison (2003) noted that there are growing pressures to evaluate curriculums and programs in education for different purposes but typically to look into the achievement of the goals. As a result, it can be observed that education borrows evaluation models from other fields like business to evaluate the extent of the achievement of its educational goals. However, the appropriateness of evaluation models is contextually dependent (McNamara, 2000) and the evaluators are faced with the task to adjust them (Fitzpatrick et al., 2004). This is the point where the use of certain evaluation model, not the model itself, presents serious limitations.

Within higher education, one of these models transported to the program evaluation is the model proposed by Donald Kirkpatrick in his seminal articles published in 1959. Historically, the purpose of the Kirkpatrick model was to assist managers for a systematic and efficient means to account for outcomes among employees and in organizational systems. Managers who need solid evidence that training would improve their sales quantity, cost effectiveness, and other business indicators quickly adapted the said model (Yardley & Dornan, 2012).

CONTACT: Michael B. Cahapay ✉ mbcahapay@up.edu.ph 📍 College of Education, Mindanao State University, General Santos City, Philippines.

ISSN-e: 2148-7456 /© IJATE 2021

The Kirkpatrick model originally comprises of four levels - *reaction, learning, behaviour, and impact*. These levels were intentionally designed to appraise the apprenticeship and workplace training (Kirkpatrick, 1976). It is recommended that all programs be evaluated in the progressive levels as resources will allow. Each of these levels have different emphases and are described based on Kirkpatrick & Kirkpatrick (2006):

- The reaction level determines the level of satisfaction of the participants or how they feel about the training program. Assessing how engaged the participant were, how they contributed, and how they responded assists evaluators to recognize how well the participants perceive the training program.
- The learning level measures the level to of knowledge, skills, and values acquired by the participants from the program. This level measures what the participants think they will be able to perform the expected change, how assured they are that they can perform it, and how driven they are to perform it.
- The behaviour level ascertains the changes in the behaviours of the participants in the work environment as a result of the program. The measurement of this level is an activity that should occur over weeks or months following the inputs that the participants received from the training program.
- The impact level examines the institutional outcomes that demonstrate a good return on investment and can be attributed to the training program. Considering the institutional outcomes, a task that can be challenging is to design a method to evaluate these outcomes which are long term in nature.

The general strengths of the Kirkpatrick model in evaluation theory and practice have been extolled by scholars. They recognize the model for its ability to provide the following: simple system or language in dealing with the different outcomes and how information about these outcomes can be obtained; descriptive or evaluative information about the kind of training that are needed, thus allows organizations to anchor the results of what they do in business points of view; and practical approach for the typically complex evaluation process (Bates, 2004). With these strengths, it cannot be denied that Kirkpatrick model has offered significant contributions to the evaluation theory and practice.

Because of the strengths, the Kirkpatrick model has become known in a wide range of evaluation studies. The application of the model has reached the different higher education fields and aspects (see Quintas et al., 2017 on instructional approach; Baskin, 2001 on online group work; Paull et al., 2016 on curriculum intervention; Abdulghani et al., 2014 on research workshops; Aryadoust, 2017 on writing course; Chang & Chen, 2014 on online information literacy course; Farjad, 2012 on training courses; Rouse, 2011 on health information management courses; Dewi & Kartowagiran, 2018 on internship program; Liao & Hsu, 2019 on medical education program; Miller, 2018 on leadership development program; Sahin, 2006 on teacher training program; Masood & Usmani, 2015 on training program; Embi et al., 2017 on blended learning environment).

The reviews of Alliger and Janak (1989), Bates (2004), and Reio et al., (2017) help understand the current state of the Kirkpatrick model by overtly tackling its inherent limitations in the general context. However, an analysis of the limitations when the model is transported to higher education evaluation has not been paid attention. Lambert (2011) supports that judging the worth of learning in the multifarious environments of higher education can be without experiences of limitations. As regards these limitations in the context of higher education, there has been a passing mention (Steele et al., 2016; Covington, 2012; Haupt & Blignaut, 2007) and a collective analysis is yet to be explored.

The intention of this paper is not to downplay the Kirkpatrick evaluation model. It intends to inform evaluators of the possible limitations in the adaptation of such a model in the evaluation in higher education programs or institution. This paper also disclaims that such limitations are not directly attributed to the model. These limitations are based on how the model is applied by evaluators in the educational field. If these limitations are given attention, evaluators will be in a better position as to making cogent considerations to proactively address the potential disadvantages of using the model. As such, they will be guided in designing appropriate methods and tools to successfully use the model and accomplish their desired goals.

Considering the issues and gaps raised in this paper, the current review presents a descriptive analysis of the limitations of the Kirkpatrick model as used in the higher education evaluation.

2. METHOD

This section presents the methods used in this study. It discusses the research design, data sources, data analysis, and analysis procedure. They are elaborated as follows.

2.1. Research design

This research is primarily conducted as a desk review. This research design involves the process of gathering relevant data from various sources (Sileyew, 2019). It may include materials such as legal codes, historical records, statistical data, published papers, news articles, review articles, and other pieces that have a descriptive or analytical purpose (Guptill, 2016). This research design is considered appropriate for this paper. It provides an cogent approach to search, collect, and analyze different materials related to the focus of this paper.

2.2. Data sources

The sources of data for this paper are considered as primary sources. They are original documents, data, or images (Guptill, 2016). These primary sources in the current study consist of books, essays, and articles accessed online. Furthermore, they were screened and included based on the following eligibilities: written in intelligible language, accessible in full text, authored by credible persons or institutions, focused on the Kirkpatrick model as used in higher education evaluation.

2.3. Data analysis

The primary sources gathered in this study were treated through document analysis. It is a technique that “requires repeated review, examination, and interpretation of the data in order to gain empirical knowledge of the construct being studied” (Frey, 2018). Moreover, it involves the creation of themes similar to how interview data are treated (Bowen, 2009). It should be noted, however, that since the themes were readily identified according to the interest of this research, the analysis process was deductively performed (Braun & Clarke, 2006).

2.4. Analysis procedure

The process of deductive analysis was carried out in this study in stages. The researcher initially acquainted himself with the data in the materials, noting down codes relevant to the limitations of the Kirkpatrick model. Then, he grouped these codes based further on the earlier identified themes of limitations of the Kirkpatrick model. The researcher repetitively reviewed the the codes and themes, returning to the original sources until final results were generated.

3. LIMITATIONS OF KIRKPATRICK MODEL AS USED IN HIGHER EDUCATION EVALUATION

This paper is mainly driven by the purpose to provide a descriptive analysis of the limitations of the Kirkpatrick model as it is used by evaluators in the higher education. The following limitations are presented and discussed.

3.1. Propensity towards lower levels of the model

Alliger and Janak (1989) reviewed articles evaluating the Kirkpatrick model. They stated a major conjecture that the levels are structured in increasing order of importance and the model is tiered. Because of this notion, they observed that in the business world, professionals tend to disregard the lower levels of the Kirkpatrick model and address only the higher ones. This is not the case when it comes to higher education.

When the Kirkpatrick model is adapted in educational evaluation, there are pieces of evidence of the tendency to restrict evaluation to the lower levels of the model (Steele, et al., 2016). When it comes to evaluation of effectiveness whether of a training program for teachers or a curriculum for the students in higher education, this limitation can be observed (e.g. see Quintas et al., 2017; Dewi & Kartowagiran, 2018; Sahin, 2006; Aryadoust, 2017). It should be noted that, as disclaimed earlier, these limitations are not caused by the model itself but how it is used in the educational field

Efforts to use the third and fourth levels of the model have been exerted when it comes to evaluation of training effectiveness for teachers in higher education. However, there seemed to be concerns as regards the scope and rigor. For example, in the study conducted by Abdulghani et al. (2014), they evaluated the effectiveness of research workshops to the faculty at a college of medicine. The researchers, however, evaluated the behavioural changes and main outcomes as a single unit in terms of the research activities of the participants. This situation asserts again the limitations as not directly caused by the model itself but how it is used in the field.

Massod and Usmani (2015) also evaluated the outcomes of a training program for teachers in selected medical institutions. The evaluation was framed within the four levels of the Kirkpatrick model. However, the results only discussed the benefits gained by the participants based on their perceptions. These perceived benefits were taken at different points of time to show impacts across the four levels. Moreover, Farjad (2012) attempted a comprehensive evaluation using the Kirkpatrick model in determining the effectiveness of training courses for university employees. The four levels, however, were just measured based on the perceptions of the employees using the survey. The use of perceptions of the participants themselves can be subjective and may decrease the reliability of the results.

The survey of higher education evaluation studies using the Kirkpatrick model in determining the effectiveness of training courses to the employees shows varied evaluation practices. Some studies were restricted on levels one and two. Other studies have tried to reach levels three and four, but they appeared to downplay the scope or diminish the rigor. It should be noted that levels three and four evaluate the workplace behaviours and the organisational impacts respectively (Kirkpatrick & Kirkpatrick, 2006). Rouse (2011) explained that level four operates at the system level or organisational impact. It attempts to identify if an increase in company revenues, client approval, or related indicators is realised as a result of the course or program inputs. Covington (2012) added that while the return of investment is an option to assess economic outcomes, in some professions such as education, optimal outcomes are not exclusively measured by monetary means.

On the other hand, Nickols (2000) explained the propensity towards the lower levels of the Kirkpatrick model in the context of evaluating the impact of the curriculums or programs on the students. He elaborated that any evaluation of change in the student behaviours, level three in the model of Kirkpatrick, will have to occur when they are already in the workplace. It is deemed logical, therefore, to assess behaviour changes in the workplace. However, in the higher education context, employing the Kirkpatrick model can be challenging because students have not normally gone for employment at this stage in their lives.

Hence, because it is difficult to follow the students in the field, many educators tend to end with just the lower levels of the model, leaving out the long-term results of the education. Even if the expectation is clearly defined, it would not be practically easy to trace the learners in the field. Sahin (2006) expressed in a study that an essential limitation when the model is used by evaluators in education is related to the evaluation evidence collected for the behaviour and impact levels. The performance of the students was not directly assessed through observation. Some indirect processes were instead employed to gauge the outcomes of the stated levels.

There are also studies (Embi et al., 2013; Moreira et al., 2018) that attempt to use the level three, but it appears that they also seem to simplify or deviate from the principle of this level. For example, Embi et al. (2013) covered the level three to evaluate transfer of skills in a blended learning environment in higher education. Their result based on student perception showed that students have applied their learning from a direct instruction method into reconstructivist learning. Wang (2018) similarly performed an evaluation study covering the four levels to gauge student learning outcomes as a result of undergoing an information organisation curriculum. Some questionnaire surveys were used specifically to evaluate the behaviour level and results level.

The same can be argued as explained earlier (Kirkpatrick & Kirkpatrick, 2006; Rouse, 2011; Covington, 2012). While the transfer of skills or change in behaviours can be better described through observation, Nickols (2000) reminded that, in using Kirkpatrick model to evaluate the impact of a higher education program to the learners, level three is supposed to measure the medium-term transfer of learned skills from the program to the work environment.

This limitation because of how evaluators use the model in the field may again put barriers and employing the model may be risky for stakeholders especially in education. Thus, the Kirkpatrick model is effectively employed at the lower levels only (Topno, 2012), whether in evaluating training effectiveness to teachers or program effectiveness on students. While higher levels have been attempted to be used in other evaluation efforts, it seemed that the Kirkpatrick model has been treated significantly simplistic. Paull et al. (2016) suggested that similar to challenges experienced in the work, education evaluators should ponder other the alternatives that may be employed to determine the outcomes based on levels three and four.

3.2. Rigidity which leaves out essential aspects of evaluand

The argument for the extreme rigidity of the levels of the Kirkpatrick model is put forward in the light of the importance of contextual factors and essential aspects of the program. This limitation is discussed by various researchers pointing out some features of the Kirkpatrick model with its four-level framework.

For one, according to Rouse (2011), the Kirkpatrick model oversimplifies effectiveness, not considering the various contextual factors within the program. This limitation was also acknowledged in the study of Lillo-Crespo et al. (2017) when they developed a framework adapting the Kirkpatrick model to evaluate the impact of healthcare improvement science. The team noted the weakness of the Kirkpatrick model as devoid of the consideration of contextual influences on the evaluation.

Yardley and Dornan (2012) also observed that, in their study in formal medical education, different levels necessitated different beneficiaries, i.e. levels one to three involve the students; level four relates to the organisations; the educators are overlooked from the system. Thus, they argued that the model does not explore multidimensional outcomes that can be ascertained through qualitative and quantitative approaches. It does not also elaborate on the underlying reasons why outcomes are the outputs of the particular inputs. It appears to gauge only the intended outcomes and disregard the unintended ones.

Furthermore, this problem was echoed by Abernathy (1999). He noted that each level tends to be particular on the questions posed and the outcomes generated, thus rigid. He precluded the levels as not appropriate to evaluate the soft outcomes and continuous education, which are typical in formal education.

3.3. Paucity of evidence on causal chains among the levels

An assumption of the Kirkpatrick model posits that all its levels are contributory (Alliger & Janak, 1989). Grounded on this assumption, scholars and practitioners postulate that, for example, reaction level has a causal influence on learning level. It is believed that the learning level further stimulates change at the behaviour level, and then leads to the desired results at the organisational level (Hilber et al., 1997; Kirkpatrick & Kirkpatrick, 2006).

This assumption can be applied in higher education, that is, what students acquire as a result of participation in the curriculum or program is supposed to cause changes in the reaction, learning, behaviour, and impact. Arthur et al. (2003) determined the relationships among the course grades, student learning, and teaching effectiveness which were reframed within reaction and learning levels. The results revealed that there is a moderate correlation between course grades and student learning. On the other hand, a low correlation was observed between learning measure and teaching effectiveness.

Moreover, Arthur et al. (2010), in their research in the field of technology education, failed to illustrate a piece of evidence of such a relationship between or among the levels. Their findings revealed that substantial relationships between the different levels are restricted. This implies that what long term outcomes learners exhibit might not necessarily be the result of the education they get in school. There could be other external factors that the model does not look into.

This result is empirically supported by Haupt and Blignaut (2007). They applied the Kirkpatrick model in their study to attempted to find out the outcomes in the learning of aesthetics in the program of design and technology education. Similarly, they were not able to show strong corroborations of the causal connections between or among the levels. They specifically were unable to show the link between levels two and three outcomes in their study.

Other related studies previously analysed (Embi et al., 2013; Moreira et al., 2018; Quintas et al., 2017; Dewi & Kartowagiran, 2018; Sahin, 2006; Aryadoust, 2017; Abdulghani et al. 2014; Usmani, 2015; Farjad, 2012; Baskin, 2001; Chang & Chen, 2014) employing Kirkpatrick model did not attempt to probe the causal links among the levels. This concern is not within the interest of these studies.

Tamkin et al. (2002) added that arguably the evaluation model of Kirkpatrick could be negatively attacked on the reasons that empirical studies conducted do not present evidence that the levels are significantly correlated. Hence, it is said to be simple of a thought and that it does not consider other essential features that affect learning. Thus, this limitation should be accounted for when conducting an evaluation using the Kirkpatrick model in higher education, and conclusions should be carefully drawn.

4. CONCLUSION

While Kirkpatrick model is gaining a reputation as a framework for program evaluation, however, it has its limitations in the field of higher education. It presents a propensity towards the use of the lower levels only, rigidity which leaves out other essential aspects of the evaluand, and paucity of evidence on the causal chains among the levels. These limitations offer opportunities and challenges for evaluators who plan to adapt the Kirkpatrick model in higher education evaluation.

First, the propensity towards the lower levels leaves a problem with the limited application of higher levels. This concern may be addressed by considering more appropriate methods and tools. For example, in the behaviour level which seeks to describe how learning has been transferred or has changed the behaviours of the participants in the workplace setting, it is strongly advised that a direct observation must be performed. It should be stressed that an evaluation of change in behaviour requires forceful evidence that goes beyond perceptions of participants usually generated from surveys. Additionally, ultimate outcomes in higher education measured by level four are not exclusively measured by monetary means. Thus, evaluators should redevelop their evaluation frameworks and redesign methods to appropriately evaluate this level.

Furthermore, to offset the argument for too much rigidity of the Kirkpatrick model, a deliberate effort should be made to integrate the contextual inputs and other essential aspects of the evaluand. This can be done by considering the individual participants, work environment, and other aspects that evaluators think are necessary to the framework. For example, a level may be contextualised to educational outcomes or some instruments may be designed to capture these contextual inputs or essential aspects in the light of such limitation. This way, while the levels of the Kirkpatrick model serve as a cogent guide in evaluation, there is room for flexibility so that evaluation will not be too fixed or detached from essential aspects of the evaluand.

Lastly, because the Kirkpatrick model is criticised for the lack of evidence showing the causal relationships among the levels, future studies should strive to prove these chains. The concept of causal relationships can be empirically established through the use of statistical tools. Thus, results in all levels must be converted to a quantitative set as much as possible. Much of the lower levels are often quantitatively measured. Where a level is qualitatively evaluated, data transformation models within the mixed method paradigm offer procedures to convert qualitative data to quantitative data. If causal relationships among the levels are provided attention in evaluation studies using the Kirkpatrick model, more comprehensive and appropriate conclusions may be drawn about the effectiveness of the curriculum or program.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Michael B. Cahapay: Conceptualization of the paper, methodology, literature search and evaluation, formal analysis, writing of the manuscript from its first to final form.

ORCID

Michael B. Cahapay  <https://orcid.org/0000-0002-0588-0022>

5. REFERENCES

- Abdulghani, H., A Al Drees, A.M., Khamis, N., & Irshad, M. (2014). Research methodology workshops evaluation using the Kirkpatrick's model: Translating theory into practice. *Medical Teacher*, 36(1), 24-29. <https://doi.org/10.3109/0142159x.2014.886012>
- Abernathy D.J. (1999). Thinking outside the evaluation box. *Training Development*, 53(2), 18-23. <https://eric.ed.gov/?id=EJ578905>
- Alliger, G., & Janak, E.. Kirkpatrick's Levels of Training Criteria: Thirty Years Later. *Personnel Psychology*, 42(2), 331-341. <https://doi.org/10.1111/j.1744-6570.1989.tb00661.x>

- Arthur, W., Jr., Bennet, W., Edens, P.S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88(2), 234-245. <https://doi.org/10.1037/0021-9010.88.2.234>
- Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Edens, P. S. (2010). Teaching effectiveness: The relationship between reaction and learning criteria. *Educational Psychology*, 23(3), 275-285. <https://doi.org/10.1080/0144341032000060110>
- Aryadoust, V. (2017). Adapting Levels 1 and 2 of Kirkpatrick's model of training evaluation to examine the effectiveness of a tertiary-level writing course. *Pedagogies: An International Journal*, 12(2), 151-179. <https://doi.org/10.1080/1554480X.2016.1242426>
- Bates, R. (2004). A critical analysis of evaluation practice: the Kirkpatrick Model and the principle of beneficence. *Evaluation and Program Planning*, 27, 341-347. <https://doi.org/10.1016/j.evalprogplan.2004.04.011>
- Baskin, C. (2001, December). Using Kirkpatrick's four-level-evaluation model to explore the effectiveness of collaborative online group work. In *Proceedings of the Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education* (pp. 37-44). Melbourne, Australia: Biomedical Multimedia Unit, The University of Melbourne.
- Bowen, G. A. (2009) Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. <https://doi.org/10.3316/qrj0902027>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- Chang, N. & Chen, L. (2014). Evaluating the learning effectiveness of an online information literacy class based on the Kirkpatrick framework. *Libri*, 64(3), 211-223. <https://doi.org/10.1515/libri-2014-0016>
- Covington, J.A. (2012). *Efficacy of webinar training for continuing professional education: applications for school personnel in k-12 settings* [Doctoral dissertation, University of North Carolina] <https://eric.ed.gov/?id=ED550661>
- Dewi. L.R., & Kartowagiran, B. (2018). An evaluation of internship program by using Kirkpatrick evaluation model. *Research and Evaluation in Education*, 4(2), 155-163. <https://doi.org/10.21831/reid.v4i2.22495>
- Embi, Z.C., Neo, T.K., & Neo, M. (2017). Using Kirkpatrick's evaluation model in a multimedia-based blended learning environment. *Journal of Multimedia Information System*, 4(3), 115-122, 2383-7632. <http://dx.doi.org/10.9717/JMIS.2017.4.3.115>
- Farjad, S. (2012). The Evaluation effectiveness of training courses in university by Kirkpatrick model. *Procedia – Social and Behavioral Sciences*, 46, 2837-2841. <https://doi.org/10.1016/j.sbspro.2012.05.573>
- Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications, Inc. <https://doi.org/10.4135/9781506326139>
- Guptill, A. (2016). Secondary sources in their natural habitats. *Writing in College*. <https://milnepublishing.geneseo.edu/writing-in-college-from-competence-to-excellence/chapter/secondary-sources-in-their-natural-habitats/>
- Haupt, G., & Blignaut, S. (2007). Uncovering learning outcomes: explicating obscurity in learning of aesthetics in design and technology education. *International Journal of Technology and Education*, 18(4), 361-374. <https://doi.org/10.1007/s10798-007-9029-1>
- Hilbert, J., Preskill, H., & Russ-Eft, D. (1997). Evaluating training's effectiveness. In L. Bassi & D. Russ-Eft (Eds.), *What works: Assessment, development, and measurement* (pp. 109–150). American Society for Training and Development.
- Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training and Development*, 13, 3-9.

- Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (2nd ed., pp. 301–319). New York: McGraw-Hill.
- Kirkpatrick D. L., & Kirkpatrick J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler Publication.
- Lambert, N. (2011). Ban happy sheets! - Understanding and using evaluation. *Nurse Education Today*, 32(1), 1-4. <https://doi.org/10.1016/j.nedt.2011.05.020>
- Liao, S. C., & Hsu, S. Y. (2019). Evaluating a continuing medical education program: New World Kirkpatrick Model Approach. *International Journal of Management, Economics and Social Sciences*, 8(4), 266-279. <http://dx.doi.org/10.32327/IJMESS/8.4.2019.17>
- Lillo-Crespo, M., Sierras-Davo, M. C., McRae, R., & Rooney, K. (2017). Developing a framework for evaluating the impact of Healthcare Improvement Science Education across Europe: A qualitative study. *Journal of Educational Evaluation in Health Profession*, 14, 28. <https://doi.org/10.3352/jeehp.2017.14.28>
- Masood, R., & Usmani, M. A. W. (2015). A study for program evaluation through Kirkpatrick's model. *Khyber Medical University Journal*, 2(7), 76-80. <https://www.kmu.edu.pk/article/view/15377>
- Miller, B.J. (2018). *Utilizing the Kirkpatrick model to evaluate a collegiate high-impact leadership development program* [Master's thesis, Texas A&M University, College Station, Texas]. <https://oaktrust.library.tamu.edu>
- Moreira I.C., Ramos, I., Ventura, S.R., Rodrigues, P.P. (2018). Learner's perception, knowledge and behaviour assessment within a breast imaging E-Learning course for radiographers. *European Journal of Radiology*, 111, 47-55. <https://doi.org/10.1016/j.ejrad.2018.12.006>
- Morrison J. (2003). ABC of learning and teaching in medicine: Evaluation. *British Medical Journal*, 326(7385), 385-387. <https://doi.org/10.1136/bmj.326.7385.385>
- Nickols, F. W. (2000). *Evaluating training: There is no "cookbook" approach*. http://home.att.net/~nickols/evaluating_training.htm
- Paull, M., Whitsed, C., & Girardi, A. (2016). Applying the Kirkpatrick model: Evaluating an Interaction for Learning Framework curriculum intervention. *Issues in Educational Research*, 26(3), 490-502. <https://www.iier.org.au/iier26/paull.pdf>
- Quintas, C., Fernandes Silva, I., & Tiexiera, A. (2017). Assessing an e-Learning and b-Learning Model - A study of perceived satisfaction. *International Journal of Information and Education Technology*, 7(4), 265-268. <https://doi.org/10.18178/ijiet.2017.7.4.878>
- Reio, T. G., Rocco, T. S., Smith, D. H., & Chang, E. (2017). A Critique of Kirkpatrick's Evaluation Model. *New Horizons in Adult Education & Human Resource Development* 29(2), 35-53. <https://doi.org/10.1002/nha3.20178>
- Rouse D. N. (2011). Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. *Perspectives of Health Information Management*, 8, 1-5. <https://www.ncbi.nlm.nih.gov/pubmed/21464860>
- Sahin, V. (2006). Evaluation of the in-service teacher training program "The Certificate for Teachers of English" at the Middle East Technical University School of Foreign Languages. [Doctoral dissertation, Middle East Technical University] <https://etd.lib.metu.edu.tr/upload/12607752/index.pdf>
- Shelton, S., & Alliger, G. M. (1993). Who's afraid of level 4 evaluation? A practical approach. *Training and Development Journal*, 47, 43-46. <https://eric.ed.gov/?id=EJ463549>
- Sileyew, K. J. (2019). Research design and methodology. In E. Abu- Taieh, A. El Mouatasim, & I.H. Al Hadid (Eds.), *Cyberspace*. IntechOpen. <https://doi.org/10.5772/intechopen.78887>

- Steele, L. M., Mulhearn, T. J., Medeiros, K. E., Watts, L. L., Connelly, S., & Mumford, M. D. (2016). How do we know what works? A review and critique of current practices in ethics training evaluation. *Accountability in Research*, 23(6), 319-350. <http://dx.doi.org/10.1080/08989621.2016.1186547>
- Tamkin, P., Yarnall, J., & Kerrin, M. (2002). *Kirkpatrick and beyond: A review of models of training evaluation* (Report No. 392). Institute for Employment Studies
- Topno, H. (2012). Evaluation of training and development: An analysis of various models. *IOSR Journal of Business and Management*, 5(2), 16-22. <https://doi.org/10.9790/487x-0521622>
- Yardley, S., & Dornan, T. (2012). Kirkpatrick's levels and education 'evidence'. *Medical Education*, 46(1), 97-106. <https://doi.org/10.1111/j.1365-2923.2011.04076.x>

The Effect of Item Pools of Different Strengths on the Test Results of Computerized-Adaptive Testing

Fatih Kezer ^{1,*}

¹Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli, Turkey

ARTICLE HISTORY

Received: May 10, 2020

Revised: Dec. 17, 2020

Accepted: Jan. 11, 2021

Keywords

Adaptive test,
Item pool,
Item difficulty,
CAT,
IRT.

Abstract: Item response theory provides various important advantages for exams carried out or to be carried out digitally. For computerized adaptive tests to be able to make valid and reliable predictions supported by IRT, good quality item pools should be used. This study examines how adaptive test applications vary in item pools which consist of items with varying difficulty levels. Within the scope of the study, the impact of items was examined where the parameter b differentiates while the parameters a and c are kept in fixed range. To this end, eight different 2000-people item pools were designed in simulation which consist of 500 items with ability scores and varying difficulty levels. As a result of CAT simulations, RMSD, BIAS and test lengths were examined. At the end of the study, it was found that tests run by item pools with parameter b in the range that matches the ability level end up with fewer items and have a more accurate estimation. When parameter b takes value in a narrower range, estimation of ability for extreme ability values that are not consistent with parameter b required more items. It was difficult to make accurate estimations for individuals with high ability levels especially in test applications conducted with an item pool that consists of easy items, and for individuals with low ability levels in test applications conducted with an item pool consisting of difficult items.

1. INTRODUCTION

The measurement and evaluation process plays a critical role in determining whether the qualities targeted to be acquired in education are realized or not. Change has undoubtedly been inevitable in measurement and evaluation just like it has been in every field throughout history. Although paper and pencil tests, which were based on the classical test theory, have been an important part of measurement and evaluation, they have certain important limitations and disadvantages. Item difficulty parameter and item discrimination parameter vary depending on the group from which data were collected; in other words, it varies according to sampling (Lord & Novick, 1968). Another limitation is that individuals' ability levels depend on item parameters. Individuals receive different scores in test batteries with different difficulty levels. One's ability may seem high in an easy test and low in a difficult test. Due to this important limitation, problems may arise in comparing the individual. Even when they could be compared,

CONTACT: Fatih Kezer ✉ fatih.kezer@kocaeli.edu.tr 📍 Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

because their ability levels are different, their ability scores could cause errors in different sizes (Hambleton, Swaminathan & Rogers, 1991). Tests developed according to traditional approaches and classical test theory usually work better with the individuals with intermediate ability levels (Crocker & Algina, 1986). When few items were designed for individuals with very low- and very high-level abilities, the test ceases to be distinctive for these ability levels, and reliable predictions cannot be made for these extreme ability levels. With existing test designs, it is not possible to know how an individual would perform with a given item set. The limitations of the theory put forth by Spearman in 1905 pioneered the formation of a new theory in 1930s. Item Response Theory (IRT) ties to eliminate limitations due to its strong assumptions (unidimensionality, local independence, model-data fit) and differences in the test algorithm. IRT is also called Latent Trait Theory (Crocker & Algina, 1986). This theory explains with a mathematical function the relationship between an individual's ability level related to the measured characteristic and the answers they give (Embretson & Reise, 2000; Hambleton & Swaminathan, 1989).

The most common item parameters in Item Response Theory are difficulty (b), discrimination (a), and chance (c). Parameter b is the ability (θ) level that corresponds to the point where the individual answers an item correctly with a 50% probability. It is also shown on the same scale as θ (Lord & Novick, 1968). Although it may theoretically take a value between $-\infty$ and $+\infty$, it usually takes in practice a value in the -3 and $+3$ range. An increase in b denotes that the item is getting more difficult and a decrease indicates that it is getting easier. When parameter b is 0, it denotes a medium-level difficulty. Item discrimination (a) parameter corresponds to the curve on the $\theta=b_i$ point. Theoretically, ranges from $-\infty$ to $+\infty$, however in practice it usually takes a value between 0 and 2. Parameter a can take a negative value, albeit rarely, and this indicates that the item works in the opposite direction. Parameter c denotes the probability of individuals giving a correct answer by guessing.

An important advantage of Item Response Theory is that item and test information functions can be obtained. Item information function shows how much information an item gives of its measured characteristic. Item information is inversely proportional to item error variant (Reid, Kolakowsky-Hayner, Lewis & Armstrong, 2007). A function that takes up a different value in every point of θ is calculated by the equation given below (Baker & Kim, 2004; Hambleton, Swaminathan & Rogers, 1991).

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

For a three-parameter logistic model, this equation is expressed as follows with item parameters:

$$I_i(\theta) = \frac{2.89 a_i^2(1-c_i)}{[c_i + e^{1.7 a_i(\theta-b_i)}][1 + e^{-1.7 a_i(\theta-b_i)}]^2}$$

As parameter a increases and parameter b gets closer to zero, $I(\theta)$ value increases as well. Parameter b getting closer to θ increases $I(\theta)$. The total of item information functions gives the test information function that shows how much the test gives information about the measured characteristic (Hambleton, Swaminathan & Rogers, 1991; Reid et al., 2007).

$$I(\theta) = \sum_{i=1}^n I_i(\theta)$$

Given the item and test information functions, item characteristics in forming a test is important to be able to have a valid and reliable measuring. IRT provides significant support to measuring processes with its mathematical basis. The invariance characteristic of IRT enables item and test parameters to be independent from the group, and it enables predicted ability levels to be independent from the test. As such, it is possible to compare measuring results of different groups. Being able to calculate the reliability not for a single item but for each of them and for each ability level separately, and also being able to calculate errors separately for each individual enables a shorter test with quality items (Adams, 2005; Crocker & Algina, 1986; Embretson & Reise, 2000; Magnusson, 1966). With its strong mathematical structure, IRT is convenient for various applications. The most important of these are test design, item mapping, test equating, test and item bias studies and computerized adaptive test applications.

In classical tests, a fixed number of items are designed to be applied to all individuals. Adaptive tests, on the other hand, are based on the principle that items appropriate to an individual's ability are used. Thus, the test is cleared of inappropriate items so that it becomes both shorter and more reliable. With the advancement of technology, adaptive tests have begun to be applied more, and computerized adaptive tests (CAT) have gained more importance. In the application of CAT to individuals by selecting items from a large item pool, there are different methods (two-stage testing, self-selecting testing, pyramidal multistage testing, alternating testing, stradaptive testing, multilevel format) (Glas & Linden, 2003; Hambleton & Swaminathan, 1989; Thompson & Weiss, 1980; Vale & Weiss, 1975; Weiss, 1985). Adaptive test strategies are designed to use item information obtained through item information function (Brown & Weiss, 1977; Maurelli & Weiss, 1981; Weiss & Kinsbury, 1984).

The main aim of CAT is to apply the item cluster that gives most information for each individual. To this end, individuals are given different item sets, and based on the answers given to these item sets, an ability estimation is done. Contrary to CTT, CAT is based on IRT and CAT's test logic is based on large item pools item parameters which are known beforehand. Item pool can consist of different item types (Embretson & Reise, 2000; Sukamolson, 2002; Wainer et al., 2000). This testing method requires an item pool which is comprised of items that have high discrimination and that are distributed in a balanced manner on the difficulty-ability level ($b-\theta$) so that it can make estimations for individuals at different ability levels (Geordiadou, Triantafillou & Economides, 2006; Veldkamp & Linden, 2010; Weiss, 1985, 2011). In practice, it is not that easy to form an item pool whose item parameters take value in a large range. In this study, it was examined how estimation of ability changes when item pools consist of items with different characteristics, what kind of differences in the testing would application changing parameter b providing parameters a and c remain in the same range make. Moreover, it examined how estimations of ability changes by creating conditions in which parameter b takes value between narrow and wide ranges and where there is conglomeration at different points from easy to difficult.

2. METHOD

This study is designed as a basic research model in which the psychometric qualities of application results of computerized adaptive tests with items culled from item pools with different difficulty levels, are examined. Basic research refers to those studies that are conducted based on theories, by developing assumptions, testing them, and scientifically interpreting their results (Karasar, 2016).

2.1. Simulation Design

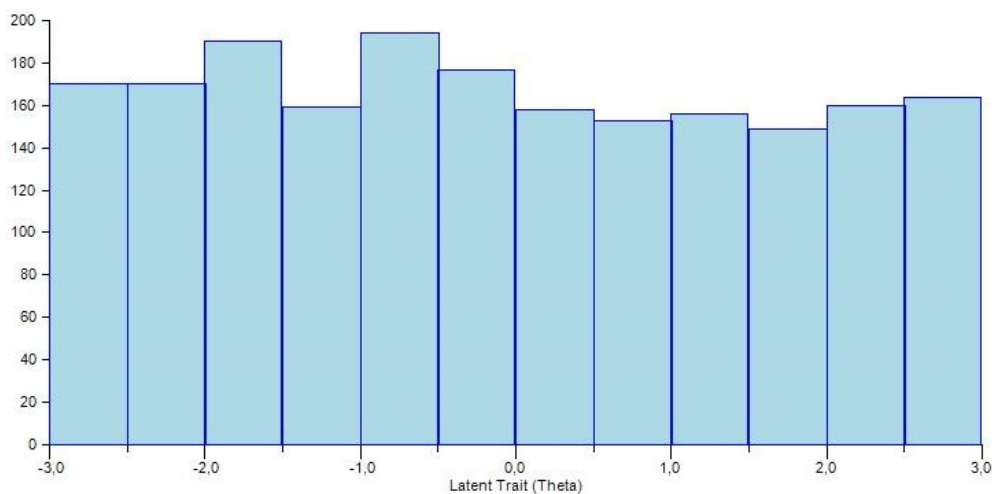
In line with the aim of the study, data were generated in simulation with SIMULCAT Monte-Carlo simulation to compare different item pools. Developed by Kyung T. Han in 2020, SimulCAT is a software to carry out simulated adaptive test applications. When algorithms and codes of practice of adaptive tests are considered, one needs large item pools developed according to item response theory as well as estimated ability parameters from large groups. In this respect, simulative data that could represent each special condition were used in this study. The study was conducted based on a three-parameter model. First, ability parameters were estimated so that they represent a 2000-people group. To estimate the ability, θ (theta) was defined within the -3 and +3 range. Descriptive statistics for estimated ability parameters are presented in [Table 1](#).

Table 1. Descriptive statistics of ability scores.

Statistics	Value
N	2000
Mean	-0.073
Median	-0.148
Minimum	-3.000
Maximum	3.000
Range	6.000
Standard Deviation	1.728
Variance	2.985
Skewness	0.091
Std. Error of Skewness	0.055
Kurtosis	-1.188
Std. Error of Kurtosis	0.109

As can be seen in [Table 1](#), mean of the ability parameters generated at (-3, +3) range was found to be -0.073, and its standard deviation 1.718. The same ability parameters (2000-people) were used for all conditions. Distribution related to estimated ability parameters are presented in [Figure 1](#).

Figure 1. Distribution of ability parameters.



Eight different conditions were formulated to be able to examine estimated parameters from item pools with different difficulty levels. There are 500 items in each item pool. To see the effect of average difficulty levels, discrimination (a) and chance (c) parameters were defined within the same range so that other conditions remain the same. Parameter a was kept within 0.25 and 2.00, and parameter c within 0.00 and 0.20. Difficulty parameter (b) was defined as a range for each 3 conditions: it was between -3 and +3 for the first condition, -2 and +2 for the second condition and was between -1 and +1 for the third condition. Other than the three ranges, five different conditions were also determined according to average difficulty. In these five different conditions, parameter b was defined as -2.5, -1.5, 0.0, 1.5, and 2.5, respectively, keeping standard deviation as 1.5. Item parameters related to these eight conditions are summarized in Table 2.

Table 2. Item parameters (defined/generated) for eight different conditions.

	Defined			Generated					
	b	a	c	b		a		c	
				\bar{X}	Sd	\bar{X}	Sd	\bar{X}	Sd
1 st Condition	(-3.0,+3.0)			0.017	1.763	1.121	0.498	0.100	0.058
2 nd Condition	(-2.0,+2.0)			0.013	1.173	1.139	0.502	0.100	0.058
3 rd Condition	(-1.0,+1.0)			0.026	0.594	1.150	0.504	0.101	0.058
4 th Condition	$\bar{X}=2.5$	(0.25,2.0)	(0.0,0.2)	2.373	1.567	1.098	0.498	0.098	0.058
5 th Condition	$\bar{X}=1.5$			Sd=1.5	1.417	1.536	1.124	0.524	0.101
6 th Condition	$\bar{X}=0.0$	Sd=1.5		0.023	1.406	1.139	0.495	0.102	0.057
7 th Condition	$\bar{X}=-1.5$	Sd=1.5		-1.599	1.494	1.138	0.503	0.103	0.057
8 th Condition	$\bar{X}=-2.5$	Sd=1.5		-2.577	1.514	1.122	0.522	0.096	0.059

In the adaptive test application design, Maximum Fisher Information (MFI) was used, which is the most common method for item selection management. As initial ability parameter, (-0.5, +0.5) range was determined. Maximum Likelihood Method (MLE) was selected for all conditions as the estimation of ability method. Maximum Likelihood Method is based on selecting the item that gives out most information about an individual. As the termination rule, a common rule was likewise selected for the eight conditions. Standard error which is smaller than 0.30 was determined as the test termination rule. Half the amount of the item pool – 250 items – was decided to be an upper termination rule because too many items would be needed for the estimation of ability if item pool is not appropriate. While conducting the test in inappropriate item pools, the test was stopped when half of the pool is reached. 25 repetitions were made for estimations.

2.2. Data Analysis

In the evaluation of test findings, Root Mean Squared Difference (RMSD) and BIAS values were used. RMSD is a statistic that denotes the difference between estimations of ability (Boyd, Dodd & Fitzpatrick, 2013). BIAS is a difference statistic between the ability parameter average value and its real value. RMSD and BIAS are calculated by using the following formula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2}{N}} \quad BIAS = \frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)}{N}$$

Moreover, test lengths were also checked in the ability parameter ranges for eight different conditions. The aim was to have a detailed examination of how long the test would take for individuals at different ability levels in the response cluster. Therefore, RMSD and BIAS values at ability ranges were examined.

3. RESULT / FINDINGS

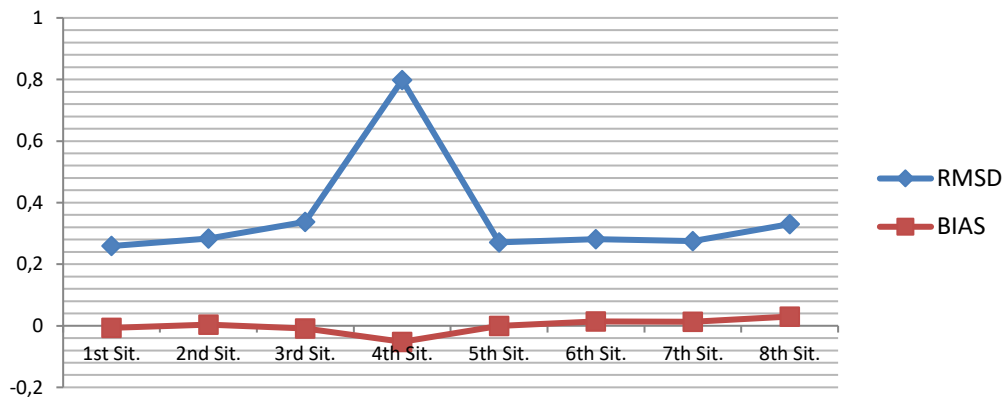
In line with the aim of this study, RMSD and BIAS values for ability parameters obtained from adaptive tests, which were conducted with item pools with different difficulty levels, were calculated and presented in [Table 3](#).

Table 3. RMSD and BIAS values concerning estimation of ability.

Condition	b	RMSD	BIAS
1 st Condition	(-3.0,+3.0)	0.259	-0.007
2 nd Condition	(-2.0,+2.0)	0.283	0.004
3 rd Condition	(-1.0,+1.0)	0.338	-0.009
4 th Condition	$\bar{X}=2.5$ Sd=1.5	0.798	-0.052
5 th Condition	$\bar{X}=1.5$ Sd=1.5	0.271	-8X10 ⁻⁵
6 th Condition	$\bar{X}=0.0$ Sd=1.5	0.281	0.014
7 th Condition	$\bar{X}=-1.5$ Sd=1.5	0.275	0.013
8 th Condition	$\bar{X}=-2.5$ Sd=1.5	0.330	0.030

Since 25 repetitions were done in estimations of parameter, obtained results were turned into a report by taking their average. As can be seen in [Table 3](#), RMSD values vary between 0.259 and 0.798. Except for the 4th condition, RMSD values were in a narrower range (0.259-0.338). The lowest RMSD value was obtained, as expected, from the condition in which the difficulty parameters of items in the item pool were between -3 and +3. This value increased when the range of parameter b comparatively narrowed. Apart from when the average was 2.5 in item pools which were formed by considering, the averages of parameter b, no significant difference was detected. Distribution related to RMSD and BIAS values are shown in [Figure 2](#).

Figure 2. Distribution of RMSD and BIAS values concerning estimations of ability.



The array of RMSD values according to their size were found to be $RMSD_{Cnd.1} < RMSD_{Cnd.5} < RMSD_{Cnd.7} < RMSD_{Cnd.6} < RMSD_{Cnd.2} < RMSD_{Cnd.8} < RMSD_{Cnd.3} < RMSD_{Cnd.4}$. Similarly, BIAS values concerning different conditions varied absolutely between 0.00008-0.052. Lengths of the simulated adaptive tests were considered separately in θ ranges. Distribution concerning the test lengths are given in [Table 4](#).

Table 4. Lengths of the simulated adaptive tests.

θ	N	1 st Cnd.	2 nd Cnd.	3 rd Cnd.	4 th Cnd.	5 th Cnd.	6 th Cnd.	7 th Cnd.	8 th Cnd.
-3.0< θ <-2.5	170	13.01	55.06	250.00	250.00	147.55	13.57	12.42	11.19
-2.5< θ <-2.0	170	11.91	15.84	182.84	250.00	22.84	12.78	12.56	10.38
-2.0< θ <-1.5	190	11.58	10.91	46.72	134.38	13.39	12.23	12.55	10.58
-1.5< θ <-1.0	159	11.36	11.20	16.42	17.58	13.69	12.08	11.66	11.05
-1.0< θ <-0.5	194	11.23	11.36	11.10	14.29	12.70	12.11	11.45	10.89
-0.5< θ <0.0	177	10.97	11.01	10.89	14.33	12.10	12.58	11.84	11.64
0.0< θ <0.5	158	11.69	10.57	10.65	13.39	11.53	11.73	11.61	14.15
0.5< θ <1.0	153	11.35	10.23	9.88	12.85	11.58	11.30	12.73	17.97
1.0< θ <1.5	156	11.27	11.09	12.97	11.82	11.19	11.21	14.06	27.53
1.5< θ <2.0	149	11.12	10.82	26.77	11.25	11.51	12.67	20.33	138.72
2.0< θ <2.5	160	12.86	15.68	76.79	11.25	11.56	12.46	61.95	250.00
2.5< θ <3.0	164	12.44	35.32	198.61	11.40	11.31	13.12	238.54	250.00

Likewise, RMSD and BIAS values calculated for different conditions for each ability range are given in Table 5.

Table 5. RMSD and BIAS values according to ability ranges.

	θ Area	-3.0	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5
	N	170	170	190	159	194	177	158	153	156	149	160	164
1 st Cnd.	Bias	-0.02	-0.03	0.00	0.03	-0.03	0.00	-0.01	-0.01	-0.03	-0.02	-0.02	0.06
	RMSD	0.30	0.27	0.25	0.25	0.26	0.27	0.23	0.27	0.26	0.25	0.25	0.26
2 nd Cnd.	Bias	-0.08	-0.03	0.00	0.02	-0.01	-0.01	0.00	0.00	-0.01	-0.01	0.09	0.09
	RMSD	0.36	0.31	0.31	0.25	0.24	0.25	0.27	0.23	0.24	0.25	0.35	0.30
3 rd Cnd.	Bias	-0.12	-0.05	-0.07	-0.05	-0.03	-0.01	0.01	0.02	0.06	0.08	0.01	0.09
	RMSD	0.60	0.36	0.31	0.28	0.26	0.25	0.26	0.26	0.33	0.32	0.27	0.41
4 th Cnd.	Bias	-0.41	-0.08	-0.15	0.00	0.01	-0.02	0.04	0.02	0.01	-0.01	0.02	-0.01
	RMSD	2.03	0.83	1.36	0.30	0.25	0.25	0.25	0.29	0.26	0.25	0.25	0.25
5 th Cnd.	Bias	-0.06	0.00	0.02	-0.01	0.01	0.02	0.01	0.01	0.00	0.00	-0.02	0.01
	RMSD	0.40	0.27	0.26	0.26	0.26	0.27	0.25	0.24	0.25	0.26	0.24	0.26
6 th Cnd.	Bias	-0.06	0.05	0.02	0.02	-0.03	0.02	0.01	0.05	-0.03	0.01	0.06	0.05
	RMSD	0.46	0.28	0.24	0.25	0.26	0.25	0.25	0.24	0.26	0.24	0.25	0.32
7 th Cnd.	Bias	0.00	0.01	0.00	0.02	-0.01	0.02	0.02	0.03	0.01	0.00	0.03	0.03
	RMSD	0.25	0.28	0.26	0.25	0.24	0.24	0.26	0.28	0.27	0.32	0.33	0.32
8 th Cnd.	Bias	-0.02	0.02	-0.03	0.03	0.00	0.03	0.05	0.00	0.01	0.05	0.06	0.20
	RMSD	0.26	0.23	0.25	0.25	0.25	0.26	0.29	0.31	0.30	0.32	0.43	0.64

When Table 3 and Table 4 are examined, it can be seen in which ability range item pools with different characters would work more ideally. In the item pool where parameter b is between -3.0 and +3.0 (1st condition), the test was completed, as expected, at a more reasonable time. RMSD and BIAS values were similar and low in each range. In the 6th condition, it was seen that the test length was reasonable for every ability level when parameter b was heaped up around the intermediate difficulty level ($\bar{X}=0.0$, $Sd=1.5$). Keeping in mind the ability (θ)-difficulty (b) relationship of IRT, it can be said that when difficulty was kept at moderate level, more decisive estimations are done for a large ability range. In the 2nd and 3rd condition in which parameter b was kept within a limited range, it was seen that more items were needed to decisively estimate ability as one moves towards the ends where ability level is high or low. In the 2nd condition, number of items needed at extreme ability levels moved up to 55. In the adaptive test simulation ran in the item pool with parameter b at the (-1.0, +1.0) range, which

is a more limited range, (3rd condition), test lengths went outside of acceptable limits in extreme ability levels. The second termination rule of the study – stopping the test when half of the item pool is reached – worked in these three extreme ability levels, and the test was stopped before it could become consistent. This was reflected in RMSD and BIAS values. RMSD value increased to 0.60 in the (-3.0, -2.5) ability range. There was a similar case in item pools which were formed as normal distribution within a certain parameter b . Except for the 6th condition ($\bar{X}=0.0$ $Sd=1.5$), more items were needed in ranges where parameters b do not correspond to ability levels. As can be seen [Table 4](#), 5th condition-7th condition or 4th condition -8th condition worked adversely and were more decisive in different ability levels. In item pools which were designed by determining parameter b approximately as $\bar{X}=2.5$, the test was stopped by reaching the defined maximum item number without the estimation falling below the standard error value at the $-3<\theta<-2$ range. Similarly, in the 8th condition, the test was stopped as maximum item number was reached at $2<\theta<3$ range. It was observed that RMSD and BIAS values increased in inappropriate ability levels in parallel to test length.

4. DISCUSSION and CONCLUSION

Although classical paper and pencil tests are prevalently used in education and psychology, they give way to electronic exams with the advancements in technology and assessment theories. Item response theory (IRT) provides various important advantages for exams carried out or to be carried out digitally. For computerized adaptive tests to be able to make valid and reliable predictions supported by IRT, good quality item pools should be used (Hambleton, Swaminathan & Rogers, 1991; Weiss, 1985). In adaptive test designs, from 50% to 80% could be saved in test length (Bulut & Kan, 2012; Comert, 2008; Iseri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; McDonald, 2002; McBride & Martin, 1983; Olsen, Maynes, Slavvson & Ho, 1989; Oztuna, 2008; Scullard, 2007; Smits, Cuijper & Straten, 2011). With CAT, each individual can get a test appropriate for his or her ability level. Moreover, the speed of the test can be adaptive for the individual. Because it is computerized, individuals can take the test at different times where as classical paper and pencil tests everyone should sit in at the same time. Different question formats can be easily used within a test. Test results can be assessed immediately, and test standardization is easier. As an important point, a test that works effectively and properly at every ability level is designed from test that bespeaks to intermediate-level individuals. In order to do a computerized adaptive test that has these advantages, one needs large item pools of which item parameters are estimated beforehand. It is not always easy to write items that has these qualities. Quality of the pool is an important factor that affects efficiency of application. This study examined what kind of results one would get in CAT applications of item pools which have items with different characteristics. Within the scope of the study, the impact of items was examined where the parameter b differentiates while keeping the parameters a and c are kept in fixed range. At the end of the study, it was seen that tests run by item pools with parameter b in the range that matches the ability level end up with fewer items and have a more decisive prediction. Similar studies in literature also underscore when the θ - b relationship is high, more effective CAT applications are carried out (Chang, 2014; Eggen & Verschoor, 2006; Dodd, Koch & Ayala, 1993). When parameter b takes value in a narrower range, estimation of ability for extreme ability values that are not compatible with parameter b required more items. What is more, accurate estimations could not be done with a decent number of items for extreme ability values in much narrower ranges ($-1<b<+1$). Since one could not go below the desired standard deviation, it was difficult to make accurate estimations for individuals with high ability levels especially in test applications conducted with an item pool that consists of easy items, and for individuals with low ability levels in test applications conducted with an item pool consisting of difficult items. These results underline that when generating an item pool in adaptive test applications, one should be incredibly careful.

To make CAT more effective and functional, it can be said that the dimension of the item pool should be as such that would cover all values of b (Chang, 2014). Using items with inappropriate difficulty levels without considering the characteristics of the target group would put adaptive test applications in jeopardy from test length to estimation of ability. The effect of items' levels of difficulty on adaptive test applications can be tested by different item discrimination values at different ability ranges. To this end, examining item parameters would guide teachers and test designers when they form item pools. Moreover, knowing the characteristics of the item pool and its effects could help test designers in constructing correct control mechanisms in test algorithm.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Fatih Kezer: All chapters are written by the author.

ORCID

Fatih Kezer  <https://orcid.org/0000-0001-9640-3004>

5. REFERENCES

- Adams, R. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Marcel Bekker Inc.
- Boyd, A. M., Dodd, B. & Fitzpatrick, S. (2013). A comparison of exposure control procedures in cat systems based on different measurement models for testlets. *Applied Measurement in Education, 26*(2), 113-115.
- Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research, 49*, 61–80.
- Chang, H. H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika, 1*-20.
- Cömert, M. (2008). *Bireye uyarlanmış bilgisayar destekli ölçme ve değerlendirme yazılımı geliştirilmesi [Computer-aided assessment and evaluation analysis adapted to the individual]* [Unpublished master's thesis]. Bahçeşehir University.
- Crocker, L., & Algina, J. (1986). *Introduction classical and modern test theory*. Harcourt Brace Javonovich College Publishers.
- Dodd, B. G., Koch, W. R., & Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*(1), 61-77.
- Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*(5), 379-393.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Technology, 37*(2), 261–278.

- Glas, C. A., & Linden, W. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27(4), 247–261.
- Hambleton, R. K., & Swaminathan, H. (1989). *Item response theory: Principles and applications*. Kluwer Nijhoff Publishing.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications Inc.
- Iseri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures* [Unpublished doctoral dissertation]. Middle East Technical University.
- Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished doctoral dissertation]. Middle East Technical University.
- Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kâğıt-kalem testi uygulamasının karşılaştırılması [Comparison of adaptive (individualized) test application and traditional paper-pencil test application in ability estimation]* [Unpublished doctoral dissertation]. Hacettepe University.
- Karasar, N. (2016). *Bilimsel araştırma yöntemi [Scientific Research Method]*. Nobel Yayın Dağıtım.
- Kezer, F. (2013). *Comparison of the computerized adaptive testing strategies* [Unpublished doctoral dissertation]. Ankara University.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison - Wesley.
- Magnusson, D. (1966). *Test theory*. Addison-Wesley Publishing Company.
- Maurelli, V. A., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.
- McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military design. In Weiss, D.J. (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. Academic Press.
- McDonald, P. L. (2002). *Computer adaptive test for measuring personality factors using item response theory* [Unpublished doctoral dissertation]. The University Western of Ontario.
- Olsen, J. B., Maynes, D. D., Slavvson, D., & Ho, K. (1989). Comparison of paper administered, computer administered and computerized adaptive achievement tests. *Journal of Educational Computing Research*, 5(31), 311-326.
- Öztuna, D. (2008). *An application of computerized adaptive testing in the evaluation of disability in musculoskeletal disorders* [Unpublished doctoral dissertation]. Ankara Üniversitesi Sağlık Bilimleri Enstitüsü.
- Reid, C. A., Kolakowsky-Hayner, S. A., Lewis, A. N., & Armstrong, A. J. (2007). Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counselling Bulletin*, 50(3), 177-178.
- Scullard, M.G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory* [Unpublished doctoral dissertation]. University of Minnesota.
- Smits, N., Cuijpers, P., & Straten, A. (2011). Applying computerized adaptive testing to the CES-D Scale: A simulation study. *Psychiatry Research*, 188, 145–155.
- Sukamolson, S. (2002). Computerized test/item banking and computerized adaptive testing for teachers and lecturers. http://www.stc.arts.chula.ac.th/ITUA/Papers_for_ITUA_Proceedings/Suphat2.pdf
- Thompson, J. G., & Weiss, D. J. (1980). *Criterion-related validity of adaptive testing strategies* (Research Rep. No. 80-3). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

- Vale, C. D., & Weiss, D. J. (1975). *A study of computeradministered stradaptive ability testing* (Research Rep. No. 75-4). University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Veldkamp, B. P., & Linden. W. J. (2010). Designing item pools for adaptive testing. In Linden, W.J., and Glas, C.A.W.(Eds.). *Elements of adaptive testing*. Springer.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: a primer*. Lawrence Erlbaum Associates, Publishers.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53(6), 774-789.
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1–27.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361-375.

Defining Cut Point for Kullback-Leibler Divergence to Detect Answer Copying

Arzu Ucar ^{1,*}, Celal Deha Dogan ²

¹Hakkari University, Faculty of Education, Department of Educational Sciences, Hakkari, Turkey

²Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Sep. 06, 2020

Revised: Dec. 25, 2020

Accepted: Jan. 16, 2021

Keywords:

Copy detection,
Cut point,
ROC analysis,
Answer copying,
Kullback-Leibler
divergence.

Abstract: Distance learning has become a popular phenomenon across the world during the COVID-19 pandemic. This led to answer copying behavior among individuals. The cut point of the Kullback-Leibler Divergence (KL) method, one of the copy detecting methods, was calculated using the Youden Index, Cost-Benefit, and Min Score p-value approaches. Using the cut point obtained, individuals were classified as a copier or not, and the KL method was examined for cases where the determination power of the KL method was 1000, and 3000 sample size, 40 test length, copiers' rate was 0.05 and 0.15, and copying percentage was 0.1, 0.3 and 0.6. As a result, when the cut point was obtained with the Min Score p-value approach, one of the cutting methods approaches, it was seen that the power of the KL index to detect copier was high under all conditions. Similarly, under all conditions, it was observed that the second method, in which the detection power of the KL method was high, was the Youden Index approach. When the sample size and the copiers' rate increased, it was observed that the power of the KL method decreased when the cut point with the cost-benefit approach was used.

1. INTRODUCTION

Due to the COVID 19 pandemic period, some exams are required to be administered online, and this situation may increase the examinees' motivation for cheating. Therefore, examinees who cheat and do not cheat should be distinguished to minimize the measurement error that may arise from the copying. Cheating behavior risks the validity of the inferences about students' competence and skills. Cheaters should be detected to minimize the systematic error that may be caused by cheating behavior. Cheating is one of the aberrant behaviors of examinees. Numerous statistical techniques have been developed to detect aberrant response patterns and test scores of examinees. Those techniques detect anomalies associated with different cheating types. There are two main types of cheating behavior. Individual cheating occurs when the student cheats during the exam from a source (other examinees, books, notes, smartphones, etc.). On the other hand, group cheating occurs when at least two examinees cheat in cooperation during or before the exam. Group cheating usually happens when some of the test items are revealed, and a group of examinees shares test items with each other before or

CONTACT: Arzu Uçar ✉ arzukapcik@gmail.com 📍 Hakkari University, Faculty of Education, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

during the exam. Although research on methods used to detect cheating has primarily focused on individual cheating, some methods are used to identify group cheating recently (Belov, 2013; Wollack & Maynes, 2017).

Many studies involve the use of multiple statistical methods to detect individual and group cheating (Karabatson, 2003; Meijer & Sijtsma, 2001; Meijer & Tendeiro, 2014; Wollack, 2006). The methods used to detect individual cheating can be classified as answer copying and similarity analysis, person-fit statistics, relationships between scores on subsets of items within the test, and model approaches (IRT models embedding aberrant behaviors (He, Meadows & Black, 2018). Answer copying and similarity analysis include numerous methods such as Angoff's B and H indices, K index, g2 index, ω index, S1, and S2 indices, VMIs (Variable Match) indices ξ and ξ^* indices (Belov, 2011), Wesolowsky's Z similarity index (Wesolowsky, 2000), Generalized Binomial Test (GBT) index (Shu, Henson and Luecht, 2013) and M4 (Maynes, 2014). Person fit statistics differ according to the type of items (dichotomous and polytomous) and the type of model (parametric, non-parametric). Kullback-Leibler Divergence, MPI (Matched Percentile) index, IRI (Irregularity) index, Z-test statistics are the methods used to detect copiers based on the relationships between the scores on subsets of items within the test. DG (Deterministic, Gated IRT Model) model is a commonly used technique in the model approach.

Kullback-Leibler Divergence (KL) is a measure of information used in psychometric practice. It is used as an item selection method in Computerized Adaptive Testing (Chang & Ying, 1996). However, it is also used to detect individuals who cheat (Belov, 2014a, 2014b). KL gives the difference between the two distributions. For instance, we used a test to obtain ability distributions before and after manipulation. Hence previous exam results indicated that the examinees do not cheat. The posterior ability distribution of the person who cheats was compared with the posterior ability distribution of the person who did not cheat. Then, we obtained a different value. The greater value of the difference is the greater difference between the individuals' performance in both tests (Belov & Armstrong, 2010). There are many reasons for the difference in distributions. However, what we are interested in is the differentiation that occurs due to cheating.

KL has been a commonly used technique to detect individual copiers because it can be practically used when we have preknowledge about the examinees' ability. KL is one of the methods to detect the copiers. To implement the KL method, we need to find out the cut point used during the individuals' classification under various conditions. However, no standard cut point can be used to classify students with KL values, which interprets KL results vague. Also, no study focuses on defining cut points for KL. Therefore, in this study, it was aimed to obtain the cut point of KL values with two different approaches (Min score P-Value, ROC) and compare the performances (power) of those approaches under various conditions.

1.1. Purpose of the Study

In this study, we aim to define the cut point for Kullback-Leibler Divergence in different conditions. Following are the research questions:

1. What is the cut point for Kullback-Leibler Divergence based on
 - a) Youden Index
 - b) Cost-Benefit approach

in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%) and copying percentages (10%, 30%, 60%)?

2. What is the cut point for Kullback-Leibler Divergence based on the Min Score p-value approach in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%), and copying percentages (10%, 30%, 60%)?

3. What is the power of Kullback-Leibler Divergence based on

- a) Youden Index
- b) Cost-Benefit approach
- c) Min Score p-value approach

in different sample sizes (N=1000, N=30000), copiers' rate (5%, 15%) and copying percentages (10%, 30%, 60%)?

2. METHOD

2.1. Research Design

This research is a simulation study in which some variables are manipulated. We design the levels of the variables considering the previous similar studies and real-life conditions. While in previous studies, test difficulty was defined into three levels (easy, medium, and difficult), we decided to fix this variable as the medium because it reflects real-life conditions (Sunbul & Yormaz, 2018; Zopluoglu, 2016).

The copier's ability and the source is another variable that might affect the power of the copy index (Sotaridona & Meijer, 2002; Steinkamp, 2017; Sunbul & Yormaz, 2018). In the study of van der Linden and Sotaridona (2006), the indexes' power was found high for the cases when low ability individual copies the responses from the high ability one. High ability individuals rely on their knowledge in the tests and answer the items on their own. On the other hand, low ability individuals are more likely to copy someone else's answers (Voncken, 2014). Therefore, during the exams, they copy the answers from their peers. In the light of this information, in this study, we decided to fix the ability of the copier as low and the source of the copier as high because of the real-world scenario that we are high likely to experience.

The copier's ability and the source is another variable that might affect the power of the copy index (Sotaridona & Meijer, 2002; Steinkamp, 2017; Sunbul & Yormaz, 2018). We fixed the copier's ability as lower and that of the source as upper because in real-world generally lower ability individuals copy from the individuals who have the upper ability.

In previous studies, the test length was commonly defined as 40 and 80 items. Because in the real-world, large-scale tests often include approximately 40 items in a sub-test, we decided to fix the test length as 40 (Sotaridona & Meijer, 2002, 2003, Sunbul & Yormaz, 2018; Yormaz & Sunbul, 2017; Wollack, 1997, 2003; Zopluoglu, 2016).

Regarding the related literature, the copier ratio is manipulated as 5% and %15 (Steinkamp, 2017). In the previous studies comparing the power and type 1 error of the copy index, both small and large sample sizes were utilized (from 50 to 10000). However, to be prevented from biased estimations about the item and person parameters, we manipulated sample size as 1000 and 3000 (Hurtz & Weiner, 2018; Sunbul & Yormaz, 2018; van der Linden & Sotaridona, 2006; Yormaz & Sunbul, 2017; Wollack, 2003; Zopluoglu, 2016; Zopluoglu & Davenport, 2012). Based on the relevant literature, in this study, we manipulated copying percentage as lower (10%), medium (30%), and upper (60%). Considering the manipulated variables, we tested 12 conditions (sample size-2 x copiers' percentage-2 x copying percentage -3 = 12). Table 1 presents the manipulated and fixed conditions in the study.

Table 1. Simulation Design Conditions and Levels.

Condition	Number of Levels	Level Values
Sample Size	2	1000, 3000
Copiers' Percentage	2	5%, 15%
Copying Percentage	3	10%, 30%, 60%
Test Difficulty*	1	Medium
Person Parameter of Source/Copier*	1/1	Upper-Lower
Test Length*	1	40

*fixed variable

2.2. Simulation Data

The Rasch model is one of the Item Response Theory models. It has some advantages, such as being mathematically less complex and easy to apply. Moreover, it is the most frequently used model in the exam programs because encountering parameter estimation problem is less. Therefore, we used the Rasch model in this study. The ability of 10000 participants and the difficulty parameter of 40 items was produced under the standard normal distribution $N(0,1)$. Considering the population's abilities and the difficulty parameters of the test items, dichotomous (1-0) response matrices were simulated based on the Rasch model. For the simulations, we utilized the "mirt" package (Chalmers, 2019) in the R program.

Sunbul and Yormaz (2018) denoted the ability level of the copiers as (-3.00, -1.50), (-1.50, 0.00), and the ability of the source as (0.00, 1.50), (1.50, 3.00). We denoted the ability of copiers in a wider range. In this way, we reduced the interference with the ability level of the copier. In addition to this, since the performance of similarity indices in identifying copiers increases with the increase of the difference between the ability levels of the copier and the source, we selected the ability of the source individuals (1.51, 3) from the individuals with high ability in order to ensure that the difference between the abilities of the copier individuals and the source individuals is greater (van der Linden & Sotaridona (2006)). Therefore, the individuals with low (-3, 0), medium (0.01, 1.50), and high (1.51, 3) abilities were randomly selected from the population (Sunbul & Yormaz, 2018). Low, medium, and high ability levels respectively include 20%, 60%, and 20% of the sample.

Copiers in the sample were randomly assigned among low ability individuals. The sources, who are individuals that the copiers copied their answers from, were randomly assigned among high ability individuals. In this study, we assigned only one copier for each source. Responses of the individuals, who are assigned as copiers, were manipulated so that their responses become similar to the sources' responses. Data simulation is repeated 100 times per each condition.

2.3. Analysis

The Kullback-Leibler divergence, one of the common methods, was utilized to detect copiers (Kullback & Leibler, 1951). KL reveals the difference between the two distributions, is calculated with the expression in the equation:

$$D(g||h) = \int_{-\infty}^{+\infty} g(x) \ln \frac{g(x)}{h(x)} dx \quad (1)$$

KL values were obtained by estimating the individuals' abilities twice before and after manipulation and comparing those two distributions. For the analysis, the 'irtoys' (Partchev, 2017) and 'LaplaceDemon' packages (Singmann, 2020) were used in the R program.

We used two methods to find the cut point for KL values. Firstly, to find the cut point, for every 100 iterations, the lowest KL values among the copiers were selected and created a new distribution of the lowest KL values. We repeated this process for each condition separately, and in the end, we obtained 12 distributions. We defined the cut point separately for each distribution based on the 0.05 alpha value (We call this approach as Min Score p-value). Secondly, ROC analysis (Swets & Pickett, 1982; Swets & Swets, 1979) was utilized for all KL values to define the cut point. ROC analysis can classify the data as binary or multi-category. In this study, data were classified as copier and non-copier based on the ROC curves. These curves are used to determine the relationship between sensitivity (Se) and specificity (Sp). The ROC curve is obtained by coordinates (1-Sp (c); Se (c)) for all possible cut points c; where Se (c) and Sp (c);

$$Se(c) = P(T_+|D = 1) = P(T \geq c|D = 1), \quad (2)$$

$$Sp(c) = P(T_-|D = 0) = P(T < c|D = 0). \quad (3)$$

In the formulas, the T values higher than the cut points mean that the individual is a copier. Sensitivity is the degree of defining a copier correctly. On the other hand, specificity is the degree of identifying a non-copier correctly. The ROC analysis presents a graph showing the specificity and the sensitivity (1-specificity) values in the x and y-axis and a curve regarding those values. The graph makes the interpretation easier. In the end, ROC analysis gives us the area under the ROC curve (AUC), which shows the correctness of cut points and the mean of all possible cut points. Thus, it is much more beneficial to evaluate all cut points considering AUC (Bamber, 1975; Swets, 1979). AUC values vary between 0.5 (non-informative) and 1 (excellent). However, ROC analysis offers several cut points criteria using assumptions based on sensitivity and specificity measures or functions defined as a linear combination of both measures. Besides, ROC curve criteria allow the selection of optimum cut points based on the risks and benefits of right and wrong decisions due to the classification outcome. We used some of these several cut points criteria. One of them is Youden Index, and the other is the Cost-Benefit method.

The Youden index (Youden, 1950) is one of the most common indicators used to evaluate the ROC curve. Youden index is the maximum difference between true positive and false positive rates (Krzanowski & Hand, 2009).

$$YI(c) = Se(c) + Sp(c) - 1 \quad (4)$$

The benefits and risks of each type of decision are combined with the prevalence of classification to find Se and 1-Sp values in the ROC curve; this provides the minimum average risk (maximum average benefit) in a given diagnosis (McNeill, Keeler, & Adelstein, 1975; Metz, 1978; Metz, Starr, Lusted & Rossmann, 1975; Swets & Swets, 1979). In a situation where there are two possible alternative decisions, the expected risk of classification use C can be expressed as follows:

$$C(c) = C_0 + C_{TP}p Se(c) + C_{TN}(1 - p) Sp(c) + C_{FP}(1 - p) (1 - Sp(c)) + C_{FN}p(1 - Se(c)) \quad (5)$$

C_{TP} , C_{TN} , C_{FP} , C_{FN} represent the average risks of the results from the decision type, and C_0 represents the overhead risk. We used the 'OptimalCutpoints' package (Raton-Lopez & Rodriguez-Alvarez, 2019) in R to compute cut points for KL values. In the end, we compute the power ratios of the cut points obtained.

3. RESULT / FINDINGS

3.1. Results

The cut point of KL values calculated under various conditions was calculated for the ROC analysis (Youden Index and Cost-Benefit) and the p-value of the minimum score (Min Score p-value). Table 2 shows the calculated cut points for different conditions.

It is observed that the cut points based on the Min Score p-value approach ranged from 0.00000000059 to 0.00000545898. For the Youden Index, the cut point obtained were in the range of 0.00000926385-0.00009678113. On the other hand, the cut points obtained with the Cost-Benefit approach varied between 0.00001011724 and 0.00035431080. The lowest cut point was obtained as 0.00000000059 in the Min Score p-value approach. (Sample size was 1000, copiers' rate 0.05, and copying percentage was 0.6. Table 3 presents the Power of KL

method to detect copiers for the cut points obtained by Youden Index, Cost-Benefit, and Min Score p-value approaches

Table 2. *Cut point of KL values of the table.*

Sample Size	Copiers' Rate	Copying Percentage	Min Score p-value	Youden Index	Cost-Benefit	
1000	0.05	0.1	0.00000305008	0.00002678292	0.00002918371	
		0.3	0.00000545898	0.00002854412	0.00003208120	
		0.6	0.00000000059	0.00003000205	0.00003379222	
	0.15	0.1	0.00000121844	0.00009678113	0.00034437004	
		0.3	0.00000000188	0.00006357387	0.00035431080	
		0.6	0.00000000380	0.00008453689	0.00034498847	
	3000	0.05	0.1	0.00000044474	0.00000986877	0.00001011724
			0.3	0.00000073166	0.00000926385	0.00001039373
			0.6	0.00000070757	0.00000973595	0.00001085132
0.15		0.1	0.00000049923	0.00002917059	0.00012223349	
		0.3	0.00000042948	0.00003221728	0.00011512144	
		0.6	0.00000037981	0.00002460582	0.00012614019	

When using the cut point obtained with the Youden Index, the power of detecting the copiers of the KL method was observed as the lowest 0.6311 under 1000 sample size, 0.15 copiers' rate, and 0.6 copying percentage conditions. On the other hand, the highest power (0.8328) was obtained under a 1000 sample size, 0.05 copiers' rate, and 0.3 copying percentage conditions.

Table 3. *Power of KL Methods Based on Cut Points Method.*

Sample Size	Copiers' Rate	Copying Percentage	Youden Index	Min Score p-value	Cost-Benefit
1000	0.05	0.1	0.8084	0.9221	0.7866
		0.3	0.8328	0.8980	0.8120
		0.6	0.8097	1.0000	0.7868
	0.15	0.1	0.6959	0.9441	0.3975
		0.3	0.7028	0.9964	0.3479
		0.6	0.6311	1.0000	0.2994
3000	0.05	0.1	0.8168	0.9547	0.8116
		0.3	0.8079	0.9325	0.7884
		0.6	0.8108	0.9306	0.7910
	0.15	0.1	0.7058	0.9496	0.3830
		0.3	0.7191	0.9533	0.4331
		0.6	0.7513	0.9588	0.4126

When using the cut points based on the Min Score p-value approach, the power of detecting the copiers of the KL method was observed as the lowest 0.8980 under 1000 sample size, 0.05 copiers rate, and 0.3 copying percentage conditions. On the other hand, the highest power (1.000) was obtained under a 1000 sample size and 0.6 copying percentage conditions. For the Cost-Benefit approach power of detecting the copiers of the KL method was observed as the lowest 0.2994 under 1000 sample size, 0.15 copiers' rate, and 0.6 copying percentage

conditions. On the other hand, the highest power (0.81) was obtained under 3000 sample size, 0.05 copiers' rate, and 0.1 copying percentage conditions. Moreover, comparing three methods to define cut points regarding all conditions Min Score p-value approach has the highest power rates while the Cost-Benefit approach has the lowest rates.

Figure 1. The Conditions' Interaction Effects for Power of KL Methods Based on Cut Points Methods.

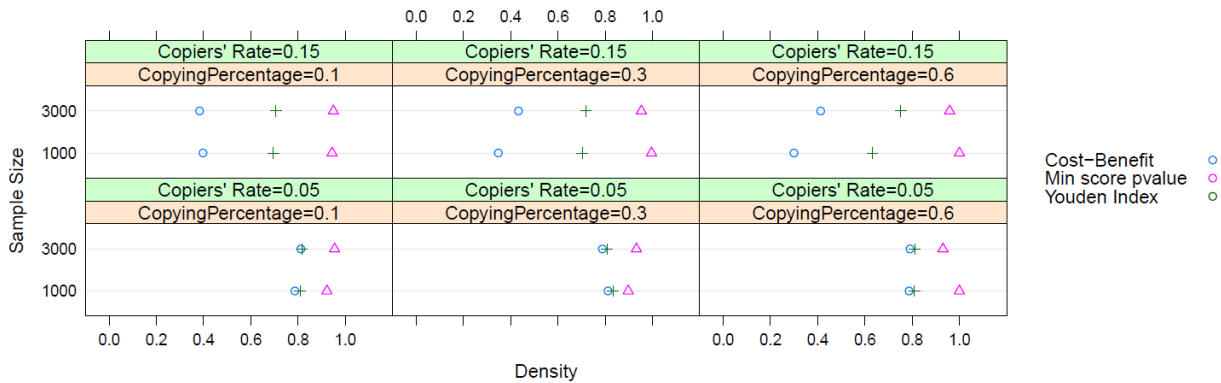


Figure 1 shows the interaction effect plot for the power of the KL method to detect the copier. Regarding the cut point obtained with the Min Score p-value approach, the KL method performed better than other approaches under all conditions. Youden Index method produced the second-best values, and the Cost-Benefit approach produced the worst values regarding the power of copy detection.

4. DISCUSSION and CONCLUSION

The Kullback-Leibler Divergence method was used to detect the copiers under various sample sizes, copiers' rates, and copying percentages. Cut points for the KL method were obtained using three approaches (Min Score p-value approach, Youden Index, Cost-Benefit approach). The power of the KL method was computed for the cut points obtained by three approaches. The findings were compared under the manipulated conditions (sample size, copiers' rate, and copying percentage).

Findings showed that the KL method's performance to detect copiers was higher under all conditions when the Min Score p-value approach was used. Especially in cases where the sample size was 1000, and the copying percentage was 0.60, the KL method correctly detected all the copiers. On the other hand, in the cases where the copiers' rate was 0.05 Youden Index and Cost-benefit approaches produced similar values.

Individuals are classified using the Cost-Benefit approach in clinical practice. There is a procedure to be performed for individuals diagnosed after classification. The Cost-Benefit approach determines cut points for this procedure to be both more useful and less cost outcome (Metz, 1978; Zou, et al., 2013). Because the procedure will be performed for each individual to be classified as false positive, otherwise it increases the cost. However, for the individual classified as a false negative, the procedure should not be applied because it will not provide a significant result. The study results revealed that the cut points obtained as a result of the analysis were higher than the cut points in other approaches to minimize the cost. When the difference between cheating individuals' distributions is less than the cut points, these individuals could not be identified as cheating individuals. Therefore, when the Cost-Benefit approach was used to define the cut point, negative relation was obtained between the copiers' rate and the KL method's power. In other words, the more copiers we had in the sample, the less power the KL method had to detect copiers. However, the copiers' rate did not affect KL methods' power when we used the cut point obtained by the Min Score p-value approach. When

the Min Score p-value approach was used to define the cut point, the KL method performed better in detecting the copiers.

When the difference between posterior ability distributions of individuals is high, the KL method with Min Score p-value approach performs better since it uses the minimum KL score of copiers in the computation process. On the other hand, in cases where there are no copiers in the sample, the Min Score p-value approach may detect individuals as copiers, although they are not (false positive). In other words, Min Score p-value Approach might inflate the type 1 error. The Youden index might perform better than the Min Score P-value approach to control the type 1 error.

In contrast to the Cost-Benefit approach's criteria, such as misclassification-cost and the minimum difference value as in the Min Score P-value approach, the Youden index displays a balancing approach. As can be seen from the findings, the cut points obtained according to the other two methods are located between both methods' cut points. In other words, the Youden index makes the classification in a balanced way by maximizing/minimizing a particular combination of sensitivity and specificity. Therefore, the cut points obtained with the Youden index is higher than the cut points obtained with the Min Score p-value approach (Raton-Lopez, et al., 2014). So, when we use the Min Score p-value approach to define the cut point, the KL method's power increases. The cost-Benefit approach decreases the type one error more than other methods do. In order to decide the cut point methods to be used, the researcher should consider the benefits and risks they will take after the decision (Lindahl & Danell, 2016).

Findings showed that the KL method's performance to detect copiers was higher under all conditions wthe hen Min Score p-value approach was used. To detect copiers with the KL method, cut scores are;

- minimum 0.00000000059 maximum 0.00000545898 based on Min Score p-value approach.
- minimum 0.00000926385 maximum 0.00009678113 based on Youden Index.
- minimum 0.00001011724 maximum 0.00035431080 based on Cost-Benefit approach.

In this study, we manipulated and the sample size, copiers' rate, copy percentage. Item difficulty parameters, sources, and copiers' abilities indexed are fixed. So different findings might be obtained when conditions are adjusted in different ways. The standard cut points of KL used by researchers are essential to detect copiers in tests developed in accordance with various measurement theories. Thus, by using various measurement theories, standard cut points of KL can be obtained from different simulation studies. In addition to various measurement theories, results for the type one error and power of KL are needed, when using standard cut points calculated for different values of α under various conditions (sample size, test length, measurement theories, ability distribution, etc.). When using the standard cut points calculated for different α values, new studies investigating the type one error and power of KL can be planned.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Arzu Uçar: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Celal Deha Doğan:** Methodology, Writing the original draft, Supervision, and Validation.

ORCID

Arzu Uçar  <https://orcid.org/0000-0002-0099-1348>

Celal Deha Doğan  <https://orcid.org/0000-0003-0683-1334>

5. REFERENCES

- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology*, 12, 387-415. [https://doi.org/10.1016/0022-2496\(75\)90001-2](https://doi.org/10.1016/0022-2496(75)90001-2)
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement*, 34, 379–392. <https://doi.org/10.1177/0146621610370453>
- Belov, D. (2011). Detection of Answer Copying Based on the Structure of a High-Stakes Test. *Applied Psychological Measurement*, 35(7), 495-517. <https://doi.org/10.1177/0146621611420705>
- Belov, D. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, 50, 141-163. <https://doi.org/10.1111/jedm.12008>
- Belov, D. (2014a). *Detection of Aberrant Answer Changes via Kullback–Leibler Divergence* (Report No. RR 14-04). Law School Admission Council.
- Belov, D. I. (2014b). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2, 37-58. <http://dx.doi.org/10.7333%2Fjcat.v2i0.36>
- Chalmers, P. (2020). Multidimensional item response model (mirt) [Computer software manual]. <https://cran.r-project.org/web/packages/mirt/mirt.pdf>
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229. <https://doi.org/10.1177/014662169602000303>
- He, Q., Meadows, M., & Black, B. (2018). *Statistical techniques for studying anomaly in test results: a review of literature* (Report No: Ofqual 6355-5). Office of Qualifications and Examinations Regulation.
- Hurtz, G., & Weiner, J. (2019). Analysis of test-taker profiles across a suite of statistical indices for detecting the presence and impact of cheating. *Journal of Applied Testing Technology*, 20(1), 1-15. <http://www.jattjournal.com/index.php/atp/article/view/140828>
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2
- Krzanowski, W., & Hand, D. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC Press.
- Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86. <https://www.jstor.org/stable/2236703>
- Lindahl, J., & Danell, R. (2016). The information value of early career productivity in mathematics: a ROC analysis of prediction errors in bibliometrically informed decision making. *Scientometrics*, 109, 2241-2262. <https://doi.org/10.1007/s11192-016-2097-9>
- Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N.M. Kingston & A.K. Clark (Eds.), *Test Fraud: Statistical Detection and Methodology* (pp. 52-80). Routledge Research in Education.
- McNeill, B., Keeler, E., & Adelstein, S. (1975). Primer on Certain Elements of Medical Decision Making, with Comments on Analysis ROC. *The New England Journal of Medicine*, 293, 211-215. https://www.researchgate.net/publication/22346698_Primer_on_Certain_Elements_of_Medical_Decision_Making

- Meijer, R., & Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement*, 25, 107-135. <https://doi.org/10.1177/01466210122031957>
- Meijer, R., & Tendeiro, J. (2014). *The use of person-fit scores in high stakes educational testing: How to use them and what they tell us*. (Report No. RR 14-03). Law School Admission Council.
- Metz, C. (1978). Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, 8, [https://doi.org/10.1016/S0001-2998\(78\)80014-2](https://doi.org/10.1016/S0001-2998(78)80014-2)
- Metz, C., Starr, S., Lusted, L., & Rossmann, K. (1975). Progress in Evaluation of Human Observer Visual Detection Performance Using the ROC Curve Approach. In C. Raynaud & A. E. Todd-Pokropek (Eds.), *Information processing in scintigraphy* (pp. 420-436). Orsay.
- Partchev, I. (2017). A collection of functions related to item response theory (irtoys) [Computer software manual]. <https://cran.r-project.org/web/packages/irtoys/irtoys.pdf>
- Raton-Lopez, M. & Rodriquez-Alvarez, X. M. (2019.). Computing optimal cut points in diagnostic tests (OptimalCutpoints) [Computer software manual]. <https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf>
- Raton-Lopez, M., Rodriquez-Alvarez, X. M., Suarez- Cadarso, C., & Sampedro-Gude, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cut points in Diagnostic Tests. *Journal of Statistical Software*, 61(8), 1-36. <https://www.jstatsoft.org/v061/i08>
- Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response. *Psychometrika*, 78, 481-497. <https://doi.org/10.1007/s11336-012-9311-3>
- Singmann, H. (2020). Complete Environment for Bayesian Inference (LaplaceDemon) [Computer software manual]. <https://cran.r-project.org/web/packages/LaplacesDemon/LaplacesDemon.pdf>
- Sotaridona, L., & Meijer, R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement*, 39(2), 115-132. <https://www.jstor.org/stable/1435251>
- Sotaridona, L., & Meijer, R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40(1), 53-70. <https://www.jstor.org/stable/1435054>
- Steinkamp, S. (2017). Identifying aberrant responding: Use of multiple measures [Doctoral dissertation]. https://conservancy.umn.edu/bitstream/handle/11299/188885/Steinkamp_umn_0130E_18212.pdf?sequence=1&isAllowed=y
- Sunbul, O., & Yormaz, S. (2018). Investigating the performance of omega index according to item parameters and ability levels. *Eurasian Journal of Educational Research*, 74, 207-226. https://ejer.com.tr/public/assets/catalogs/en/11_EJER_SYormaz.pdf
- Swets, J. (1979). ROC Analysis Applied to the Evaluation of Medical Imaging Techniques. *Investigative Radiology*, 14(2), 109-121.
- Swets, J., & Pickett, R. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. Academic Press.
- Swets, J., & Swets, J. (1976). ROC approach to cost/benefit analysis. In KL. Ripley & A. Murray (Eds.), *Proceedings of the Sixth IEEE Conference on Computer Applications in Radiology*. IEEE Computer Society Press.
- van der Linden, W., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3), 283-304. <https://www.jstor.org/stable/4122441>
- Voncken, L. (2014). *Comparison of the Lz^* Person-Fit Index and ω Copying-Index in Copying Detection*. (First Year Paper). Universiteit van Tilburg. <http://arno.uvt.nl/show.cgi?fid=135361>
- Wesolowsky, G. (2000). Detecting excessive similarity in answers on multiple choice exams.

- Journal of Applied Statistics*, 27(7), 909-921. <https://doi.org/10.1080/02664760050120588>
- Wollack, J. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307-320. <https://doi.org/10.1177/01466216970214002>
- Wollack, J. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40(3), 189–205. <https://www.jstor.org/stable/1435127>
- Wollack, J. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19(4), 265-288. https://doi.org/10.1207/s15324818ame1904_3
- Wollack, J., & Maynes, D. (2017). Detection of test collusion using cluster analysis. In G. Cizek & J. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 124-150). Routledge.
- Yormaz, S., & Sunbul, O. (2017). Determination of Type I Error Rates and Power of Answer Copying Indices under Various Conditions. *Educational Sciences: Theory & Practice*, 17(1), 5-26. <https://doi.org/10.12738/estp.2017.1.0105>
- Youden, W. (1950). Index for Rating Diagnostic Tests. *Cancer*, 3, 5-26. [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)
- Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement*, 40, 592–607. <https://doi.org/10.1177/01466216166664724>
- Zopluoglu, C., & Davenport, E. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72(6), 975-1000. <https://doi.org/10.1177/0013164412442941>
- Zou, K. H., Yu, C.-R., Liu, K., Carlsson, M. O., & Cabrera, J. (2013). Optimal Thresholds by Maximizing or Minimizing Various Metrics via ROC-Type Analysis. *Academic Radiology*, 20(7), 807–815. <https://doi.org/10.1016/j.acra.2013.02.004>

The Problem of Measurement Equivalence or Invariance in Instruments

Tulin Otbicer-Acar ^{1,*}

¹Parantez Education Research Consultancy&Publishing, Ankara, Turkey

ARTICLE HISTORY

Received: Feb. 18, 2020

Revised: Nov. 15, 2020

Accepted: Jan. 27, 2021

Keywords:

Measurement equivalence,
Measurement invariance,
Cross-validation.

Abstract: The purpose of this study is to discuss the validity of equivalence in the sample groups of young and adult; females and males in the scale of assessing the attitudes towards foreign language skills and to offer the researchers that will use this scale certain evidence based on data. No measurement equivalence/invariance was found in adult and young groups. Consequently, measurement equivalence / invariance based on gender variable was not present, either. The absence of measurement equivalence/invariance is in fact a fundamental proof that the measurement instrument is specific to the group that it is intended for. For this reason, researchers should evaluate cross-validity or multi-group analyses on the basis of the traits that are measured using the measurement instrument. It is not always negative not to have measurement equivalence/invariance during the process of gathering validity evidences.

1. INTRODUCTION

Numerous instruments of measurement have been developed by researchers to measure the psychological structures of individuals, such as interest, attitude, success, anxiety, and motivation. A measurement instrument is sometimes considered within the scope of adaptation studies. Developing or adapting an instrument is a time consuming and rigorous process in which whether the measurement instrument is capable of measuring the same conceptual structure in different groups and cultures signifies the validity of the instrument. In validity studies, it is desirable for the structure that is being measured under measurement by the instrument to be invariant and unbiased. When the measurements vary among the subgroups of the populations that are measured or among different populations, there is a certain amount of *bias*. The potential for bias in test items is the most significant element. They arise from systematic errors. Also, other sources should be taken into consideration for the validity of the instruments of measurement. The sources of bias are studied under the categories of construct bias, method bias (namely sample bias, administration bias, and instrument bias) and, item bias (Vijver & Tanzer, 2004).

Item bias is typically referred to as Differential Item Functioning (DIF). However, educational experts, test developers make a difference between the concept of item bias and DIF. The concept of item bias has a negative meaning in everyday life and it is associated with a negative idea. The difference between technical use of item bias and the everyday use of it is uncertain.

CONTACT: Tulin Otbicer Acar ✉ totbicer@gmail.com 📠 Parantez Education Research Consultancy Publishing, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

The conceptual difference between DIF and item bias is as follows (Hambleton et al., 1991, p.109):

Investigation of bias involve gathering empirical evidence concerning the relative performances on the test item of members of minority group of interest and members of the group that represents the majoriy. Empirical evidence of differential performance is necessary, but not sufficient, to draw the conclusion that bias is present; this conclusion involves an inference that goes beyond the data. To distinguish the empirical evidence from the conclusion, the term differential item functioning (DIF) rather than bias is used commonly to described the empirical evidence obtained in investigations of bias.

It is understood that item response theory (IRT) and structural equation modeling (SEM) are used in the studies that are intended to determine the systematic errors that interfere in the results of measurement. The traits measured in these two methods are defined as latent traits. According to the IRT, item bias is determined by DIF. DIF is a function that is used to determine whether the probability of responding an item differs among subgroups in each skill level of the psychological structure that is to be measured by an item (Lord, 1980; Embretson & Reise, 2000). Likelihood Ratio according to IRT (Thissen et al., 1988), Lord's chi-square test (Lord, 1980), and Raju's area measures (Raju, 1988) are among the techniques that are used in the literature to determine DIF. In addition, there are techniques of DIF determination such as Mantel-Haenszel, Logistic Regression, and SIBTEST in the classical test theory based on observed scores in metrology (Gierl et al., 1999).

A different approach, according to DIF techniques in IRT, is measurement equivalence/invariance. In the literature, the term 'measurement invariance' is usually used as the synonym of measurement equivalence (Davidov et al., 2014). Wadenberg and Lance (2000, p.5) stated that "measurement equivalence-ME (or alternately, measurement invariance-MI) across populations". In addition, measurement equivalence is also called structural equivalence (Kankaraš & Moors, 2010). Measurement equivalence denotes similarity of observed variables and latent structures among groups (Koh & Zumbo, 2008). A method based on covariance structures is used in measurement equivalence research (French & Finch, 2008). This covariance-based method is resolved by SEM analyses. Studying multiple group equivalence by SEM method corresponds to the concept of measurement method. According to the definition made by Byrne (2008), measurement equivalence or invariance (ME/I) implies that the measurement instrument has the same psychological meaning and theoretical structure in the groups of interest. It is an approach that is based on restriction of structural parameters (factor loadings, observed variable error variances, error covariances) produced by multiple group invariance – an extension of Confirmatory Factor Analysis – CFA). This approach is associated with measurement equivalence/invariance. Two types of techniques are used for measurement equivalence/invariance in SEM. The first one is Multi-Group Confirmatory Factor Analysis (MGCFA, see Jöreskog, 1971; Cheung & Rensvold, 2002) where the equivalence of covariance structures is tested. The second one is Mean and Covariance Structure (MACS, see Byrne et al., 1989; Little, 2010; Yuan & Bentler, 1997) where the equivalence of mean and covariance structures is tested. Both MGCFA and MACS are cross-validation techniques. These analyzes are resolved by SEM. MACS analysis is used to assess differences between group in terms of the constructs' mean, variances and covariances. MGCFA tests the invariance of estimated parameters across groups.

There are numerous studies focusing on whether several instruments of measurement that measure different psychological characteristics ensure measurement equivalence/invariance in different subgroups (see Akyıldız, 2009; Asil & Gelbal, 2012; Baumgartner & Steenkamp, 1998; Lomax, 1983; Mullen, 1995; Önen, 2007; Uyar & Doğan, 2014; Yoo, 2002). It is observed that these studies offer a comparison of the models that are made up of restricted

parameters. The steps of the measurement equivalence/invariance that shows a series of progressivity depending on the number of restricted parameters are as follows (Byrne & Stewart, 2006):

Model 1 – Configural invariance: The first stage. Factor loads, regression constants and error variances are released among groups. However, the number of factors, and the factor loading pattern are defined similarly among groups. Therefore, structural invariance is ensured among groups. Measurement of the same structure is measured among groups.

Model 2 – Weak factorial invariance or Metric invariance: Factor loads are restricted in addition to the first stage. When metric invariance is not ensured, the items in the groups are not considered to be interpreted at the same level. Factor loads correspond to the Discrimination Parameter, and non-uniform DIF of factor load is present among groups (Steinmetz et al., 2009).

Model 3 – Strong factorial invariance or scalar invariance: This is the stage where the regression constant equates between groups. On the other hand, for straightforward interpretation of latent means and correlations across groups, both the factor loadings and intercepts should be the same across groups (Van de Schoot, Lugtig & Hox, 2012, p.490). Variance of regression constant among groups signifies presence of uniform DIF on the items and means that scalar invariance is not present (Kankaraš & Moors, 2010).

Model 4 – Strict factorial invariance: At this stage, critical error variances have been restricted as well. In this model, the error variances of the second group stabilize on the error variances of the first group.

Cheung and Rensvold (2002, p.236) stated that “the statistic for testing the hypothesis is the difference between the fit of the constrained model and that of a less constrained model. Many fit indices are obtained for each of the four models mentioned above. The most frequently used criterion is that the difference between the values of RMSEA and CFI – fit indices – in comparison of models is smaller than 0.01 (Byrne & Stewart, 2006; Hirschfeld & Brachel, 2014). Since RMSEA, CFI, and SRMR are not affected by the sample size of fit indices, these indices are suggested to be taken into consideration in comparing these models (Hu & Bentler, 1999). Similarly, the chi-square difference between the two models, the insignificance of the chi-square difference test, and the difference between the degrees of freedom are considered as an indication of the invariance between the models. Byrne and Stewart (2006, p.305) noted that “ $\Delta\chi^2$ value is as sensitive to sample size and nonnormality as the chi-square statistic itself, thereby rendering it an impractical and unrealistic criterion on which to base evidence of invariance.”

An examination of the literature reveals that multi-group analyzes are also called cross-validation (Fiala et al., 2002; Gandek et al., 1998). It is obvious that these techniques provide extra data in gathering data for validity. However, it should be noted that reducing the sample size makes a major disadvantage in cross-validation or multi-group studies. Varoquaux (2018, p.68) stated that “the shortcoming of small samples are more stringent and inherent as they are not offset with large effect sizes”.

1.1. Aim of the Study

Sample characteristics of subjects must be taken into consideration for the future usage of the same scale. Otherwise, measurements errors are likely to appear. So, the scope of present study is to discuss the validity of equivalence in the sample groups of young and adult, and females and males in the scale of assessing the attitudes towards foreign language skills and to offer the researchers that will use this scale certain evidence based on data. It should be noted that this study was carried out for evidence of measurement equivalence/invariance for the scale developed. The empirical outcomes of this study will make important contributions to both psychological test developers and psychometrists.

2. METHOD

2.1. Research Design

In this research, measurement equivalence/invariance was investigated for gender and two groups. Thus, present research is a descriptive research. Descriptive research is “current events and that the research questions and problems raised are based on the appreciation of the present phenomena, events, or state of affairs” (Ariola, 2006, p.47). The scale developed for 15-16 year old people cannot be applied to the scale developed for 18-60 year old people without being tested and applied. Like age variable, gender variable plays a significant role in measurement equivalence/invariance due to the fact that gender difference embraces both biologic and cultural implications.

2.2. The Characteristics of Participants

The researcher collected data on from 563 participants to test the equivalence in the group of adults aged 18 to 60 in the scale which was developed for the student groups of 15-16 years of age for determining attitudes towards foreign language skills. 15-16-year-old students were high school students who continued secondary education. Therefore, this group of students was named *young* in this study.

The scale was administered to the participants in Turkey. Some of the participants are employed and some are out of employment. They belong to various occupational groups such as academicians, entrepreneurs and business people. The scale was administered online and in printed form. Missing values were excluded from the data set. Therefore, the data set includes 481 participants – 275 young students (57.2%) and 206 adults (42.8%). The frequency of gender distributions by groups and the result of the chi-square test is given in [Table 1](#).

Table 1. Gender distribution by groups.

			Group		Total
			Young	Adult	
Gender	Female	Count	136	109	245
		%	55.5%	44.5%	100.0%
	Male	Count	139	97	236
		%	58.9%	41.1%	100,0%
Total	Count	275	206	481	
	%	57.2%	42.8%	100.0%	

$\chi^2=0.564$ Sig.=0.453

55.5% of the female participants are young and 44.5% are adults. 58.9% of male participants are young and 41.1% are adults. 49.5% of the young are females. 52.9% of the adults are females. Statistically, no significant difference was found between the gender distributions of the individuals according to the groups ($p>0.05$). For both young and adult groups, according to the gender of the participants, the difference between age averages was tested by t test for independent samples. The results are given in [Table 2](#).

Table 2. Results of the difference between age averages according to the gender.

Group		N	Mean	Std. Deviation	t value	df	p
Young	Female	136	15.54	0.50	.073	273	.942
	Male	139	15.53	0.50			
	Total	275	15.53	0.50			
Adult	Female	106	27.34	7.14	-.279	197	.780
	Male	93	27.63	7.74			
	Total	199	27.48	7.41			

The average age of the young group is 16-years old. The average age of the adult's group is 28-years old. Statistically, there was no significant difference between the average age of male and female young ($p>0.05$). Also, there was no significant difference between the mean age of male and female adults ($p>0.05$).

2.3. Instrument

The developed scale comprises 29 items that are structured on a 5-point scale ranging from 1 to 5. The original purpose of the scale was to identify the attitudes of 15- and 16-year-old students towards Foreign Language Skills. For 15-16-years old students, the reliability-validity analyses of the development process of the scale are available in the reference Acar (2016). The implementation scale is given in the [Table A1](#). The scale has 4 sub-factors: reading, writing, speaking, and listening. In this study (for 481 participants), the internal consistency of the scale's Cronbach Alpha reliability is 0.923 for the adult group; 0.900 for the young group; 0.922 for the females' group and 0.899 for the males' group. Sub-scales reliabilities were showed in [Table 3](#). It is observed that the internal consistency of the subscales is at appropriate values.

Table 3. Cronbach's Alpha coefficients.

	Sub-Groups			
	Young	Adult	Female	Male
Reading	0.768	0.729	0.782	0.724
Writting	0.758	0.756	0.783	0.744
Speaking	0.758	0.623	0.722	0.692
Listening	0.804	0.786	0.789	0.793

Item-total correlations are shown in [Table A2](#). The variation between 0.140 and 0.655 in the subscales of item total correlations was measured. No item was removed in this study, although the number of items in the subscales was relatively low. Due to the fact that the purpose of the research is the measurement invariance of the instruments. In addition, the reliability and validity of the scale were tested in another sample, too.

2.4. Data Analysis

For measurement equivalence/invariance, all procedures were based on the analysis of MACS within the framework of CFA modeling. The LISREL (Jöreskog&Sörbom, 2003) program was used for all analyses. First of all, the dataset was completely cleared of missing values. It was observed that item scores ranged from 1 to 5, and there were no extreme values. Through confirmatory factor analysis, four sub-dimensional scale was tested for the all data before multi-group CFA was carried out. The multivariate assumption of normality was not met. Because, Mardia's measure of multivariate skewness and kurtosis was not found significant ($\chi^2=2664.719$ $p<0.000$). Thus, the observed scores of scale items were converted into normal scores in LISREL. Estimations of parameters were carried out through maximum likelihood. Asymptotic covariance matrix was used for parameter estimations. Fit indices was presented [Table 4](#).

Root mean square approximation error was calculated as (RMSEA) = 0.074. Van de Schoot, Lugtig and Hox (2012, p.488) stated that "the RMSEA is insensitive to sample size, but sensitive to model complexity". Bialosiewicz, Murphy, and Berry (2013) pointed out that an RMSEA around 0.10 is acceptable. Standardized root mean square residual was calculated as (S-RMR) = 0.068; comparative fit index was calculated as (CFI) = 0.93; normed fit index was calculated as (NFI) = 0.91 and relative fit index was calculated as (RFI) = 0.90. Chi-square statistics of the similarity rate was calculated as χ^2 (371) = 1339.65 $p<0.01$ and χ^2 / df is 3.61.

Table 4. Goodness of fit indices.

Goodness of fit indices	Cut off value *	Values
χ^2/df	<5 Moderate <3 Perfect fit	1339.65/371= 3.61
GFI	>0.90	0.79
CFI	>0.90	0.93
NFI	>0.90	0.91
RFI	>0.85	0.90
S-RMR	< 0.08	0.068
RMSEA	< 0.08	0.074

* Resources: Kline, 2011; Bentler, 1980

Goodness-of-fit index was calculated as (GFI)= 0.79 and only this index was found below 0.90. GFI involve terms that adjust for degrees of freedom. Thus, GFI is highly dependent on sample size. In addition, Cheung and Rensvold (2002) showed that number of items per factor and number of factors in the model affect GFI values. Bollen and Long (1983) pointed out, "the test statistics and fit indices are very beneficial, they are no replacement for sound judgment and substantive expertise". It was observed that 4-factor structure attitude scale concerning the English language skills was acceptable according to the standard criteria. Baumgartner and Homburg (1996, p.153) suggest that general rules of thumb (e.g., that GFI be greater than 0.9) may be misleading because they ignore such contingencies. χ^2/df and RMSEA seem to be effective in controlling for model complexity by assessing fit per degree of freedom. t values indicating the significance of the relationship between the items and the latent variable are presented in the [Figure A1](#).

3. FINDINGS

In the invariance studies, the RMSEA value is not interpreted alone. According to, literature for comparison of the four models, difference values or difference tests (for example $\Delta\chi^2$, Δ GFI, Δ CFI, Δ TLI, Δ BBI or Δ RMSEA) are used. Rijkeboer and van den Bergh (2006) suggested the use of Chi-Square difference test which is the most efficient one with respect to both goodness-of-fit and parsimony. The choice of difference tests remains at the expertise the researcher. The dataset was divided in two groups – namely, females and males, then the measurement equivalence/invariance of the scale for determining the attitudes towards foreign language skills was tested on the basis of the gender variable, and the results of the fit indices were specified in [Table 5](#).

Table 5. Measurement equivalence/invariance based on gender variable.

Models	χ^2	df	RMSEA	CFI	Δ CFI	Δ RMSEA	$\Delta\chi^2$	Δ df	p
1: Configural invariance	2527.59	781	0.097	0.92	-	-	-	-	-
2: Metric invariance	2579.36	806	0.096	0.92	0.00	-0.001	51.77	25	0.001
3: Scalar invariance	2731.79	834	0.097	0.91	-0.01	0.001	152.43	28	0.000
4: Strict factorial invariance	2759.81	835	0.098	0.91	0.00	0.001	28.02	1	0.000

When comparing *Model 2* versus *Model 1*, Cheung and Rensvold (2002, p.251) pointed out “a value of smaller than or equal to -0.01 indicates that the null hypothesis of invariance should not be rejected”. A comparison of *Model 1 - Model 2*, *Model 2 - Model 3*, and *Model 3 - Model 4* reveal that Δ RMSEA and Δ CFI values were in appropriate ranges. However, p value of the chi-square difference test was found to be significant. It is seen that $\Delta\chi^2$, Δ CFI, and Δ RMSEA values provide different interpretations. In this study, final comments are made according to $\Delta\chi^2$ values. It was observed that metric, scalar, and strict factorial invariances could not be

ensured in the multi-group analysis based on the gender variable. At this stage, it was suggested to test whether there are any items that contain uniform and non-uniform DIF.

According to the system of progressivity, it is not significant to skip to the next stage when a stage is not appropriate. It is observed that in certain studies, partial invariance models are attempted to be used where invariance cannot be ensured (Murayama et al., 2009; Milfont & Fischer, 2010). However, partial invariance models were not used in this study. Measurement equivalence/invariance of the scale of determining the attitudes towards foreign language skills in young and adult groups and the results of the fit indices are given in [Table 6](#).

Table 6. Measurement equivalence/invariance based on the group variable.

Models	χ^2	df	RMSEA	CFI	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$	Δdf	p
1: Configural invariance	2502.04	781	0.096	0.92	-	-	-	-	-
2: Metric invariance	2565.62	806	0.095	0.92	0.00	-0.001	63.58	25	0.000
3: Scalar invariance	2459.67	835	0.103	0.91	-0.01	0.008	105.95	29	0.000
4: Strict factorial invariance	2459.67	835	0.103	0.91	0.00	0.000			

A comparison of Model 1 and Model 2 reveals that $\Delta RMSEA$ and ΔCFI values didn't not exceed the 0.01 threshold. However according to chi-square difference test ($\Delta\chi^2$), metric invariance was not ensured for factor number, factor loading pattern, and factor loads among young and adult groups for the scale for determining the attitudes towards foreign language skills. It is seen that $\Delta\chi^2$, ΔCFI , and $\Delta RMSEA$ values provide different interpretations. In this study, final comments are made according to $\Delta\chi^2$ values. Therefore, it was found that certain items in young and adult groups may be biased. This result offers a clue in identifying the items that contain uniform DIF. Since metric invariance was not ensured it was understood that factors do not mean the same in different groups.

When *Model 2* was compared with *Model 3*, it was revealed that $\Delta RMSEA$ and ΔCFI values didn't not exceed the threshold. However, according to chi-square difference test, scalar invariance was not ensured for factor number, factor loading pattern, factor loads, and regression constants among young and adult groups for the scale for determining the attitudes towards foreign language skills. Therefore, it was found again that certain items in young and adult groups may be biased. These results offer clues on identifying the items that contain non-uniform DIF. In other words, the mean values of latent structures vary among the groups. It is not appropriate to make a comparison between the means of young and adults.

According to chi-square difference test, scale item equivalence could not be ensured on the basis of groups. The results of the discriminant analysis were used to decide which group the developed scale is appropriate for. The correct classification ratio, equality of covariance matrices, and log determinant tables were evaluated according to the discriminant analysis results. According to the discriminant analysis, the correct classification ratio of original and predicted group memberships was 81.1% for young and 66% for adults. Also, 74.6% of original grouped cases correctly classified. The results indicate a higher classification consistence for the young group. An examination of Box'M results in the equation of covariance matrices leads to rejection of the equation of covariance matrices in young and adult groups ($F(2; 592459, 45) = 833.362$ sig=0.000). Log Determinant values are given in [Table 7](#).

Table 7. Log determinants.

Group	Rank	Log Determinant
Young	29	-1.781
Adult	29	-7.182
Pooled within-groups	29	-2.353

In the multi-group model, log determinant values provide an indication of which groups' covariance matrices differ most. For each group, its log determinant is the product of the eigenvalues of its within group covariance matrix. In this research, the log determinant value for adult group is very small relative to that of the young group. Therefore, it is fair to say that scale items are suitable for the young group that was developed initially.

4. CONCLUSION and RECOMMENDATIONS

In the study, analysis of MACS was used to test for measurement invariance of the scale items across group and gender variables. $\Delta\chi^2$, ΔCFI , and $\Delta RMSEA$ values provided different interpretations. In this study, final comments were made according to chi-square difference test values. No measurement equivalence/invariance was found in adult and young groups. Consequently, measurement equivalence/invariance based on gender variable was not presented, either. Female and male datasets include adult and young as well. For this reason, it is a predictable result that measurement equivalence/invariance is absent for groups, and that the measurement equivalence/invariance based on the gender of the same individuals is not in compliance. The finding bears similarity to the finding of Feingold (1992) who emphasized that cognitive abilities arise from gender differences.

Little (2010, p.53) said that "the nature of sociocultural differences and similarities can be confidently and meaningfully tested among the constructs' moments in each sociocultural sample". But in this study, the measurement equivalence/invariance of the scale in different cultures was not tested. Since the scale was intended to measure the attitudes of 15-16-year-old Turkish students towards foreign language skills, it is restricted with the psychological traits of Turkish students. It is considered that the reasons for the absence of measurement equivalence/invariance in the scale include different interests, motivations, and attitudes towards foreign language skills among adult and young groups. Since the young group is made up of individuals who receive formal education, it is quite likely that they have different perceptions of foreign language skills compared to adults. Students' respective success in English courses is considered to have an impact on their attitude to foreign language skills. On the other hand, adults' perspective of foreign language skills is generally influenced by their occupational development, because they are not in formal education anymore due to their age.

Metric and scalar invariance was not present based on groups of adults and young, and on genders. There is evidence of the presence of uniform and non-uniform DIF items. However, a detailed study on DIF was not conducted due to the purpose of this study. The measurement instrument may be redesigned later. Certain items may be added, removed, or modified depending on the psychological traits of the implementation group. Equivalence trait of the measurement instrument may be abandoned in different groups. In this respect, the scale may be used for the target group for which it was originally intended.

It is not always negative not to have measurement equivalence/invariance during the process of gathering validity evidence. The absence of measurement equivalence/invariance is, in fact, a fundamental proof that the measurement instrument is specific to the group that it is intended for. For this reason, researchers should evaluate cross-validity or multi-group analyses on the basis of the traits that are measured using the measurement instrument. The validity of the instruments is the evidence gathering process. An ad-hoc measurement instrument should not be developed or used. It is recommended that any kind of information be used in gathering evidence and data for examination of the instruments of measurement.

Scale developing is a process. The most important stage of this process is ensuring the validity of the measurement instrument. Validity analysis should be examined through different techniques. In this process, items can be regulated. The qualifications of the application group may vary. Even the application area of the scales may expand.

Acknowledgments

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests and Ethics

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

Authorship contribution statement

Tulin Otbicer Acar: All chapters are written by the author.

ORCID

Tülin Otbicer Acar  <https://orcid.org/0000-0001-7976-5521>

5. REFERENCES

- Acar, T. (2016). Measurement of attitudes regarding foreign language skills and its relation with success. *International Journal of Evaluation and Research in Education*, 5(4), 310-322. <https://doi.org/10.11591/ijere.v5i4.5959>
- Akyıldız, D. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 6(1), 18-47. <https://dergipark.org.tr/tr/pub/yyu/efd/issue/13711/165993>
- Ariola, M. M. (2006). *Principles and methods of research*. Rex Book Store.
- Asil, M., & Gelbal, S. (2012). Crosscultural equivalence of the PISA student questionnaire. *Education and Science*, 37, 236-249.
- Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13(2), 139-161.
- Baumgartner, H., & Steenkamp, J.-B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters*, 9, 21-35. <https://doi.org/10.1023/A:1007911903032>
- Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31(1), 419-456.
- Bialosiewicz, S., Murphy, K., & Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants*. <http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8>
- Bollen, K. A., & Long, J. S. (Eds.). (1983). *Testing structural equation models*. Sage.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466. <https://doi.org/10.1037/0033-2909.105.3.456>
- Byrne, B. M. (2008) Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, 20, 872-882.
- Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: the_macs approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, 13(2), 287-321. https://doi.org/10.1207/s15328007sem1302_7
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40(1) 55-75. <https://doi.org/10.1146/annurev-soc-071913-043137>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.
- Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, 111(2), 304–341. <https://doi.org/10.1037/0033-2909.111.2.304>
- Fiala, W. E., Bjorck, J. P., & Gorsuch, R. (2002). The religious support scale: construction, validation, and cross-validation. *American Journal of community Psychology*, 30, 761-786. <https://doi.org/10.1023/A:1020264718397>
- French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 96-113. <https://doi.org/10.1080/10705510701758349>
- Gandek, B., Ware J. E., Aaronson N. K., Apolone, G. B., Brazier J. E., et al. (1998). Cross validation of item selection and scoring for the SF-12 health survey in nine countries: results from the iqola project, international quality of life assessment. *Journal of Clinical Epidemiology*, 51(11), 1171-1180. [https://doi.org/10.1016/S0895-4356\(98\)00109-7](https://doi.org/10.1016/S0895-4356(98)00109-7)
- Gierl, M., Khaliq, S. N., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: prevalence and policy implications*. Paper Presented at the Symposium entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education, Canada
- Hambleton, R. K., Swaminathan, H., & J. H. Rogers. (1991). *Fundamentals of Item Response Theory*. Sage Publications.
- Hirschfeld, G., & Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R-A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, 19(7), 1-12.
- 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. <https://doi.org/10.1080/10705519909540118>
- Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409-426. <https://doi.org/10.1007/BF02291366>
- Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross cultural studies. *Psihologija*, 43(2), 121-136. <https://doi.org/10.2298/PSI1002121K>
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. The Guilford Press.
- Koh, K., & Zumbo, B. D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods*, 7(2), 471-477. <https://doi.org/10.22237/jmasm/1225512660>
- Little, T.D. (2010). Mean and covariance structures (MACS) analyses of crosscultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32(1), 53-76. https://doi.org/10.1207/s15327906mbr3201_3
- Lomax, R. G. (1983). A guide to multiple-sample structural equation modeling. *Behavior Research Methods & Instrumentation*, 15, 580-584. <https://doi.org/10.3758/BF03203726>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. <https://doi.org/10.4324/9780203056615>
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111-130. <https://doi.org/10.21500/20112084.857>

- Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26, 573-596. <https://doi.org/10.1057/palgrave.jibs.8490187>
- Murayama, K., Zhou, M., & Nesbit, J. C. (2009) A cross-cultural examination of the psychometric properties of responses to the Achievement Goal Questionnaire. *Educational and Psychological Measurement*, 69(2), 266-286. <https://doi.org/10.1177/013164408322017>
- Önen, E. (2007). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin incelenmesi: Epistemolojik inançlar envanteri üzerine bir çalışma [Examination of measurement invariance at groups' comparisons: A study on epistemological beliefs inventory]. *Ege Eğitim Dergisi*, 8(2), 87-109. <https://dergipark.org.tr/tr/pub/eggeefd/issue/4913/67270>
- Raju, N. S. (1988). The area between two item response functions. *Psychometrika*, 53, 495-502. <https://doi.org/10.1007/BF02294403>
- Rijkeboer, M. M., & van den Bergh, H. (2006). Multiple group confirmatory factor analysis of the young schema-questionnaire in a Dutch clinical versus non-clinical population. *Cogn. Ther. Res.*, 30, 263–278. <https://doi.org/10.1007/s10608-006-9051-8>
- Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality and Quantity*, 42, 599-616. <https://doi.org/10.1007/s11135-007-9143-x>
- Thissen, D., Steinberg, L., & Wainer, H (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 147-169). Lawrence Erlbaum Associates, Inc.
- Uyar, Ş., & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA turkey sample]. *Uluslararası Türk Eğitim Bilimleri Dergisi*, 2, 30-43.
- Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 180, 68–77.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012) A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486-492. <https://doi.org/10.1080/17405629.2012.686740>
- Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée*, 54(2), 119-135. <https://doi.org/10.1016/j.erap.2003.12.004>
- Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology & Marketing*, 19(4), 357-368. <https://doi.org/10.1002/mar.10014>
- Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, 92(438), 767-774. <https://doi.org/10.1080/01621459.1997.10474029>

6. APPENDIX

Table A1. Form of the Attitude Scale Regarding English Language Skills.

		Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
1	I can answer the questions asked, after listening to a dialogue.					
2	I listen to a tourist if I encounter one.					
3	I look up the words in the dictionary, whose English meanings I don't know.					
4	I make an effort to watch an English movie or listen to English news or music.					
5*	I'm anxious about writing a letter, petition or resume in English.					
6	When I listen to a text or music in English, I make an effort to understand its meaning.					
7*	Writing in English in English exams, makes me anxious.					
8*	I close the English pages I encounter while making a search in the search engines.					
9*	I get bored with English listening activities.					
10	Speaking English, increases my self-confidence.					
11*	Speaking English, makes me anxious.					
12	I like reading English story books.					
13	I read a lot, in order to learn English words.					
14*	It is boring for me to listen to someone speaking English.					
15	I care about summarizing what I've heard in English, and writing them correctly.					
16*	I immediately walk away when I see someone speaking English.					
17*	I don't prefer having foreign friends to speak English with.					
18	I enjoy speaking English.					
19	I'd like to be a listener in a conference where English is spoken.					
20*	Reading and perceiving what is written in English, does not take an important place in my daily life.					
21*	I can't express my opinions easily while writing an English text.					
22*	Writing in English, is not important in daily life.					
23	I'd like the English reading activities to be more.					
24	I do not hesitate from answering the questions asked in English.					
25	I pay attention to the grammar rules while writing in English.					
26*	It is not important to speak English fluently.					
27*	I don't like reading equipment manuals that are written in English.					
28	I can write an English text about myself.					
29	I try to speak in accordance with the grammar rules.					

*Reverse coded items.

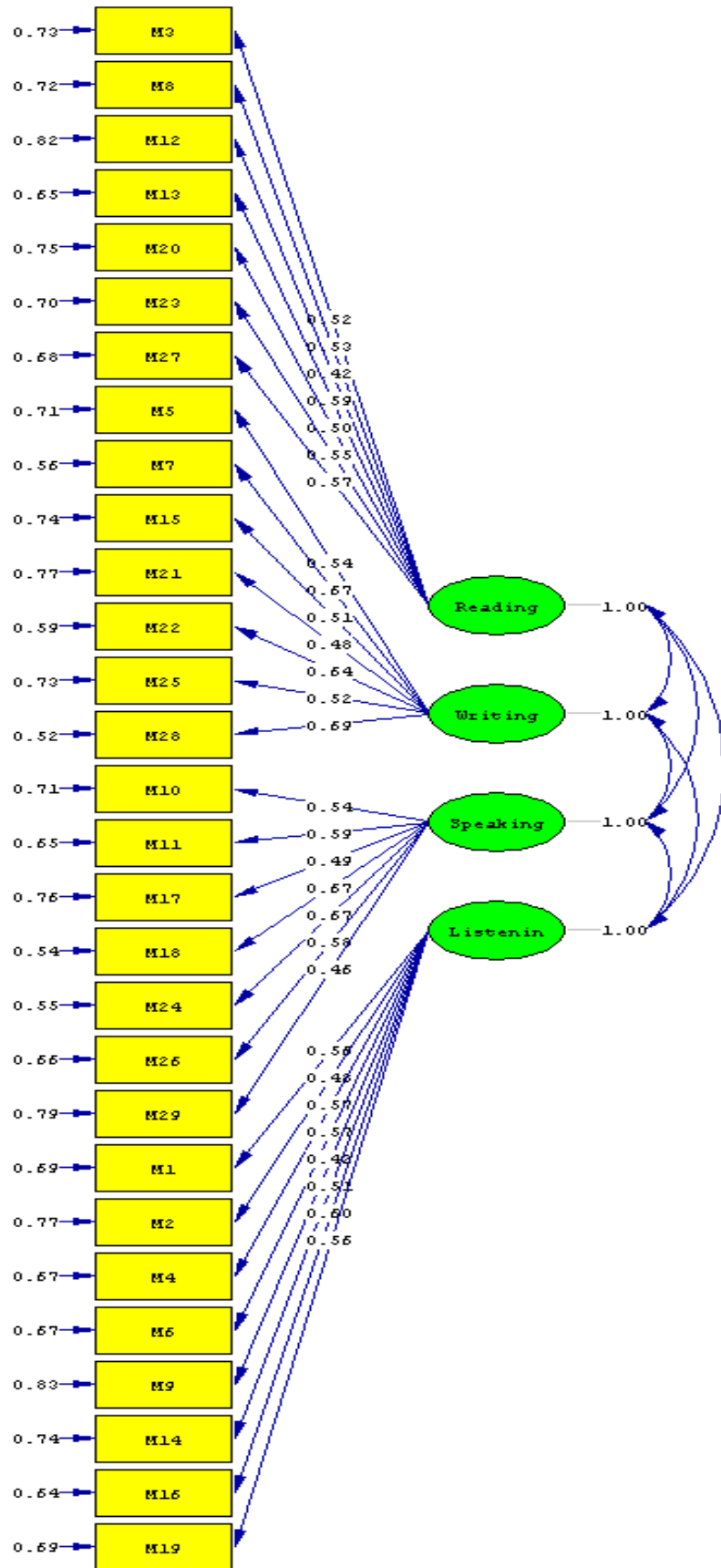
Attitude items related to the reading skill=3,8,12,13,20,23,27

Attitude items related to the writing skill =5,7,15,21,22,25,28

Attitude items related to the speaking skill =10,11,17,18,24,26,29

Attitude items related to the listening skill =1,2,4,6,9,14,16,19

Figure A1. The Path Diagram which is Factor Load per Item for All Dataset.



Chi-Square=1339.65, df=371, P-value=0.00000, RMSEA=0.074

Table A2. *Item-Total Correlations.*

Sub-scales	Item No	Young Group	Adult Group	Female Group	Male Group
Reading	m3	.395	.348	.427	.341
	m8	.470	.428	.486	.383
	m12	.613	.633	.655	.562
	m13	.585	.531	.567	.580
	m20	.386	.398	.409	.367
	m23	.530	.330	.428	.508
	m27	.436	.425	.573	.307
Writing	m5	.406	.640	.504	.548
	m7	.498	.647	.606	.568
	m15	.476	.290	.475	.348
	m21	.457	.601	.573	.503
	m22	.451	.313	.413	.361
	m25	.537	.381	.420	.476
	m28	.511	.456	.562	.405
Speaking	m10	.535	.334	.493	.408
	m11	.455	.339	.422	.413
	m17	.437	.317	.345	.383
	m18	.577	.611	.592	.572
	m24	.466	.447	.543	.414
	m26	.431	.140	.296	.311
	m29	.434	.214	.361	.318
Listening	m1	.390	.375	.457	.322
	m2	.503	.409	.457	.489
	m4	.551	.538	.566	.511
	m6	.541	.426	.469	.499
	m9	.489	.604	.515	.520
	m14	.585	.601	.559	.588
	m16	.523	.417	.444	.504
	m19	.549	.552	.505	.568